

“© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Uncertainty Analysis for the Keyword System of Web Events

Junyu Xuan, Xiangfeng Luo, *Member, IEEE*, Guangquan Zhang,
Jie Lu, *Senior Member, IEEE*, and Zheng Xu

Abstract—Webpage recommendations for hot web events can assist people to easily follow the evolution of these web events. At the same time, there are different levels of semantic uncertainty underlying the amount of webpages for a web event, such as recapitulative information and detailed information. Apparently, the grasp of the semantic uncertainty of web events could improve the satisfactoriness of webpage recommendations. However, traditional hit-rate-based or clustering-based webpage recommendation methods have overlooked these different levels of semantic uncertainty. In this paper, we propose a framework to identify the different underlying levels of semantic uncertainty in terms of web events, and then utilize these for webpage recommendations. Our idea is to consider a web event as a ‘system’ composed of different keywords, and the uncertainty of this keyword system is related to the uncertainty of the particular web event. Based on Keyword Association Linked Network (KALN) web event representation and Shannon Entropy, we identify the different levels of semantic uncertainty, and construct a semantic pyramid to express the uncertainty hierarchy of a web event. Finally, a semantic pyramid-based webpage recommendation system is developed. Experiments show that the proposed algorithm can significantly capture the different levels of the semantic uncertainties of web events and it can be used for webpage recommendations.

Index Terms—web mining, web event, social event, uncertainty analysis, webpage recommendation

I. INTRODUCTION

A Web event could be a hot story or a social activity which attracts broad attention on the web and there could be an extraordinary number of webpages covering this web event. For example, the *Libya War (in 2011)* is a web event with thousands of webpages, blogs and posts. The large scale of webpages makes it impossible for users to grasp the evolution of a web event through manually surfing these webpages. Current researches on web events mainly focus on detecting them from the amount of webpages [1]–[5] and do the automatic summarization by selecting appropriate sentences [6]–[8]. In this paper, we focus on the uncertainty analysis of the web events and its application to webpage recommendations.

J. Xuan and X. Luo are with the School of Computer Engineering and Science, Shanghai University, China (e-mail: Junyu.Xuan@student.uts.edu.au; luoxf@shu.edu.cn).

J. Xuan, G. Zhang and J. Lu are with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS), Australia (e-mail: Guangquan.Zhang@uts.edu.au; Jie.Lu@uts.edu.au).

Z. Xu is with the Third Research Institute of the Ministry of Public Security, Shanghai 200031, China, and also with Tsinghua University, Beijing 100084, China (e-mail: xuzheng@shu.edu.cn).

Uncertainty is a big concept which is used to encompass many sub-concepts [9], [10]. According to different sources, uncertainty is categorized as: epistemic uncertainty [11], linguistic uncertainty [12], decision uncertainty [13] and variability uncertainty [14]. Of these, variability uncertainty refers to the diversity or heterogeneity of knowledge [10]. Uncertainty analysis is first defined as a process to quantify the uncertainty of a risk estimate and estimate the effect of this uncertainty on the outcomes [9], [10]. In the literature, there are many methods proposed for uncertainty analysis. For example, analytical methods include Delta method [15] and Point Estimation Method [16]; probabilistic methods include Monte Carlo simulation [17] and probability bounds/boxes [18]; graphical methods include Bayesian networks [19] and Loop Analysis [20]; and fuzzy methods include Fuzzy set [21] and Fuzzy cognitive maps [22]. To the best of our knowledge, there have not been any works proposed for the uncertainty analysis of web events.

A web event also has its semantic uncertainty. As shown in Fig. 1, a web event can be considered as a system composed of different keywords, and this keyword system, like other systems, has its own uncertainty. In this paper, the uncertainty of the keyword system is seen as the uncertainty (a kind of variability uncertainty) of this web event. This uncertainty is the measurement of the states of keyword systems which is the relative weights of different subtopics of a web event. For example, on Mar. 20, 2011, the web event *Libya War* has two subtopics: *Chinese stock market* and *Military attack*. If they have similar weights in this web event, which may be expressed by the same number of webpages or same number of people, this web event is not certain; If they have different weights in this web event, which may be expressed by the different number of webpages or different number of people concerning them, this web event is more certain than the former case. Since this uncertainty is a measure of the keyword system of a web event and keywords are the basic semantic atoms of a web event, it can also be called a *semantic uncertainty*. Note that the web event can be seen as a topic, like *Libya War*, and this web event/topic may have some subtopics.

Although there are many works on webpage recommendations, the semantic uncertainty of web events is seldom considered. The literature of webpage recommendations can be roughly classified into two categories: non-content-based methods and content-based methods. For non-content-based methods, the rating-based recommendations rely on the webpage-user ratings [23] that come from the user feedback. However, it is impractical to collect the feedback for webpages

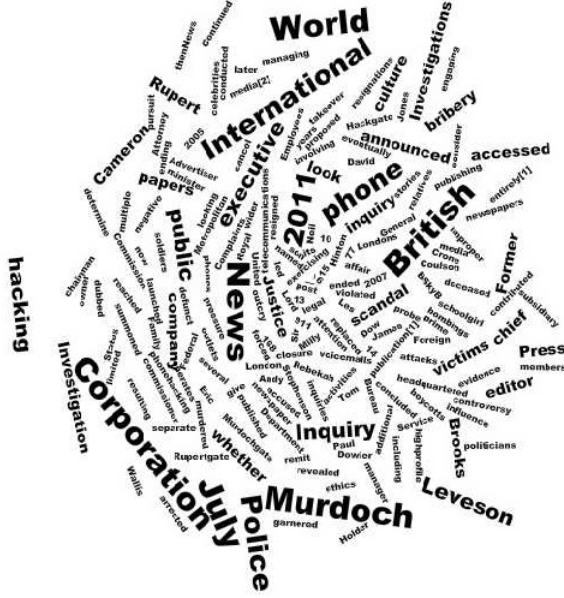


Fig. 1: An example keyword system of the web event ‘News International phone hacking scandal’. The content/semantic of this web event is determined by this keyword system, and its evolving trend is related to the states of this system.

of web events. Other methods, such as association rules [24] and Markov models [25], focus on capturing the sequential relations from the scanning/session history. These non-content-based methods do not consider the content of webpages. For the content-based methods, the texts of webpages are represented as vectors by VSM [26]. The recommendation is based on the matching between the user profiles and the webpages. Some other clustering-based methods [27] and ontology-enhanced methods [28] are used to improve performance. The problem is that ontology is difficult to construct from the dramatically evolving web events and the clustering is not enough. Therefore, it would be better to incorporate more analysis of the contents of web events.

The uncertainty analysis for a web event can assist websites to recommend appropriate webpages of web events to their visitors. Through the uncertainty analysis of the keyword system of a web event, we can unveil which part of the contents of a web event are active and attractive. For example, as mentioned above, there are about 7000 webpages covering the *Libya War* in a simplified Chinese web environment in one day. In order to know what this event, it is difficult and impractical for a user to read all of these webpages. There are two kinds of information in these webpages. 1) One is the certain part information, which will not change drastically with the evolution of a web event and can serve to provide the main content of the web event. For example, *Libya*, *anti-government*, *armed conflict* will exist in webpages most of time. 2) The other is the uncertain part information, which will markedly change with the evolution of a web event and will provide more details about the web event. For example, *stock* and *economy* only exist in webpages for a limited amount of time. For the past web events, we can easily distinguish

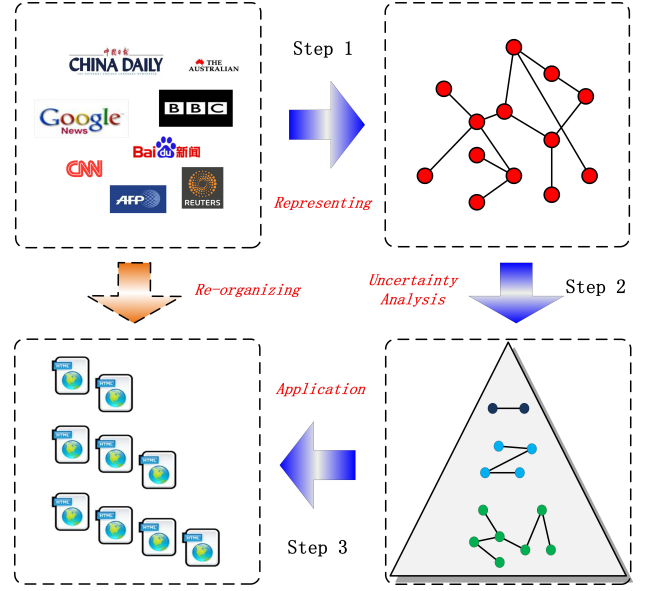


Fig. 2: The proposed method and framework of this paper. At first, the original webpages about a web event are collected, which may come from different sources, e.g. China Daily, BBC and Google news. Secondly, a flat KALN is constructed as the computational model for preserving the semantics of this web event. Then, a semantic pyramid is constructed for the uncertainty analysis. Finally, according to the mapping relations between webpages and keywords, the webpages with different uncertainties are recommended to the users.

certain and uncertain information through statistics. But, for a currently ongoing web event, we can only predict it by current data alone, especially in the initial stage. Thus, we need to do the uncertainty analysis for the web events based on their content. The current problem is how to define and perform the uncertainty analysis for web events, and how to apply this uncertainty analysis to the webpage recommendations for web events.

In this paper, we propose a method which can automatically analyze the semantic uncertainty of web events and the whole framework for this is shown in Fig. 2. The ultimate goal is to distinguish the different levels of the semantic uncertainties of keywords in a web event, which can help to recommend appropriate webpages to users. In order to achieve this goal, three important sub-questions need to be addressed: 1) How to semantically represent a web event at each time stamp with the webpages in hand? 2) How to correctly distinguish the levels of a web event with different semantic uncertainty at a given time? 3) How to utilize different level of uncertainties of a web event to recommend appropriate webpages? This paper answers these sub-questions by the proposed three steps shown in Fig. 2. In Step 1, a web event is represented by a Keyword Association Link Network (KALN) to preserve the semantics of a web event at a given time. This lays the foundation for further semantic uncertainty analysis. In Step 2, based on Shannon Entropy, an algorithm is proposed to identify the hierarchical uncertainty of the KALN, and then construct a semantic pyramid to express the hierarchical uncertainty of web events. Here, the keyword semantic uncertainty of a web event is measured by its weights distribution via Shannon Entropy.

This keyword weight distribution is defined to reflect not only the statistical property of keywords but also the local or global network structures of KALN. According to the properties of the different level of semantic uncertainties of a web event, a semantic pyramid is constructed to incorporate different levels of uncertainty semantics. In this semantic pyramid, the higher level has lower uncertainty semantics and the lower level has higher uncertainty semantics. In Step 3, each webpage of a web event can be attached to the constructed semantic pyramid through its keywords that have already been attached on the semantic pyramid. By the appropriate ranking strategies considering the keywords' positions in the semantic pyramid, the appropriate webpages can be selected to furnish the former proposed three categories requirements. This recommendation is based on the content of webpages not other information, like distinctive menu items and navigation indicators [26], [29].

The main contributions of this paper are summarized as follows:

- 1) The keyword association link network (KALN) has been mined and is considered as a new semantic representation to capture the semantic uncertainty of a web event at a given time;
- 2) Three strategies are proposed to define the weights of keywords in a KALN at a given time by considering three properties, i.e., document frequency, local network structure and global network structure;
- 3) The semantic uncertainties of a web event/KALN and all keywords are measured by utilizing Shannon entropy on the keyword weight distribution;
- 4) Based on the uncertainty of each keyword, the semantic pyramid has been constructed to express the different level semantic uncertainties of a web event. Three webpage ranking methods are proposed to recommend webpages by the constructed semantic pyramid.

The rest of this paper is organized as follows. In Section II, we do Step 1 in Fig. 2 which is to mine and analyze the flat KALN. The following two sections are designed for Step 2 in Fig. 2. Section III discusses three strategies to measure the impact of event keywords to the semantic uncertainty of a web event, and Section IV analyzes the hierarchical semantic uncertainty of a web event and constructs the semantic pyramid. Step 3 in Fig. 2 focuses on how to utilize the constructed the semantic pyramid to recommend webpages and this is shown in Section V. Section VI shows the experimental results and Section VII concludes this study.

II. FLAT KALN MINING AND UNCERTAINTY MEASURING

In this section, we will introduce Step 1 in Fig. 2 in detail, where a basic flat keyword network representation for the web events is proposed and constructed. Suppose we have a collection of webpages about a web event, e , which could come from news websites, blogs or forums. In this paper, this web event at a given time, t , is represented as,

Definition 1 (Keyword Association Link Network, KALN): A Keyword Association Link Network, Ω , which is composed of the keywords (as nodes) and their association relations (as

links) between keywords, is defined as,

$$KALN_t^e = \{S_{k,t}^e, S_{r,t}^e\} \quad (1)$$

where $S_{k,t}^e$ is the keyword set of web event, e , and $S_{r,t}^e$ is the association relation set of keywords at time, t , which are both extracted from the webpages of this event at time t .

1) *KALN Construction:* Given a collection of webpages about an event at a given time, t , by utilizing existing keyword extraction algorithms (i.e., Term Frequency and Inverse Document Frequency [30], tf-idf), we can get the nodes (keywords) of KALN from this data set. Once the nodes are fixed, the next step is to link these nodes by extracting the association rules between them. There are many state-of-the-art works on this subject. Since they are not the main concern of this paper, we will just select the Apriori algorithm [31] to get the association rules from the webpages. Finally, we connect keywords together by association rules to form the KALN.

Association rule/relation mining is a basic task of data mining and text mining. In Apriori algorithm [31], there are two weights given to each associated relation, like 'nuclear \rightarrow radiation', including support and confidence. Support is defined as,

$$support = \frac{W_{nuclear \cup radiation}}{W_{all}} \quad (2)$$

where $W_{nuclear \cup radiation}$ is the number of webpages containing words nuclear and radiation and W_{all} is the number of all webpages. With these association rules/relations, we can link the keywords together to form KALN.

Apparently, the more precise the keywords and the association rule extraction algorithms are, the better the event is described, and the KALN can express more about the real semantic uncertainty of an event.

Before the uncertainty analysis of web events, it is necessary to have a deep understanding of their representation KALN. A KALN is an expression of an event's semantics at a given time, which is composed of the keywords and association relations between them. Some other models or methods choose the distribution of keywords in the webpages to represent web events. In fact, not only the keywords but also their association relations should be considered in describing an event, because they are both basic semantic elements of an event and they almost play the same role on the semantic expression of web events. The reason why we call the constructed KALN as flat KALN is because we do not identify the uncertainty hierarchy in this section.

With the above definition in hand, we can consider the evolution of a web event as the variations of the KALN. Meanwhile, the semantics with different uncertainty hidden in these webpages can preserve more than the model which only considers keywords, because the association relations of keywords are considered here. Finally, the different level semantic uncertainty of KALN at a given time can be identified.

2) *Using Entropy as A Frame to Measure the Uncertainty of KALN:* The Entropy has been used to measure the uncertainty of a system. Here, we consider KALN as a system composed of keywords with different properties. Actually, a keyword in KALN has many different properties, such as Term Frequency

(TF), Document Frequency (DF) and Node Degree (ND). It should be noted that the association relations between keywords (i.e., network structure) can also be reflected by properties of keywords through the structure of a KALN. For example, the Node Degree (ND) can reflect the network structure of KALN. We combine the different properties (i.e., TF, DF and ND) of keywords together to generate a new property for reflecting all the properties simultaneously. This combined property is defined as,

Definition 2 (Keyword Weight, I): The Keyword Weight of a keyword in a web event is an integrated expression of the status of this keyword in this web event through combining its different properties.

So, if we select or combine different properties as the Keyword Weight of keywords to rank the keywords, the KALN's semantic uncertainty, namely web event semantic uncertainty, will be different. How to obtain the Keyword Weight, I , through the different properties of keywords will be discussed in the next section.

In a KALN, there are plenty of keywords with different abilities in terms of describing an event. For example, in the event 'Japan earthquake', the word 'earthquake' exists from the beginning to the end. By contrast, the word 'rescue' only exists for a period of time. How to identify the different abilities of these keywords based on their Keyword Weight values and then to measure the semantic uncertainty of a web event? Here, we introduce the entropy of KALN,

Definition 3 (Keyword Distribution Entropy, H_{KALN}): The keyword distribution entropy is the measurement of semantic uncertainty of KALN, which can be computed by,

$$H_{KALN} = - \sum_{j=1}^{N(KALN)} p_j \cdot \log p_j, \quad p_j = I_j / \sum_{k \in S(KALN)} I_k \quad (3)$$

where H_{KALN} is the entropy value of KALN, $N(KALN)$ is the keyword number of KALN, $S(KALN)$ is the keyword set of KALN, I_j is Keyword Weight value of keyword j in the whole KALN and p_j is the proportion of keyword j 's weight to the sum of all keywords in the KALN. If the number of keywords, $N(KALN)$, is fixed, the value scope of H_{KALN} is $(0, \log N]$.

From this definition, we can see that the Keyword Distribution Entropy is based on the famous Shannon Entropy. To be specific, its evaluation in Eq. 3 follows the form of Shannon Entropy, but the p_j is defined as the weight of keyword j . The computation of weights of keywords, I_j , will be discussed in Section 4. More discussion about Eq. (3) can be found in [32].

III. THREE STRATEGIES TO OBTAIN THE KEYWORD WEIGHT

As discussed in the previous section, the semantic uncertainty of web events is defined by the keyword distribution entropy, which is determined by the values of the keywords' weights. Apparently, different Keyword Weight computation strategies will lead to different keyword distributions and then lead to different keyword distribution entropies. In this section, three different strategies, which have taken various properties

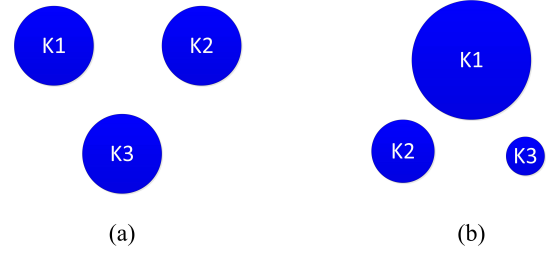


Fig. 3: Schematic figure of the ratio of Document Frequency of keywords in order to express the semantic uncertainty of a web event. Two graphs/KALNs in this figure represent two statuses of a web event. Each circle in the graph represents a keyword in a KALN, and the circle size denotes the value of DF of a keyword. The right KALN (b) is more certain than the left one (a), and the semantics of the right web event are more certain than left one, because it has a dominant keyword (K1) which has a great impact relative to the others in the left KALN.

of keywords into consideration, are designed to compute the weights of keywords. This section is for Step 2 in Fig. 2, where the hierarchy of the keywords is identified in order to consider the properties of keywords.

A. Strategy One: Document Frequency

The basic property of keywords is the Document Frequency (DF), which means the frequency of a keyword as shown in a document set. DF reflects the keyword's volume. It is easy to understand that the more frequently this keyword is used to cover a web event, the more important this keyword is relative to this web event, as shown in Fig. 3. The computation of the weight of keyword j can be simply defined by,

$$I_j = \alpha_j / \sum_{i \in S(KALN)} \alpha_i \quad (4)$$

where α_j is the Document Frequency of keyword j and $\sum_{i \in S(KALN)} \alpha_i$ is the summation of document frequencies of all keywords in a KALN.

For example, the DF of keyword 'earthquake' is bigger than keyword 'evacuate' in the web event 'Japan earthquake'. The 'earthquake' will impact more on the semantic uncertainty of web event 'Japan Earthquake' than 'evacuate'.

B. Strategy Two: Document Frequency & Node Degree

The property 'DF' only considers the volume of a keyword. However, a keyword in KALN has not only the node weight, but it also has the structural property. The structure of KALN reflects the structure of knowledge of a web event at a given time, which is as important as the DF property. In the structural perspective, different keywords have different weights according to their structural property in a KALN. Here, the Node Degree (ND)¹ is adopted to measure the structural property of keywords by,

$$I_j = \frac{1}{2} \left(\frac{\alpha_j}{\sum_{i \in S(KALN)} \alpha_i} + \frac{\beta_j}{\sum_{i \in S(KALN)} \beta_i} \right) \quad (5)$$

¹[http://en.wikipedia.org/wiki/Degree_\(graph_theory\)](http://en.wikipedia.org/wiki/Degree_(graph_theory))

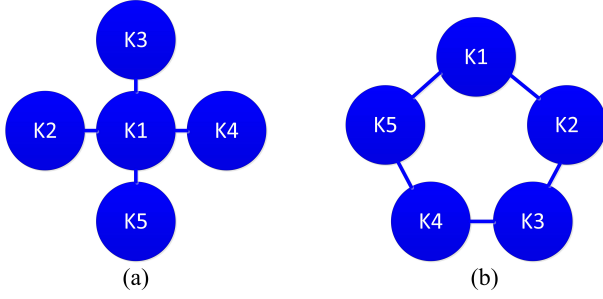


Fig. 4: Schematic figure of the function of Node Degree (ND) in order to express the semantic uncertainty of web events. Two networks/KALNs in this figure represent two statuses of a web event. Each circle in the graph represents a keyword in a KALN. The left one is more certain than the right one and the semantics of the left web event are more certain, because it has a dominant keyword (K1) which has a bigger Degree (ND) than the others but all the keyword Degrees (ND) in the right KALN are equal.

where β_i is the Node Degree of keyword j and $\sum_{i \in S(KALN)} \beta_i$ is the summation of Node Degree of all the keywords.

In order to use the Eq. (3), the formula of Eq. (5) is designed to make sure that p_j is a probability ($p_j > 0, \sum p_j = 1$). The DF and the ND are simply set to have same status in the Equation.

In order to thoroughly understand the ND of a keyword, we discuss it in detail next. At first, the reason why we choose the Degree is that it is a simple and fundamental parameter of a complex network and it reflects the local network structural property. As shown in Fig. 4, the keyword which has a big Degree means it is a dominant aspect of this web event. ND of a keyword is formatted by its associated relations with other keywords. It is another measurement dimension relative to DF. Although a keyword does not have a big DF, it may also be a dominant aspect if it has a big Degree. The bigger the ND is, the more important this keyword is. However, ND only reflects the local structural information of a keyword. It cannot ‘see’ the entire network structure. It is just a measure of relations surrounding a keyword. In the next subsection, we will provide a way to measure the global structural information of a keyword.

C. Strategy Three: Document Frequency & Power Law Contribution

Bianconi [33] think that the appearance of the power-law distribution of Degree reflects the tendency of social, technological, and especially biological networks toward “ordering”. Further, Barabasi [34] even suggests that the power-law phenomenon can be viewed as the hidden pattern of everything we do in our life. Since a web event is actually a social activity, it should also have the trend toward “ordering”, thereby satisfying the power-law distribution. So, a hypothesis is proposed:

A web event will reach its final certainty status only when its keywords’ weight distribution in KALN satisfies the “power-law distribution”.

For example, at the beginning of the web event *Japan earthquake*, there are many subtopics canvassed by people.

Algorithm 1 Computation of a keyword’s PLC

Input: Keywords’ degrees in KALN

Output: PLC of a keyword PLC_k

- 1: Obtain the straight fitting line of the Degree log-log curve, $Ax + By + C = 0$;
- 2: Evaluate the fitting error err through Eq. (6);
- 3: Remove keyword k ;
- 4: Obtain a new Degree log-log curve and its new straight fitting line;
- 5: Evaluate the new fitting error err_{new} through Eq. (6);
- 6: $PLC_k = |err - err_{new}|$.

As the time passes, however, there will be a limited number of main subtopics which will emerge. These subtopics will grab the most attention. The remaining subtopics will be less frequently discussed.

According to this hypothesis, a new measurement different from ND, named Power-Law-Contribution (PLC), is proposed to measure the global structural information of a keyword. The more a keyword contributes to the KALN’s degree power-law distribution, the more important this keyword is. The procedure of measuring a keyword’s PLC is shown in Algorithm 1.

In the Algorithm 1, the Degree log-log curve is the curve of data pairs $\langle \log(d), \log(num_d) \rangle$ where d is the value of Node Degree and num_d is the number of nodes with Node Degree d in KALN. The Degree log-log curve [35] is a classical method to evaluate the power-law distribution of Node Degree. We first do the straight line ($Ax + By + C = 0$, x denotes value of $\log(d)$ and y denotes the value of $\log(num_d)$) fitting for the Degree log-log curve, then the fitting error is evaluated as,

$$err = \sum_d \frac{|A \log(d) + B \log(num_d) + C|}{\sqrt{A^2 + B^2}}. \quad (6)$$

After removing a keyword k , the Degree log-log curve will change due to the missing of keyword k . Then, we re-fit a straight line for Degree log-log curve, and re-evaluate the fitting error, err_{new} . Finally, the PLC of keyword k is $PLC_k = |err - err_{new}|$.

Following this Algorithm 1, each keyword’s PLC can be computed. Then, we can compute the final weight of keywords as follows,

$$I_j = \frac{1}{2} \left(\frac{\alpha_j}{\sum_{i \in S(KALN)} \alpha_i} + \frac{\gamma_j}{\sum_{i \in S(KALN)} \gamma_i} \right) \quad (7)$$

where γ_j is the PLC of keyword j and $\sum_{i \in S(KALN)} \gamma_i$ is the summation of PLC of all keywords. The formula of Eq. (7) is used to make sure that p_j is a probability ($\sum p_j = 1$). The DF and the PLC are set to have the same status in the Equation here.

The aim of PLC is similar with Node Degree (ND), and the keyword with a big ND tends to have a big PLC, too. But there are significant differences between these two measurements of network structural information. At the theoretical level, the PLC, which considers the global network structure, is

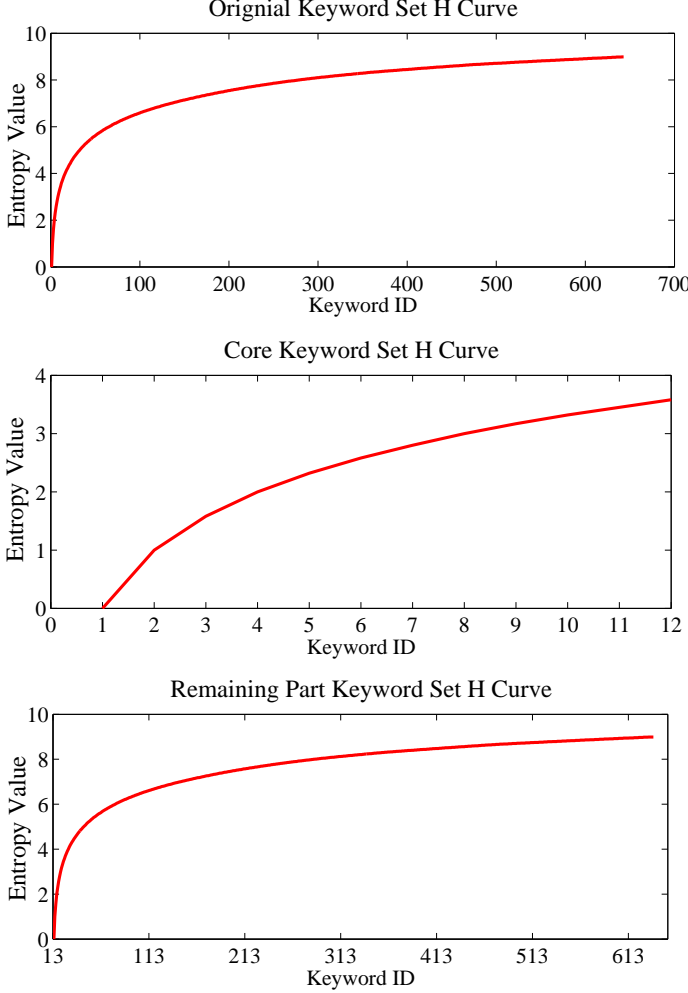


Fig. 5: The entropy values with the variation of the keyword set of KALN with the first strategy. The new keyword is added into the keyword set in the descending order of DF. The contribution of a keyword to the semantic uncertainty is reflected by the change of entropy value after a keyword is added into KALN. Top subfigure is all the keywords, middle one is the core keyword set of the initial keyword set and bottom one is the remaining part of the initial keyword set. The x-axis is the keyword number and the y-axis is the entropy value. (Web event: Japan earthquake, Date: Mar. 9, 2011)

from the node/keyword contribution of the KALN's power-law distribution, but ND, which considers the local network structure, is from the Degree of the node/keyword. At the computational level, the PLC computation not only has a relation with the Degree of the keyword, the Degrees of neighbors are also considered. The keyword with a big Degree does not definitely have big PLC, and vice versa.

D. Measurement of influence of each keyword to the KALN entropy

After three different strategies of computing the weight of keywords are introduced, we can utilize the Keyword Distribution Entropy, H_{KALN} , to evaluate the influence of each keyword to the KALN entropy (i.e., the semantic uncertainty of web events). In order to do that, a procedure is designed

to increase the number of keyword in KALN (i.e., the value of KALN entropy) one by one with the Keyword Weight in descending order. Then, a series of entropy values of KALNs with a different number of keywords is obtained and forms a curve. The change between two points on the curve just reflects the influence of the keyword at the second point of the semantic uncertainty of this web event. The procedure is described in Algorithm 2,

Algorithm 2 Compute KALN entropy with a given keyword

Input: A keyword set, S_k assigning each keyword a weight value, I_j , by Eq. (4), (5) or (7)

Output: The values of keywords' influences to the KALN entropy

- 1: Initial an empty KALN without keywords
 - 2: **while** S_k is not empty **do**
 - 3: Select the keyword with maximum I_j from S_k to add into KALN
 - 4: Remove this keyword from S_k
 - 5: Compute the entropy, H_i , of KALN by Eq. (3)
 - 6: **end while**
 - 7: Finally, we get $\langle H_0, H_1, \dots, H_n \rangle$.
-

The advantage of this descending order is that the final entropy value curve is relatively smooth and monotonously as compared to a situation where keywords are randomly added.

The reason why we construct this series of entropies is that the change of two consecutive entropy values in the series can show the impact on the former KALN by adding a new keyword. Apparently, the bigger the change is, the bigger impact from adding a new keyword and the more powerful this keyword can be in terms of characterizing this event at a given time. This change reflects the influence of a keyword to the semantic uncertainty of the web event. Indeed, the bigger the change in terms of entropy values, the bigger the influence of adding a new keyword. As shown in Fig. 5, the shape of the entropy value curve, which is strictly ascending and ascending rate continually becomes small, well matches the properties of the keyword distribution entropy.

After the introduction of three ranking strategies (i.e., computation strategies of Keyword Weight) of keywords of a web event, it would appear that different keywords have different contributions to the KALN's entropy and semantic uncertainty of this web event. There is actually a hierarchical structure underlying the web event. Our next task is to identify, discuss and formally define it.

IV. SEMANTIC PYRAMID OF WEB EVENTS

This section is for Step 2 in Fig. 2, where three layers of keywords are recognized from the flat keyword network. We firstly analyze the hierarchical property of KALN through the entropy value curve constructed in the previous section. Then, the semantic pyramid of a web event is constructed and discussed to express the hierarchical uncertainty of this web event. How we utilize this pyramid will be given in the next section.

A. Hierarchical property of KALN

As discussed in Definition of Keyword Distribution Entropy, there are two factors which influence the semantic uncertainty of the KALN: the number of keywords and their ratios. In this section, we assume the number of keywords is fixed and focus on the second factor.

As the procedure of entropy computation discussed above, the change of two points' values on the entropy curve reflects the importance degree of the latter keyword relative to the former KALN. The bigger the curvature that a point has, the smaller the contribution this keyword makes to the former KALN and the more certain the former KALN is. One explanation is that if the characteristic reflected by the new added keyword has little impact on the semantic uncertainty of a web event, this characteristic or this keyword can be ignored. Then, the initial keyword set can be split into two parts.

After the splitting, it can be found that the same keyword may have different impacts on the uncertainty of two different parts. The reason for this phenomenon is that the KALN has the hierarchical property of semantic uncertainty. The low layer keywords with high semantic uncertainty may have a big impact on the low layer KALN, but it has a small impact on the high layer KALN with low semantic uncertainty at the same time. For example, in the event *Japan Earthquake*, the keyword *earthquake* belongs to high layer KALN and the keyword *evacuate* belongs to a lower layer KALN. The keyword *evacuate* has a big impact on the sub event *Embassy and Corporation Evacuation* but has small impact on the whole event *Japan Earthquake*.

In order to identify this hierarchical property of the semantic uncertainty of KALN, the entropy value curve constructed in the previous section is analyzed. Here, we use the curvature of a point on the curve to denote the influence degree of the keyword added at this point. In mathematics, curvature of a plane curve reflects the bend sharpness of a curve. To the entropy value curve, this means how much the addition of a new keyword makes or contributes to the characterization of an event. The bigger the curvature of a point is, the smaller contribution that the keyword has added at this point to the semantic uncertainty of the web event. The core keyword set contributes more to the whole entropy value of the flat KALN as compared to the remaining keyword set. This means that the semantic uncertainty of a web event mainly depends on the core keyword set. When people comprehend these core semantics, they can understand the web event in general. A detailed algorithm is introduced in [32].

B. Construct semantic pyramid

According to the different curvatures of points, the initial keyword set can be divided into two parts with the boundary of the average value of curvatures of all keywords. As shown in Fig. 6, we can get a binary tree by dividing the keyword set recursively, whereby the nodes are the different keyword sets and the root node is the initial keyword set. In this tree, the positions of different keyword sets reflect their different semantic uncertainty levels. The higher the position of a keyword set, the more certain it is. Every node's left child

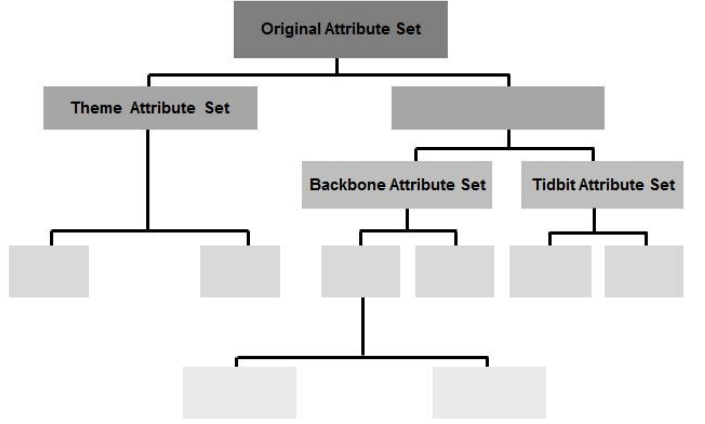


Fig. 6: The binary tree representation of a web event. The root of this tree is the initial keyword set of a web event and the left child of it is the core set of the root and the right child is the left set. The marked three nodes in this tree are selected to construct the semantic pyramid of the web event. The more upper the keyword set is, the more this keyword set contributes to the semantic uncertainty of this web event and the more likely it represents the main part of this web event.

node contains the core keywords of this node and can well characterize this node. Through this tree, we have a thorough understanding about the semantic uncertainty of keywords in KALN at this given time. Here, we just split the initial keyword set into three layers, as shown in Fig. 7.

At first, we get two Split Points for division:

$$\begin{aligned} \vartheta_{p1} &= \frac{\sum_{i \in S_k^{KALN}} \vartheta_i}{N^{KALN}} \\ \vartheta_{p2} &= \frac{\sum_{i \in S_k^{KALN'}} \vartheta_i}{N^{KALN'}} \end{aligned} \quad (8)$$

where S_k^{KALN} is the keyword set of KALN, N^{KALN} is the number of keywords in KALN, ϑ_{p1} is the first Split Point and the initial keyword set in KALN is split at this point. The Theme keyword set is obtained from this splitting. And the KALN is the remaining part, which is split at the second Split Point. The Backbone keyword set and Tidbit keyword set are formed through this splitting.

Then, we give each layer's definition as follows,

Definition 4 (Theme Layer KALN, Ω^I): The theme layer KALN is comprised of the keywords, which satisfy the condition that ϑ is bigger than ϑ_{p1} , and the association rules between them. This layer network is the core of the flat KALN. It expresses what this KALN or this event is referring to and has less semantic variation over time.

This layer network is located at the top of the semantic pyramid and contains the fewest keywords. However, it is these keywords that reflect the core semantics of a web event and have the biggest influence in terms of semantic uncertainty. Over time, it will have the smallest change compared to the other two layer networks, and the web event will have a large evolution if this layer network experiences little change.

Definition 5 (Backbone Layer KALN, Ω^{II}): The backbone layer KALN is comprised of keywords, which satisfy the

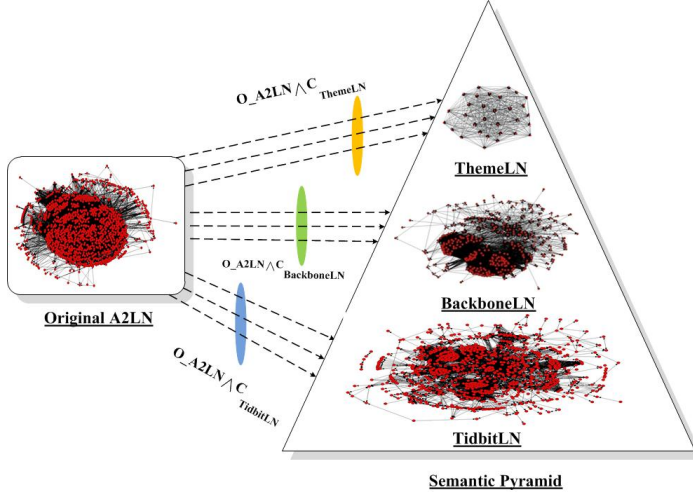


Fig. 7: The mined Semantic Pyramid (SP) from the flat KALN. The left part network is flat KALN. In the right part SP, three networks are Theme Layer Network, Backbone Layer Network and Tidbit Layer Network from top to bottom. (Web event: Japan earthquake, Date: Mar. 9, 2011, Using first strategy)

condition that ϑ is smaller than ϑ_{p2} and bigger than ϑ_{p1} , and the association rules between them. This layer network is the backbone of the flat KALN. It gives more detail than the Theme Layer network and shows the main semantics of this event. Over time, it will change more drastically than the Theme Layer.

This layer network is located at the middle of the semantic pyramid and it contains the medium number of keywords. It may contain all the sub-events' semantic uncertainty and become the backbone of web events. Over time, it will have a medium level of semantic change compared to the other two layer networks.

Definition 6 (Tidbit Layer KALN, Ω^{III}): The theme layer KALN is comprised of keywords, which satisfy the condition that ϑ is smaller than ϑ_{p2} , and the association rules between them. This layer network represents the tidbit of the flat KALN. It gives detailed semantics of all aspects of this event and is the most sensitive to time in the three layers.

This layer network is located at the bottom of the semantic pyramid and contains the most keywords. It further details the web events. As the time passes, it will experience the biggest semantic change compared to the other two layer networks, because it contributes the least to the semantic uncertainty of a web event. Even if it had a big semantic change over time, the web event would have little evolution.

Definition 7 (Semantic Pyramid of KALN, Λ): The Semantic Pyramid, Λ , is consisted of ThemeLN, Ω^I , BackboneLN, Ω^{II} , and TidbitLN, Ω^{III} , each of which is a specific semantic level of KALN.

$$\Lambda = \left\langle \begin{matrix} \Omega^I \\ \Omega^{II} \\ \Omega^{III} \end{matrix} \right\rangle \quad (9)$$

In our tests, we found that the numbers of keywords and their association relations of each layer network are approximately one order smaller than the lower layers, as Table

TABLE I: COMPLEX NETWORK PARAMETERS OF THREE LAYER NETWORKS (WEB EVENT: 'JAPAN EARTHQUAKE', DATE: Mar. 9, 2011, USING FIRST STRATEGY)

Parameters	KALN		
	Theme	Backbone	Tidbit
Number of nodes	12	48	584
Number of arcs	128	986	5346
Density	0.97	0.437	0.016
Average Degree	10.667	20.542	9.154
Clustering Coefficient	0.973	0.756	0.825
Mean Distance	1.03	1.661	3.325
Diameter	2	4	8

I shows. Of course, it depends on the DF distribution of keywords. However, the density remarkably decreases from the top layer to the bottom layer. The removal of a small number of keywords leads to a big decrease of arcs/relations between keywords, which expresses the fact that different keywords have different characters and that a small number of them can lead to a big number of arcs/ relations. This big density in the higher layer reflects the fact that certain keywords of a web event tend to connect with each other and vice versa. The diameter and the mean distance of the Theme layer network are 2 and 1.03. This means the distance between two nodes/keywords is smaller than 2. It is close to a complete graph and the addition of a new keyword almost needs to link to every node/keyword of the old network. This condition is so 'rigid' that the theme layer KALN's structure is the most semantic certain. During the evolving process, this rigid condition will make the change of this layer very difficult. So this layer KALN reflects certain semantics. But, if the change does happen, it will lead to a big change in the semantic uncertainty of a web event. By contrast, the density and the diameter of Tidbit Layer KALN are 0.016 and 8 respectively, which means this layer network is a sparse network. It is not sensitive to losing a node or adding a new node/keyword. So this layer KALN reflects semantic uncertainty. This means the missing or incoming nodes in this layer will not lead to big changes in the semantic uncertainty of a web event.

So far, we have a deep understanding of a semantic uncertainty of a KALN by constructing the Semantic Pyramid composed of three layer networks. Meanwhile, through recursively splitting the core set of initial keywords and terminating them at a desired position, the most characteristic keywords can be identified by our method. In the experiments, we will certify if the semantic uncertainty of each layer of Semantic Pyramid satisfies our assumption: The higher the layer is, the more certain it is. Next, we will show what this semantic pyramid can do.

V. SEMANTIC PYRAMID-BASED WEBPAGE RECOMMENDATION

This section corresponds to Step 3 in Fig. 2, where the recognized hierarchial keyword network is applied for webpage recommendations. For a user who wants to follow a web event, it will be impossible to read all the related

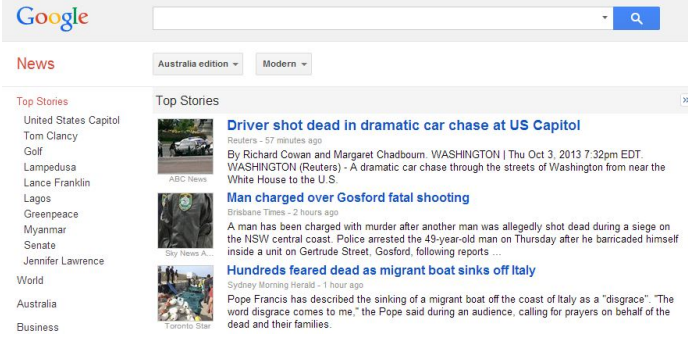


Fig. 8: The main page of *Google News*. There are a number of web events with hyperlinks to the recommended webpages. These webpages are the most representative and should express the main content of each web event well.

webpages about this web event owing to the huge number of webpages emerging each day. Fortunately in this paper, a semantic pyramid has been constructed from these webpages to represent and organize all the semantics of a web event on a given time. It can be viewed as a mental structure constructed after the reading of all the webpages by a human. Here, we automatically complete this task and then assist people by this semantic pyramid. In this section, we propose three kinds of requirements of users and corresponding recommendation algorithms based on our semantic pyramid.

A. The most certain webpages recommendation based on Theme level KALN

For a user who just starts to focus on a web event, the most certain webpages will enable them in order to quickly grasp the main semantics of this web event. For a website, like *Google News*, which aims to show the main content of each web event with limited webpages, the most certain webpages can best represent a web event, as shown in Fig. 8. For our semantic pyramid, this means the webpages contain the keywords in the Theme level KALN. A criteria is proposed for a webpage,

$$C(w) = \frac{\sum_{k_i \in \Omega^I \cap w} I_i}{\sum_{k_i \in \Lambda \cap w} I_i} \quad (10)$$

where $C(w)$ is the certainty of a webpage w and $k_i \in \Omega^I \cap w$ denote the keywords, k_i , both in ThemeLN, Ω^I , and webpage, w . After ranking by $C(w)$, the top N webpages are selected as recommendations for users or websites.

B. The most uncertain webpages recommendation based on Backbone level KALN

For the continuously updated web events, the users who have been following these web events just want to know the information that is more uncertain and which has a large potential to cause the evolutions of web events. For our semantic pyramid, this means the webpages contain the keywords in the Tidbit level KALN. A criteria is proposed for a webpage,

$$U(w) = \frac{\sum_{k_i \in \Omega^{III} \cap w} I_i}{\sum_{k_i \in \Lambda \cap w} I_i} \quad (11)$$

where $U(w)$ is the uncertainty of a webpage w and $k_i \in \Omega^{III} \cap w$ denote the keywords, k_i , both in TidbitLN, Ω^{III} , and webpage, w . After ranking by $U(w)$, the top N webpages are selected as recommendations for users.

C. The directional webpages recommendation based on Tidbit level KALN

Some users just want to know a specific aspect about a web event, and the correlated webpages should be carefully selected to recommend to them. Normally, this specific aspect is in the second or third level KALN of semantic pyramid. It is more likely to search for webpages in all the web event data. The traditional search method by existing search engines using the web event name and the specific keywords has a problem: the returned webpage with other ‘noisy’ keywords may not focus on the desired aspect about this web event. In order to do the directional webpages recommendation, a criteria is proposed for a webpage,

$$D(w) = \frac{\sum_{k_i \in w \cap K} w_{k_i}}{\sum_{n=I, II \text{ and } III} \left(\sum_{k_i \notin K \text{ and } k_i \in \Omega^n \cap w} \rho^n \cdot I_i \right)} \quad (12)$$

where $D(w)$ is the correlation of a webpage w to the desired keyword set, K , $\sum_{k_i \in w \cap K} w_{k_i}$ denotes the matching degree of w to K , k_i , and the denominator is for removing the undesired information from w . ρ^n are the coefficients of three levels in $[0, 1]$.

The whole equation aims to make the candidate webpage more focused on the desired information. The more a webpage is matched to the desired information, the better the webpage is. At the same time, the less other keywords a webpage contains, the better the webpage is. However, the ‘noisy’ keywords in different levels should be differentiated. This is because the desired information about a web event is more likely in the third level of a semantic pyramid. If the noisy keywords are also in the third level, the content of this webpage may be away from the desired information. So, the ρ^3 will be big and set as 1. By contrast, if the noisy keywords in the first level, the influence of them will be small. The ρ^1 is set 0.2 and ρ^2 is set 0.5 in this paper.

VI. EXPERIMENT AND EVALUATION

A. Data preparation

We introduce the datasets used in this paper:

- 1) **Dateset 1** It is the webpages of web event *Japan earthquake* (in 2011) from Mar. 9, 2011 to Apr. 20, 2011. The number of webpages is 6,884. All these webpages are collected from search engines, including *www.Google.com.hk* and *www.Baidu.com*², regardless of the sources of these webpages.
- 2) **Dateset 2** It is also the webpages of web event *Japan earthquake* (in 2011) from Mar. 9, 2011 to Apr. 20, 2011, but these webpages are collected using the source (i.e., news websites, blogs, forums) interface provided by

²<http://www.baidu.com> is the biggest search engine in China.

www.Google.com.hk. The numbers are 3,059 (from news websites), 4,533 (from blogs) and 3,535 (from forums).

- 3) **Dataset 3** It is a collection of web events. There are 50 hot web events in China collected through the source (i.e., news websites, blogs, forums) interface of *www.Google.com.hk*. The average length of these web events is 30 days, and the total number of webpages is 202,673.

B. Evaluation of the semantic uncertainty of each layer in KALN

The generated three layers KALN represent the different semantic uncertainty levels of a web event at a given time. Here, the experiment is conducted to evaluate the semantic uncertainty of each layer. If the keywords in a layer will always stay in this layer over time, this layer KALN can be viewed as containing stable semantics and a high certainty. Based on this evaluation metric, the four correlation coefficients are given as:

coco 1 the frequency of a keyword *at* current time stamp *vs.* the frequency of this keyword *after* current time. The meaning of this parameter is the possibility of a keyword still showing up in the future if it appears at the current time stamp;

coco 2 the document frequency of a keyword *at* current time stamp *vs.* the *sum* of document frequency of this keyword *after* current time. The meaning of this parameter is the possibility of a keyword still having big document frequency in the future if it has big document frequency at the current time stamp;

coco 3 the frequency of a keyword *before* current time stamp *vs.* the frequency of this keyword *after* current time. The meaning of this parameter is the possibility of a keyword still showing up in the future if it appears before the current time stamp;

coco 4 the document frequency of a keyword *before* current time stamp *vs.* the *sum* of document frequency of this keyword *after* current time. The meaning of this parameter the possibility of a keyword still having big document frequency in the future if it has big document frequency before the current time stamp.

If the semantics of a layer of KALN is certain, it will not change too much as time goes and then these four correlation coefficients will tend to be large. The differences of four correlation coefficients are: 1) coco1 and coco3 just considers the frequencies of the keywords, but coco2 and coco4 consider the weight of keywords. To be specific, the keywords with different weights will make different contributions to the coco2 and coco4, so this difference will make coco2 and coco4 more reasonable and relatively larger than coco1 and coco3; 2) the coco3 and coco4 compare two periods of time of the entire time span split by the given time, but coco1 and coco2 only compare the current time stamp with the time duration from next time stamp to the end. This difference will make the coco1 and coco2 more random and relatively larger comparing with coco3 and coco4.

Apparently, the bigger correlation coefficients that a layer has, the more certain is the layers of the semantics. Here, the Pearson's Correlation Coefficient is selected to compute the four correlation coefficients as,

$$r_{XY} = \frac{\sum_{i=1}^N (X_i - \bar{X}) (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}} \quad (13)$$

Then, these four correlation coefficients are computed on the three layers of the KALN respectively, which are conducted on three kinds of sources (the DF strategy) using Dataset 2, as shown in Fig. 9. Except the single web event example in Fig. 9, the results on more web events of Dataset 3 are summarized in Table II. In each cell of Table II, the average of cocos from 50 web events in the specific source and layer is given.

Based on these results, it can be confirmed that the ThemeLN's semantics are the most certain, the TidbitLN's semantics are least certain and BackboneLN's semantics are median. Comparing the results of different sources in Fig. 9, the blog is the most semantic certain and the news is the most semantic uncertain. Compared to the news, the blogs are written by people after considered thought. By contrast, the news webpages are more quick, direct and diverse and are published by journalists for the purposes of attracting more readers. There will be more keywords to be added in the keyword set of web events in news. So it would be more likely that the difference between the keyword sets of news at different time stamps is bigger than the one of blog. This finally leads to the curves of correlation coefficients of news which fluctuate more significantly than blog. The forums have relatively similar behavior to be news. Meanwhile, the coco2 and coco4 are bigger than coco1 and coco3 respectively. The reason is that the coco2 and coco4 consider the former period of time before a time stamp, while the coco1 and coco3 only focus on a single time stamp. However, it does not mean that coco1 and coco3 are inaccurate and useless. They are just a local view relative to the global view of coco2 and coco4. And the results show that they have the same trends with each other. Interestingly, the curves in Fig. 9 tend to be high in the middle area. This is due to their definitions. In their definitions, the comparisons are 'current *vs.* future' (coco1 and coco2) and 'before *vs.* future' (coco3 and coco4). So the values of them are relative to the current point, especially coco3 and coco4. At the different sections of time-axis, the keywords sets of 'before' will be different. For example, at the start point, the keyword set of 'before' is null and the keyword set of 'future' is universal set. At the end point, the keyword set of 'before' is universal set but the keyword set of 'future' is null. As for the curves, this will lead to being relatively big in the middle and small at the ends of the curves. But it does not impact the comparison of different layer networks and different sources.

C. Correlations of Three Strategies

As proposed in Section 4, three keyword ranking strategies consider different properties of keywords, which can define the weights of keywords from the statistical and structural of KALN. However, if the ranking results of the three strategies are the same, two of them will be meaningless. We therefore

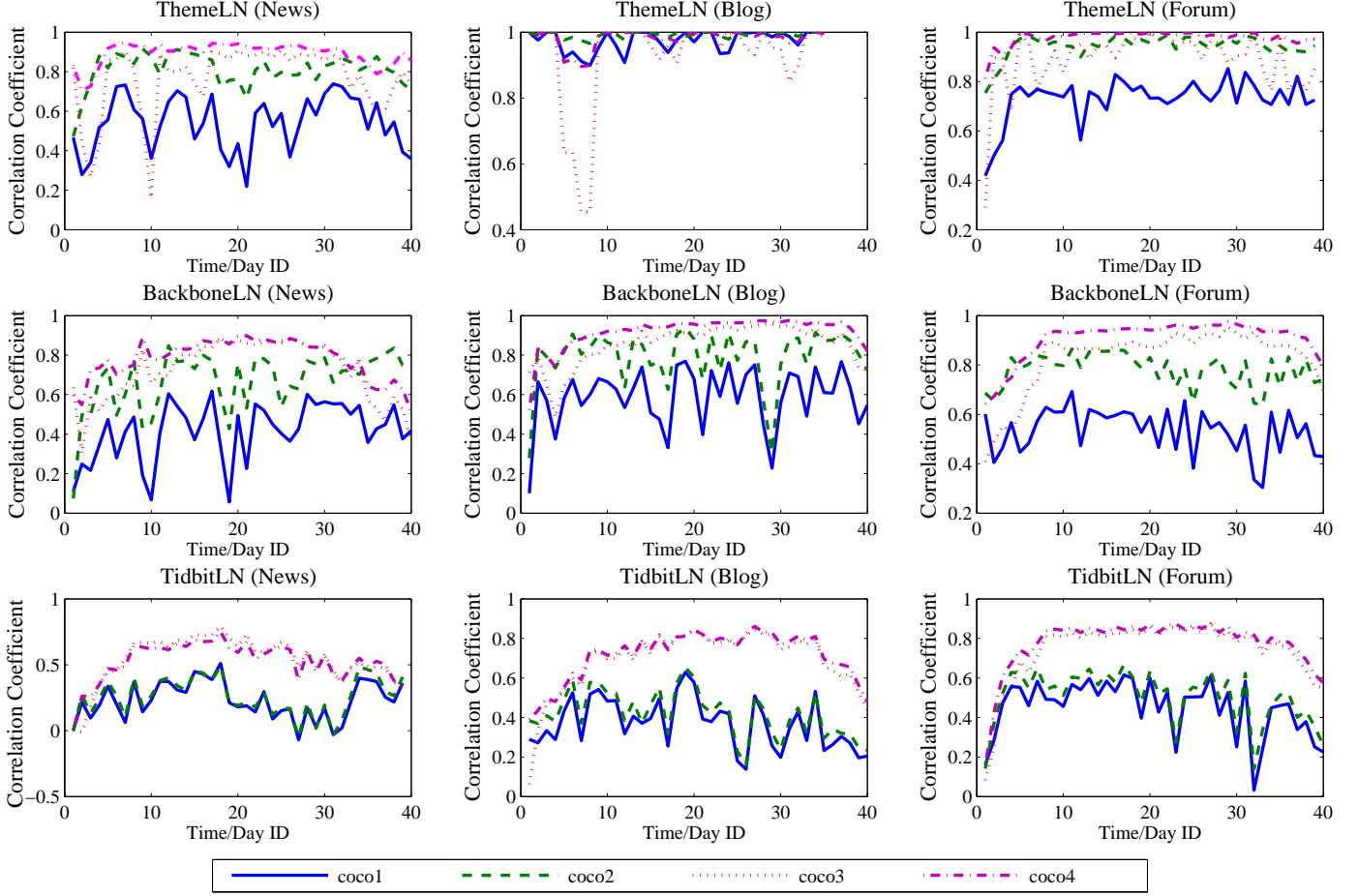


Fig. 9: Four correlation coefficients of three layers KALN, including ThemeLN, BackboneLN and TidbitLN. It can be found that four evaluation metrics have relatively similar trends. The variation range of the ThemeLN is the smallest in these three layers. The TidbitLN's variation range is the biggest. This suggests that ThemeLN's semantics is the most stable one, TidbitLN's semantics is the most unstable one and BackboneLN's semantics is the medium one. In the evolution process, the different layer networks show different behaviors. These results satisfy the definitions. (Web event: Japan earthquake. Source: News. Start Time: Mar. 9, 2011. End Time: Apr. 20, 2011, Using first Strategy)

TABLE II: FOUR CORRELATION OF THREE LAYERS KALN. (WEB EVENT NUMBER : 50, AVERAGE LENGTH : 30 DAYS, WEBPAGES NUMBER : 202,673, USING FIRST STRATEGY)

Sources	news			blog			forum		
Layers	ThemeLN	BackboneLN	TidbitLN	ThemeLN	BackboneLN	TidbitLN	ThemeLN	BackboneLN	TidbitLN
COCO1	0.47	0.53	0.13	0.92	0.55	0.35	0.7	0.48	0.35
COCO2	0.71	0.65	0.20	0.97	0.68	0.4	0.92	0.72	0.40
COCO3	0.54	0.59	0.53	0.79	0.63	0.6	0.81	0.61	0.67
COCO4	0.80	0.74	0.54	0.9	0.85	0.62	0.85	0.81	0.71

TABLE III: CORRELATION OF THREE KEYWORD RANKING STRATEGIES. S1 IS STRATEGY I; S2 IS STRATEGY II; S3 IS STRATEGY III (WEB EVENT NUMBER : 50, AVERAGE LENGTH : 30 DAYS, WEBPAGES NUMBER : 202,673)

	news	blog	forum
CO(S1,S2)	0.301889395	0.240520556	0.260074409
CO(S1,S3)	0.258960083	0.28971901	0.289943821
CO(S2,S3)	0.181041884	0.193622537	0.189510145

need to evaluate the correlations between them to know how different they are. Dataset 3 is used. The KALN of each time

point of each event is ranked by the three strategies. And then the average of all correlations of the three rankings at all the time points is computed. As Table III shown, the correlations from different sources are relatively small, which means three strategies have weak correlation. This result indicates that the three strategies have their own meanings and grasp different aspects of a web event.

D. Comparing Three Strategies in the Semantic Uncertainty of Web Events

Different strategies can form different Semantic Pyramids. In order to compare the performances of three strategies, an

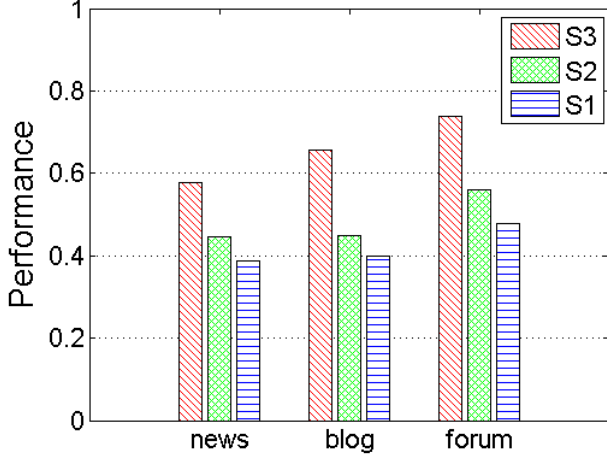


Fig. 10: Performance comparisons of three strategies on different sources, including news, blog and forum. S1, S2 and S3 stand for Strategy I, Strategy II and Strategy III, respectively. In all the sources (news, blog and forum), the Strategy III out weighs the other two strategies. (web event number: 50, average length: 30 days, webpage number: 202,673)

evaluation metric is introduced according to the definitions of three layers KALN and their uncertainty properties.

Definition 8 (Performance of Strategy, ℓ): The main idea is that if the semantic pyramid of KALN is formed, the change of upper layer KALN should be smaller than the layer KALN below and between two consecutive time points. The bigger difference between changes of each pair of layers, the better the corresponding strategy. The strategy, which can best maintain the corresponding property of each layer of the semantic pyramid, is the best one. This metric can be formalized as,

$$\ell = \sum_{level_i < level_j} \left(\frac{\frac{1}{|S_T|} \sum_{T_i \in S_T} \frac{1}{|S_{T,t}|} \sum_{t_i \in S_{T,t}} C_{t_{i-1}, t_i}^{level_i}}{\frac{1}{|S_T|} \sum_{T_i \in S_T} \frac{1}{|S_{T,t}|} \sum_{t_i \in S_{T,t}} C_{t_{i-1}, t_i}^{level_j}} \right) \quad (14)$$

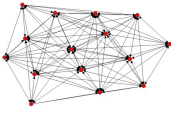
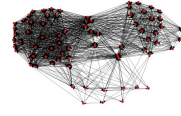
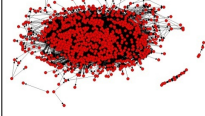
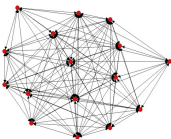
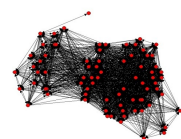
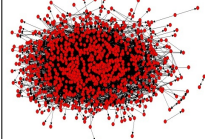
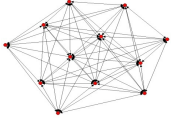
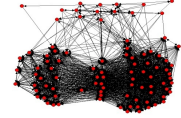
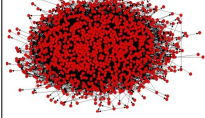
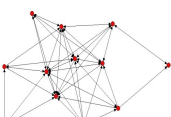
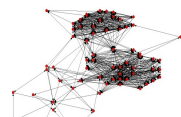
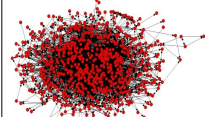
where ℓ is the performance of a strategy, S_T is the set of web events and $S_{T,t}$ is the set of time points of a web event. $C_{t_{i-1}, t_i}^{level_i}$ is the change of $level_i$ KALN of a web event between t_{i-1} and t_i .

The result on Dataset 2 is shown in Fig. 10. The performance of strategy III outweighs the other two strategies in mining the semantic pyramid. This means the final semantic pyramid is more rationally mined by this strategy than others.

E. Case study on the Japan Earthquake in 2011

In order to show the constructed semantic pyramid for web events, certain important days are selected to build the semantic pyramid for comparison with the important things that happened in those days. So, the chronicle of *Japan Earthquake* is listed in Table IV and the hierarchical semantic uncertainty analysis results which are generated by our methods are shown in Table V and Table VI. The keyword numbers of different layers in Table V are consistent with the analysis of Table 2 and the different layer networks of these days are shown

TABLE VI: THE MINED SEMANTIC PYRAMID OF FIVE DAYS CORRESPONDING TO THE CHRONICLE OF LIBYA WAR FORM THE SIMPLIFIED CHINESE WEB ENVIRONMENT

Date	Keywords of Theme LN			Keywords of Backbone LN		Keywords of Tidbit LN	
03.11.2011	Japan	time	influence	Richter	forecast	loss	
	earthquake	occur	Pacific	centre	aftershock	rescue	
	tsunami	Tokyo	atmosphere	cause	appear	arrive	
	revocation	China	maritime	locate	attention	city	
	epicenter	trigger	hypocenter	district	global	houses	
03.15.2011							
	<i>h-KALN</i>						
	tsunami	trigger	disaster	centre	lead to	sets	
	earthquake	time	population	death	seisesthesia	loss	
	nuclearstation	occur	hypocenter	exceed	earth	shock	
	international	China	government	explode	official	expert	
	influence	Japan	missing	cause	houses	public	
03.21.2011							
	<i>h-KALN</i>						
	earthquake	cause	nuclear-radiation	publish	information	report	
	influence	trigger	leak	crisis	material assets	Richter	
	tsunami	Japan	economy	event	radioactivity	nation	
	revocation	disaster	followup	Tokyo	radiation	knowledge	
	China	occur	nuclear-station	loss	claim-indemnity	threaten	
04.07.2011							
	<i>h-KALN</i>						
	earthquake	Japan	investigate	Beijing	development	death	
	tsunami	occur	nuclear-station	alarm	early-warning	pollution	
	webpage	district	archipelago	disaster	seisesthesia	radiation	
	revocation	Tokyo	TV-station	seawater	nuclear-radiation	explode	
	snapshot	China	strong-earthquake	Richter	seacoast	expert	
04.07.2011							
	<i>h-KALN</i>						

in Table VI, which have ignored the isolated nodes. Through these figures, we can directly see the different properties of each layer of the networks. Since space is limited, only 10 keywords of each layer KALN are listed in Table VI to reflect the three level semantics of the web event.

Upper discussion is for certifying that the semantic uncertainty of each layer of the semantic pyramid satisfies our former work. In order to show its usage, a simple demo is given below to show its possible application.

Based on the constructed semantic pyramid, a simple web service demo is built to show its different semantic uncertainty and function to help people to understand a web event. Some screenshots are listed at the end of this paper, in Fig. 11, Fig. 12, Fig. 13 and Fig. 14. The web address is

TABLE IV: THE CHRONICLE OF THE JAPAN EARTHQUAKE (FROM WEBSITES, BLOGS, FORUMS)

Time point (day)	Some important things happened in that day
Mar. 11 2011	A massive 8.9-magnitude quake hit northeast Japan on Friday, causing dozens of deaths, more than 80 fires, and a 10-meter (33-ft) tsunami along parts of the country's coastline; Cars, ships and buildings were swept away by a wall of water after the 8.9-magnitude tremor, which struck about 400km (250 miles) north-east of Tokyo.
Mar.15 2011	Dangerous levels of radiation leak from the Fukushima plant after a third explosion, believed to be in the number 2 reactor, and a fire, rock the complex. In a televised statement after the blast, prime minister Kan urges those within 19 miles of the area to stay indoors.
Mar. 21 2011	Results shared with the IAEA from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan showed that only six out of the 46 samples exhibited any radioactive iodine. However, concentration was reported to be below levels allowed by the Japanese food hygiene law for emergency monitoring criteria for drinking water.
Apr. 07 2011	Japan's northeastern coast, a magnitude 7.1 earthquake occurs, at least two deaths. Tokyo Electric Power Company injects Nitrogen to the No. 1 reactor containment to prevent hydrogen explosion.

TABLE V: THE EXPERIMENT RESULT OF FIVE DAYS CORRESPONDING TO THE CHRONICLE OF LIBYA WAR

Time point (day)	Total Keyword Number	Keyword Number of ThemeLN	Keyword Number of BackboneLN	Keyword Number of TidbitLN
Mar. 11 2011	1197	15	74	1108
Mar. 15 2011	1171	16	81	1074
Mar. 21 2011	989	13	50	926
Apr. 07 2011	848	12	67	769

<http://ic.shu.edu.cn:20/webevent>³. There are two dimensions on the screen. The vertical one is to control the semantic level and the horizontal one is to control the time stamp. The higher level is the most certain part of this web event. The lowest level is the most uncertain part of this web event. The nodes have been selected by humans at each level, because all the nodes cannot be shown in the screen, especially the third level (TidbitLN) in which there are around 1000 nodes. Furthermore, there is no need to exhibit all the keywords, because they will only confuse people rather than enlighten them. So, at the second and third levels, we select a limited number of nodes from the corresponding layer KALN to show the semantic uncertainty of those levels and to assist people to understand the semantic uncertainty of a web event. The analyzed web event in this web service is *Japan Earthquake*, from Mar. 9, 2011 to Mar. 29, 2011. With the help of this demo, people can form a general and hierarchical to understand about this web event by altering the date and the level of network. The automatically mined semantic pyramid is the skeleton of this web event at a given time, which traditionally can only be obtained by reading and summarizing all the webpages at this time. For the people who just want to know the general information of this web event, the ThemeLN's semantics are enough. If they want to know more about it, the BackboneLN's semantics will be more appropriate. If someone cares about the detail of this event, they can go to the TidbitLN's semantics which contains more detailed semantics about it. From this demo, we can see that the certain part of a web event will not change drastically over time and vice versa.

VII. CONCLUSION AND FUTURE WORK

A web event has different levels of semantic uncertainty. If we know about these levels, we can provide different levels

³Please make sure the web browser can access on port 20 (IE web browser is recommended).

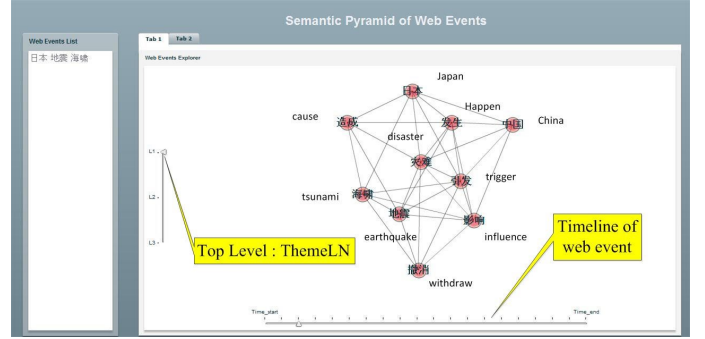


Fig. 11: The mined Semantic Pyramid of web event on ThemeLN (event: Japan earthquake). This level is the most stable one and will not change much with time (slider bar of time at the bottom of figure).

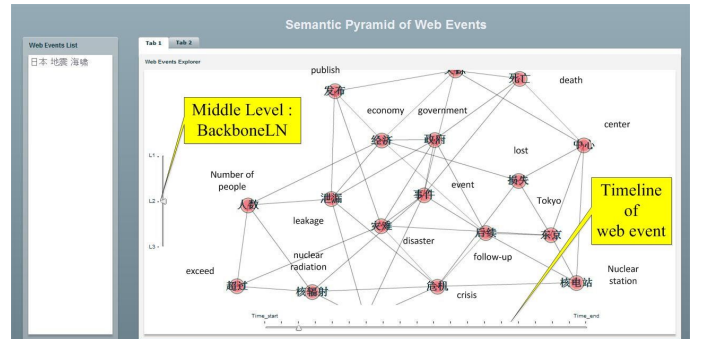


Fig. 12: The mined Semantic Pyramid of web event on BackboneLN (event: Japan earthquake). This level is the medium one.

of information to people with different requirements. In this paper, we have proposed a content-based web event representation (KALN) for preserving the semantics of web events as much as possible. As opposed to the traditional representation methods, the KALN has considered not only the keywords of web events, but also the more important association relations

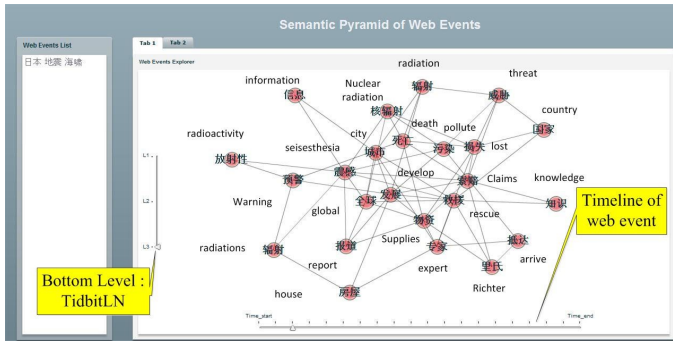


Fig. 13: The mined Semantic Pyramid of web event on TidbitLN (event: Japan earthquake). This level is the most unstable one and will change much with time (slider bar of time at the bottom of figure).

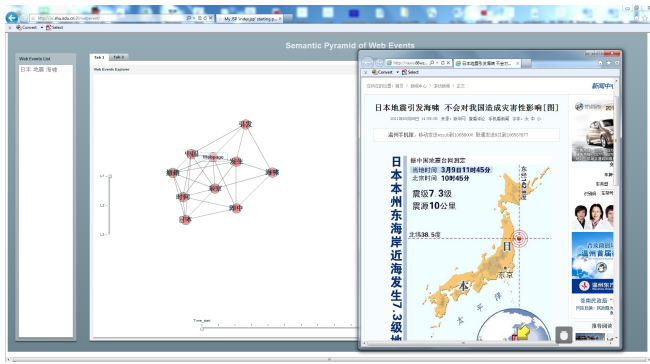


Fig. 14: Semantic Pyramid-based webpage recommendation example. If users click on the WEBPAGE label, the corresponding webpage will be given.

between them, which can preserve more of the semantics of web events. We have also proposed three strategies, including the volume property (Document Frequency), local structural information (Node Degree) and global structural information (PLC), to identify the different levels of semantic uncertainty. We have found that the strategy that considers both the Document Frequency of a keyword and the global network structure of KALN has the best ability to identify the semantic uncertainty levels. Experimental results show that the identified different levels of the semantics display different behaviors over time, so the mined Semantic Pyramid can well exhibit the different level semantic uncertainty of web events. Finally, the demo shows the possible usage of our work.

There are several interesting research points for further study based on our work. First, the dynamics between two consecutive time stamps can be measured through complex network metrics. Second, the patterns of different web events may be different, and these can be mined based on our existing work. Finally, challenging prediction work can be done. Through semantic analyzing and tracking a web event, the maximum possible status of this event can be forecast.

ACKNOWLEDGMENT

Research work reported in this paper was partly supported by the National Science Foundation of China under grant nos. 91024012, and 61071110, and the Shanghai Leading

Academic Discipline Project under grant no. J50103, and by the Australian Research Council (ARC) under discovery grant DP110103733 and the China Scholarship Council.

REFERENCES

- [1] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2003, pp. 330–337.
- [2] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar, "New event detection and topic tracking in turkish," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 4, pp. 802–819, 2010.
- [3] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1019–1028.
- [4] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 297–304.
- [5] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 497–506.
- [6] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Sequential modeling of topic dynamics with multiple timescales," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 4, p. 19, 2012.
- [7] Q. Mei and C. Zhai, "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*. ACM, 2005, pp. 198–207.
- [8] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 811–816.
- [9] M. G. Morgan and M. Small, *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, 1992.
- [10] K. Hayes, "Uncertainty and uncertainty analysis methods," *Australian Centre of Excellence for Risk Assessment (ACERA) project A*, vol. 705, 2011.
- [11] F. Gillund, K. A. Kjølberg, M. Krayen von Krauss, and A. I. Myhr, "Do uncertainty analyses reveal uncertainties? using the introduction of dna vaccines to aquaculture as a case study," *Science of the total environment*, vol. 407, no. 1, pp. 185–196, 2008.
- [12] H. M. Regan, M. Colyvan, and M. A. Burgman, "A taxonomy and treatment of uncertainty for ecology and conservation biology," *Ecological applications*, vol. 12, no. 2, pp. 618–628, 2002.
- [13] A. M. Finkel, *Confronting Uncertainty in Risk Management: A Guide for Decision-makers: a Report*. Center for Risk Management, Resources for the Future, 1990.
- [14] J. S. Clark, "Uncertainty and variability in demography and population growth: a hierarchical approach," *Ecology*, vol. 84, no. 6, pp. 1370–1381, 2003.
- [15] J.-T. Kuo, B.-C. Yen, Y.-C. Hsu, and H.-F. Lin, "Risk analysis for dam overtopping of Feitsui reservoir as a case study," *Journal of Hydraulic Engineering*, vol. 133, no. 8, pp. 955–963, 2007.
- [16] P.-S. Yu, T.-C. Yang, and S.-J. Chen, "Comparison of uncertainty analysis methods for a distributed rainfall-runoff model," *Journal of Hydrology*, vol. 244, no. 1, pp. 43–59, 2001.
- [17] M. Burgman, *Risks and decisions for conservation and environmental management*. Cambridge University Press, 2005.
- [18] M. Benzi, M. Benzi, and E. Seneta, "Francesco paolo cantelli," *International Statistical Review*, vol. 75, no. 2, pp. 127–130, 2007.
- [19] J. Pearl, "Bayesian networks," *Department of Statistics, UCLA*, 2011.
- [20] G. R. Hosack, K. R. Hayes, and J. M. Dambacher, "Assessing model structure uncertainty through an analysis of system feedback and bayesian networks," *Ecological Applications*, vol. 18, no. 4, pp. 1070–1082, 2008.
- [21] N. Çağman and S. Karataş, "Intuitionistic fuzzy soft set theory and its decision making," *Journal of Intelligent and Fuzzy Systems*, vol. 24, no. 4, pp. 829–836, 2013.

- [22] E. I. Papageorgiou and J. L. Salmeron, "A review of fuzzy cognitive maps research during the last decade," *Fuzzy Systems, IEEE Transactions on*, vol. 21, no. 1, pp. 66–79, 2013.
- [23] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.
- [24] R. Forsati and M. R. Meybodi, "Effective page recommendation algorithms based on distributed learning automata and weighted association rules," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1316–1330, 2010.
- [25] F. Khalil, J. Li, and H. Wang, "Integrating recommendation models for improved web page prediction accuracy," in *Proceedings of the thirty-first Australasian conference on Computer science-Volume 74*. Australian Computer Society, Inc., 2008, pp. 91–100.
- [26] C. Wang, J. Lu, and G. Zhang, "Mining key information of web pages: A method and its application," *Expert Systems with Applications*, vol. 33, no. 2, pp. 425 – 433, 2007.
- [27] J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.
- [28] T. Nguyen, H. Lu, and J. Lu, "Web-page recommendation based on web usage and domain knowledge," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 10, pp. 2574–2587, Oct 2014.
- [29] C. Wang, J. Lu, and G. Zhang, "Integration of ontology data through learning instance matching," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, Dec 2006, pp. 536–539.
- [30] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [31] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.
- [32] J. Xuan, X. Luo, S. Zhang, Z. Xu, H. Liu, and F. Ye, "Building hierarchical keyword level association link networks for web events semantic analysis," in *IEEE 9th International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 2011, pp. 987–994.
- [33] G. Bianconi, "Most probable degree distribution at fixed structural entropy," *Pramana*, vol. 70, no. 6, pp. 1135–1142, 2008. [Online]. Available: <http://dx.doi.org/10.1007/s12043-008-0118-9>
- [34] A. Barabasi, *Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades*. Penguin Group US, 2010. [Online]. Available: <http://books.google.com.au/books?id=ZfqVp2cMODkC>
- [35] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.



Xiangfeng Luo is a professor in the School of Computers, Shanghai University, China. Currently, he is a visiting professor in Purdue University. He received the master's and PhD degrees from the Hefei University of Technology in 2000 and 2003, respectively. He was a postdoctoral researcher with the China Knowledge Grid Research Group, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), from 2003 to 2005. His main research interests include Web Wisdom, Cognitive Informatics, and Text Understanding. He has authored or co-authored more than 50 publications and his publications have appeared in IEEE Trans. on Automation Science and Engineering, IEEE Trans. on Systems, Man, and Cybernetics-Part C, IEEE Trans. on Learning Technology, Concurrency and Computation: Practice and Experience, and New Generation Computing, etc. He has served as the Guest Editor of ACM Transactions on Intelligent Systems and Technology. Dr. Luo has also served on the committees of a number of conferences/workshops, including Program Co-chair of ICWL 2010 (Shanghai), WISM 2012 (Chengdu), CTUW2011 (Sydney) and more than 40 PC members of conferences and workshops.



Guangquan Zhang is an associate professor in Faculty of Engineering and Information Technology at the University of Technology Sydney (UTS), Australia. He has a PhD in Applied Mathematics from Curtin University of Technology, Australia. He was with the Department of Mathematics, Hebei University, China, from 1979 to 1997, as a Lecturer, Associate Professor and Professor. His main research interests lie in the area of multi-objective, bilevel and group decision making, decision support system tools, fuzzy measure, fuzzy optimization and uncertain information processing. He has published four monographs, four reference books and over 200 papers in refereed journals and conference proceedings and book chapters. He has won four Australian Research Council (ARC) discovery grants and many other research grants.



Jie Lu is a full professor and Head of School of Software at the University of Technology, Sydney. Her research interests lie in the area of decision support systems and uncertain information processing. She has published five research books and 270 papers, won five Australian Research Council discovery grants and 10 other grants. She received a University Research Excellent Medal in 2010. She serves as Editor-In-Chief for Knowledge-Based Systems (Elsevier), Editor-In-Chief for International Journal of Computational Intelligence Systems (Atlantis), editor for book series on Intelligent Information Systems (World Scientific) and guest editor of six special issues for international journals, as well as delivered six keynote speeches at international conferences.



Junyu Xuan received the bachelor's degree in 2008 from China University of Geosciences, Beijing. Currently, he is working toward the dual-doctoral degree both in Shanghai University, China and University of Technology, Sydney (UTS), Australia. His main research interests include Machine Learning, Complex Network and Web Mining.



Zheng Xu was born in Shanghai, China. He received the Diploma and Ph.D. degrees from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2007 and 2012, respectively. He is currently working in the third research institute of ministry of public security and the postdoctoral in Tsinghua University, China. His current research interests include topic detection and tracking, semantic Web and Web mining. He has authored or co-authored more than 60 publications