Information Sciences 420 (2017) 263–277

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Explicitly and implicitly exploiting the hierarchical structure for mining website interests on news events

Junyu Xuan[a,b], Xiangfeng Luo[a], Jie Lu[b,*], Guangquan Zhang[b]

[a] School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai, China
[b] Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Sydney, Australia

ARTICLE INFO

Article history:
Received 28 January 2016
Revised 14 June 2017
Accepted 15 August 2017
Available online 24 August 2017

Keywords:
Text mining
Web mining
News event
Website interest

ABSTRACT

After a news event, many different websites publish coverage of that event, each expressing their own unique commentary, perspectives, and viewpoints. Websites form around a specific set of interests to cater to different audiences, and discovering these interests can help audiences *C* especially people and organizations that are interested in news *C* select the most appropriate websites to use as their sources of information. This paper presents three methods for formally defining and mining a websites interests, each of which is explicitly or implicitly based on a hierarchial structure: website-webpage-keyword. The first, and most straightforward, method explicitly uses keyword-layer network communities and the mapping relations between websites and keywords. The second method expands upon the first method with an iterative algorithm that combines both the mapping relations and the network relations from the website-webpage-keyword structure to further refine the keyword-layer network communities. In the third method, a website topic model implicitly captures the mapping relations among the websites, webpages, and keywords. The performance of three proposed methods in website interest mining is compared using a bespoke evaluation metric. The experimental results show that the iterative procedure designed in the second method is able to improve website interest mining performance, and the website topic model in the third method achieves the best performance among the three methods.

1. Introduction

News events [20,35] that attract a great deal of attention by the public (e.g., a terrorist attack or a scandal of a famous star) are typically reported by numerous websites. *Phone-hacking of News of the World* and *9–11* are two examples. News events are composed of sub-topics. Take *9–11* as an example. At any given time, the coverage of this event focused on different sub-topics: *The influence to the US economy, American foreign policy, the history of World Trade Center*, and so on. Collectively, it is these various sub-topics that constitute what we have come to know the *9–11* event. Although sub-topics focus on different things, and are relatively independent with each others, they are often mixed together within the webpages of a news event. For example, a webpage reporting news of *9–11* event which mainly talks about *The influence to the economy of U.S.* may also contain some information about *The history of World Trade Center*. In another word, a webpage of

* Corresponding author.
E-mail addresses: xuanjunyu@shu.edu.cn (J. Xuan), luoxf@shu.edu.cn (X. Luo), jie.Lu@uts.edu.au (J. Lu), guangquan.zhang@uts.edu.au (G. Zhang).

http://dx.doi.org/10.1016/j.ins.2017.08.056
0020-0255/

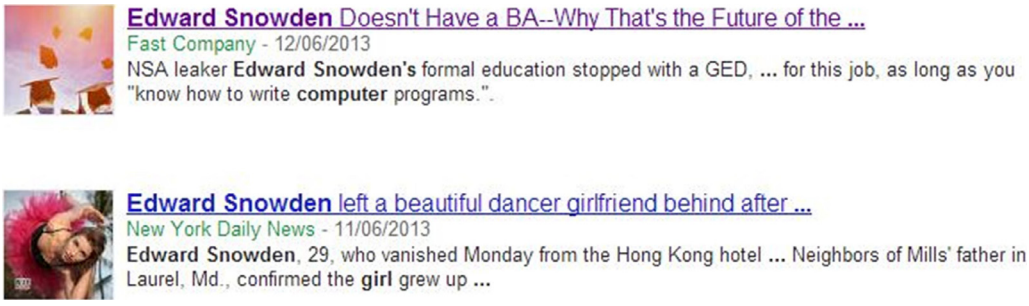

Fig. 1. Two webpages about news event *Edward Snowden Leakage* from Google search engine. Two webpages express the interests of two websites on this news event: one is the *education* of Snowden; the other is his *girlfriend*.

a news event is a mixture of its sub-topics. Similarly, a website is also a (weighted) mixture of sub-topics, and, collectively, these sub-topics express the website's interests for this news event. An illustrative example of a website's interest is shown in Fig. 1 using two pieces of information about the *Snowden PRISM* news event. The *Fast Company* website is concerned about *Snowden's education*, but the *New York Daily News* website focuses on *his girlfriend*. Websites coalesce around interests mainly to cater to different audiences. The *phone-hacking of News of the World* is another example; www2.canada.com[1] mainly focused on *Impact on ethics of internet journalism* of this news event, whereas www.guardian.co.uk[2], centered their coverage on the *Criminal charges and convictions* associated with this event.

The ability to identify a website's interests relating to news events is of practical significance: 1) **website recommendation:** Instead of visiting frequencies, website interests could be used as a criterion for website recommendations to help users and organizations easily find topics of interest for news events; 2) **precision advertising:** Since news events attract a great deal of public attention, targeted advertising on carefully selected websites with correlating interests could help companies increase sales; 3) **malicious websites detection:** If libellous information about a news event is being disseminated through one particular website, or a website generally only reports libellous information, it is possible the website may be malicious. The vast number of websites[3] makes the task of manually identifying website interests, and their intentions, impossible. However, the ability to automatically infer this information would assist with malicious website detection.

This paper presents three novel methods to automatically mine website interests based on the website-webpage-keyword hierarchical structure. Beyond the mapping relationships in this hierarchical structure, the association relations between keywords offer further significant information [24], which are hidden but can be mined from webpages. These association relations could have an impact on subtopic discovery and, as a consequence, on website interest mining. For example, if the terms *terrorist* and *attack* often co-occur, there is a high probability they relate to the same subtopic. Depending on whether this association relation is explicit or not, we propose three methods: two explicit and one implicit. The first method is straightforward, relying on the association relations in complex networks. All the keywords are linked together according to their association relations to form a keyword network. For a given news event, the keyword communities in this keyword network are considered to the subtopics of the news event. The website interests are mined with the help of mapping relations between the websites, keywords, and the discovered keyword communities. The second method builds on the first method and incorporates the mapping relations between the websites, webpages, and keywords to constrain the formation of keyword communities from the keyword network. This constraint is implemented through an iterative procedure. Its efficiency in improving website interest mining is verified in the experiments presented in Section 7. Unlike the above methods, which rely on explicit association relations, the third method uses the same information by building a website topic model from the website-webpage-keyword structures. A subtopic is considered to be a keyword distribution, and a website interest is considered to be a subtopic distribution. The performance of three proposed methods in website interest mining were evaluated and compared using real-world data.

The remainder of this paper is organized as follows. The problem and our basic idea are introduced in the next section. In Section 3, we review some related work. The three proposed methods are presented in Section 4, 5 and 6. In Section 7, the bespoke evaluation metric is described and the performance of three methods in website interest mining are quantitatively compared using real-world data. Section 8 concludes this study and discuses possibilities for further research.

2. Problem definition and basic idea

In this section, we first formally define our problem. We then introduce the basis for resolving this problem: website-webpage-keyword hierarchy.

[1] http://www2.canada.com/story.html?id=5074391.
[2] http://www.guardian.co.uk/media/2005/apr/16/pressandpublishing.crime.
[3] To the December of 2012, there are about 634 million websites on the web http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/.

2.1. Problem definition

Definition 1 (News Event)**.** Consider a news event that is composed of sub-topics,

$$e = \{t_0, t_1, \cdots, t_{T-1}\} \tag{1}$$

where t_i is i-th sub-topic of a news event e and T is the number of sub-topics related to this news event. Each sub-topic has its own relatively independent semantics,

$$t_i = < w_{k_0}, w_{k_0}, \cdots, w_{k_{N-1}} > \tag{2}$$

where k_i is the i-th keyword of a news event, N is the total number of a news event, and w_{k_i} is the weight of keyword k_i in this sub-topic.

Definition 2 (Website Interest)**.** A website interests in a given news event is the distribution of all sub-topics relating to this news event and can be represented as,

$$P_s^e = < w_{t_0}, w_{t_1}, \cdots, w_{t_{T-1}} >, \quad w_{t_i} \in [0, 1] \tag{3}$$

where w_{t_i} is the degree of website s concerning on the sub-topic t_i of a news event e. With the exception of some learge comprehensive websites, a website will normally only focus on limited number of sub-topics.

Taking *9–11* as an example again, a huge number of webpages across a number of websites reported on this news event. And, as discussed in the Introduction, a webpage may cover several subtopics, which means the website may also relate to many subtopics. However, a website generally has its own special interests, so a website specializing in economic issues has a higher probability of reporting on the *economic impacts of 9–11* and, therefore, the weight of the *economic impact* subtopic is likely to be greater than the weight of the other subtopics. Hence, the distribution of weights across different subtopics is simply a reflection of the website's specific interests in news events.

The final goal of this paper is to mine a website interests for a given news event.

2.2. Website-webpage-keyword hierarchy

In this paper, the information about a news event is organized as a three-layered network, comprising a website layer, a webpage layer and a keyword layer. Initially, keywords are linked by their association relations,

$$r^K_{k_i,k_j} = \frac{P_{k_i,k_j}}{P} \tag{4}$$

where r_{k_i,k_j} is the association relation between the keywords k_i and k_j; P_{k_i,k_j} is the number of webpages containing the keywords k_i and k_j; P is the total number of webpages. An association relation [37] is a type of weak semantic relation between keywords, which implies the possibility of two keywords both appearing within same webpages. Connecting all the keywords by their relations generates the keyword layer network shown in Fig. 2(c), ALN_K,

$$ALN_K = \begin{bmatrix} r^K_{0,0} & r^K_{0,1} & r^K_{0,2} & \cdots & r^K_{0,K-1} \\ r^K_{1,0} & r^K_{1,1} & r^K_{1,2} & \cdots & r^K_{1,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^K_{K-1,0} & r^K_{K-1,1} & r^K_{K-1,2} & \cdots & r^K_{K-1,K-1} \end{bmatrix} \tag{5}$$

Based on the keyword layer network and the mapping relations between keywords and webpages, the relations, $r^D_{d_i,d_j}$, between the webpages can be constructed [24],

$$r^D_{d_i,d_j} = \sum_{k_m \in d_i, k_n \in d_j} r^K_{k_m,k_n} \tag{6}$$

where k_m and k_n are the keywords in webpage d_i and d_j, respectively. Then, the webpage layer network shown in Fig. 2(b) is, ALN_D,

$$ALN_D = \begin{bmatrix} r^D_{0,0} & r^D_{0,1} & r^D_{0,2} & \cdots & r^D_{0,D-1} \\ r^D_{1,0} & r^D_{1,1} & r^D_{1,2} & \cdots & r^D_{1,D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^D_{D-1,0} & r^D_{D-1,1} & r^D_{D-1,2} & \cdots & r^D_{D-1,D-1} \end{bmatrix} \tag{7}$$

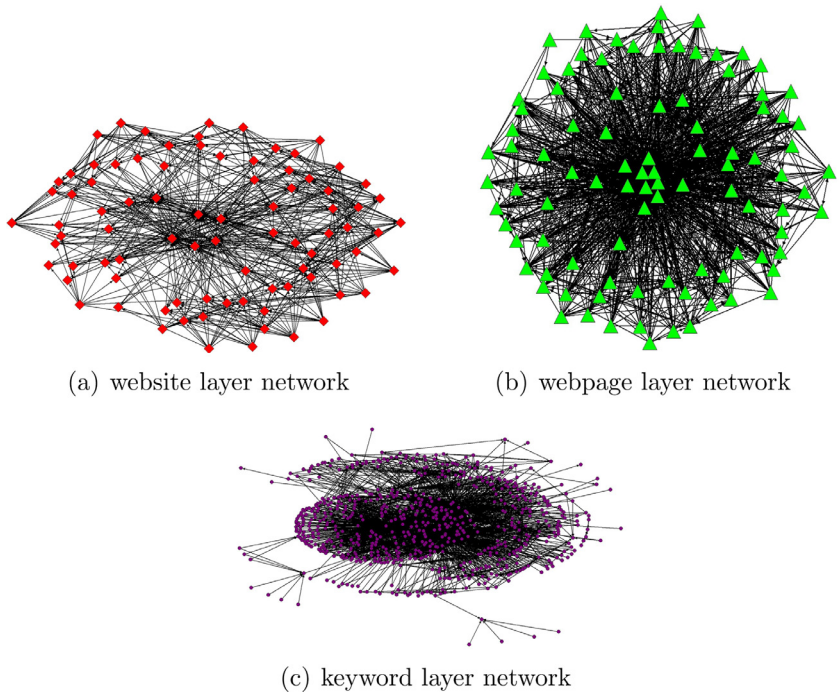

Fig. 2. An example of a three-layered network (representing the *Japan Earthquake* news event).

Similar to the webpages, the relations between the websites are computed as,

$$r^S_{s_i,s_j} = \sum_{d_m \in s_i, d_n \in s_j} r^D_{d_m,d_n} \tag{8}$$

where d_m and d_n are the webpages in websites s_i and s_j, respectively. Then, the website layer network shown in Fig. 2(a) is, ALN_S,

$$ALN_S = \begin{bmatrix} r^S_{0,0} & r^S_{0,1} & r^S_{0,2} & \cdots & r^S_{0,S-1} \\ r^S_{1,0} & r^S_{1,1} & r^S_{1,2} & \cdots & r^S_{1,S-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^S_{S-1,0} & r^S_{S-1,1} & r^S_{S-1,2} & \cdots & r^S_{S-1,S-1} \end{bmatrix} \tag{9}$$

To filter out the noisy relations in each layer of the network, a pruning method [24] is used to post-process all three layers. Only the spanning tree and some of its neighboring nodes are kept. An example is shown in Fig. 2.

2.3. Basic idea

Our idea to resolve the former problem is to use the three-layered network, comprising a website layer, a webpage layer and a keyword layer (from top to bottom), respectively. As shown in Fig. 3, a website may contain many webpages, and a webpage may contain many keywords relating to a news event. At the same time, a keyword can appear in many webpages. And, moreover, a webpage could be published by many websites in this setting because multiple webpages with extremely similar content have been merged. The reason for the merger is that similar webpages ofen result when websites reprint or forward content.

There are interdependent relations between the websites, webpages and keywords of a news event. The interest of a website for a given news event needs to be expressed by its webpages, and the content covered by a webpage about a news event need to be expressed by its keywords. In turn, the existence of keywords relies on the existence of webpages, and the existence of webpages relies on the existence of websites. So, three layered networks need to be considered in their entirety. In the next sections, different strategies are introduced that utilize these three layered networks to mine the interests appearing in news events on websites.

Some frequently used symbols in this paper are listed in Table 1.

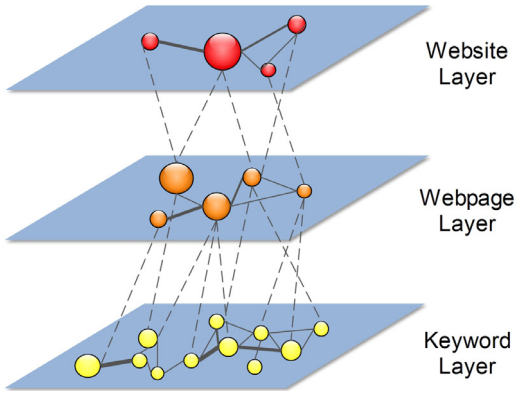

Fig. 3. A three layered network for a news event e, comprising the website, webpage, and keyword layers. The red nodes in the top layer represent websites s, the orange nodes in the middle layer represents webpages p, and the yellow nodes in the bottom layer represent keywords k. A website may contain many webpages and a webpage may contain many keywords. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Notations used in this paper.

Symbol	Description
e	a news event
s	a website
S	the number of websites of a news event
d	a webpage
D	the number of webpages of a news event
k	a keyword
K	the number of keywords of a news event
P_s	interest of a website s

3. Related works

In this section, we review related works on three parts. The first part is about the existing researches on websites; The second and third parts are about community detection and topic model which are basic tools used in this paper.

3.1. Websites

The first part of researches on website is the website evaluation. Shanshan Qi [30] suggests that a website should be evaluated from three aspects: usefulness, service quality and physical accessibility. The qualities of content and structure of websites will impact on their usage interests which means the efficiency of using these websites. And the content is more important than structure in the long run [8]. The content and structure of website is evaluated to fit better the needs of visitors by reorganizing the documents [29].

The second part of researches on website is to generate website overview.A recommender system, Pharos, is proposed to help new users understand the contents of the website by giving a overview of the whole content-centered website [39]. For contextual advertising, website hierarchies are learned by URL [13]. Different from keywords [13], the key-phrases [22] are extracted to label the website topic hierarchy [21] for providing a site-map to users.

Last part of researches on website is to compute similarity between websites. Different from SiteRank [1] and AggregateRank [10], a new content sensitive method, STRank [15], is proposed to consider the semantic and time relevance of websites rather than considering link property. Similarly, Pablo N. Mendes [25] measures the website similarity by connect the query logs with entities. In this way, semantic relatedness is added to the similarity computation rather than only considering keyword match.

3.2. Community detection

Recently, complex network is of significance to model the complex system which has been used in many fields, including biological area, physical area, information area, and so on. Community is a very important feature of complex network. A community is a subgraph of a network and a network can be seen as the union of different communities in turn. The basic principle to detect communities is to minimizing the number of links between different communities and maximizing the number of links in single community. Two classical strategies are: Aggregation and Division [12]. There are also some

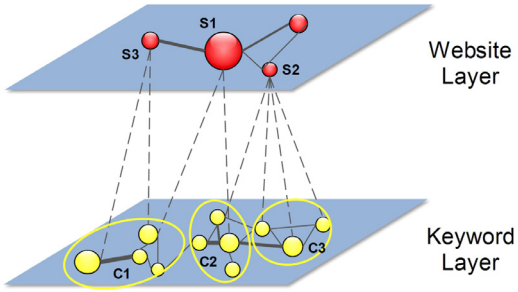

Fig. 4. Intuitive method to discover the interests of websites. In this figure, there are three sample websites and related three communities in keyword layer network. It is the keywords that connect the communities with websites.

other methods, including spectral-based method [26], statistical method [38]. Some works [11,18,19] can be read For more information on community detection,

3.3. Topic model

First topic model is probabilistic Latent Semantic Indexing (pLSI) [16], which is a probability extension of Latent Semantic Indexing (LSI) [9]. The original idea of them is from the sparse document-keyword matrix. LSI uses Singular Value Decomposition (SVD) from the dimension reduction view and pLSI builds a generative model to find the latent classes (topics). However, there is a over-fitting problem in the pLSI model, which is addressed by Latent Dirichlet Allocation (LDA) [6] using a Dirichlet prior for all the topic distributions of documents. There are also many extensions of LDA which have considered different aspects of documents. For example, the labels of document [5], time of documents [34], authors of documents [33], emotions of documents [3,32], and so on. There are also some works trying to release the independent of documents and discovered topics by considering the citation relations between documents [7,27], relations of words [36], and relations of topics [4]. However, all these works are still based on 'bag-of-words' assumption and the relations of keywords within documents are ignored. Some researchers were ware of this gap. Thomas Griffiths [14] tries to fill this gap by add syntactic relations of words in a sentence to the model. Hidden Markov Model (HMM) [2] is combined with topic model by assuming that the keywords in a document are generated under a inherent linguistic sequence.

4. Intuitive method

In this paper, the communities of keyword layer network are adopted as sub-topics of a given news event. Since each keyword is a semantic atom of a news event, the community composed of a number of keywords which have relatively close relations with each others can be seen as a sub-topic of a news event.

After giving the definitions in Section 2, the most straightforward method to obtain the interests of websites would be to detect the communities of keyword layer network and then these detected communities could be seen as the different sub-topics of a news event. The interests of websites could be computed as the membership degree on each community. The procedure of this method is as follows,

1. Construct keyword network of a number of webpages published on a number of websites;
2. Do community detection on this keyword network;
3. Compute the membership degree of each website on the detected communities.

An example is given in Fig. 4. There are three communities in keyword layer network of a given news event. Website S_1 has relations with two communities C_1 and C_2, website S_2 has relations with two communities C_2 and C_3, and website S_3 has a relation with community C_1.

$$\begin{aligned} P_{S_1} &=< w_{C_1}, w_{C_2} > \\ P_{S_2} &=< w_{C_2}, w_{C_3} > \\ P_{S_3} &=< w_{C_1} > \end{aligned} \tag{10}$$

where w_{C_i} is the membership degree of each website on a community/sub-topic. This method has considered the relations between keywords and the website-keyword mapping relation which is summarized in Algorithm 1. However, the relations between keywords, websites with webpages are ignored.

5. Iterative method

The communities of keyword layer network are only based on the keyword relations between each other, a horizontal relation in Fig. 3. This relation implies that the keywords, which have close association relations with each others, will be

Algorithm 1: Intuitive method website interests mining.

Input: Keyword network ALN_K and websites-keywords map $M_{s,k}$
Output: The community of three layer networks: $\{C_i^K\}$, $\{C_i^D\}$ and $\{C_i^S\}$
set $\{C_i^K\} = communitydetection(ALN_K)$;
return $\{C_i^K\}$, $\{C_i^D\}$ and $\{C_i^S\}$;

more likely to describe same sub-topic of a news event. Actually, the webpage layer network will also influence the formation of communities at keyword layer. When the keywords are in the same webpage by the mapping relations between keywords and webpages, it is also possible that they are talking about same sub-topic of a news event. However, the relations in the keyword layer network, ALN_K, do not take the mapping relations into consideration, which only consider the statistical values of association relations on all the webpages. For example, two keywords, k_i and k_j, have a small association relation which means that they do not frequently show in the webpages simultaneously. However, if two webpages which contain keywords, k_i and k_j, respectively and they are in the same community of webpage layer network, keywords, k_i and k_j, have also a large probability to talk about same sub-topic of a news event. Similarly, the communities of webpage layer are also influenced by the mapping relations between webpages and websites. Inspired by their inter-dependency and inter-limitation relations of websites, webpages and keywords, an iterative algorithm is proposed to optimize the formation of keyword communities/sub-topics, as shown in Algorithm 2.

Algorithm 2: Iterative method website interests mining.

Input: Three layer networks: ALN_K, ALN_D and ALN_S
Output: The communities of three layer networks: $\{C_i^K\}$, $\{C_i^D\}$ and $\{C_i^S\}$
set $cong = max\_value$; $cong^{new} = max\_value - 1$;
set $\{tC_i^K\} = \{C_i^K\} = communitydetection(ALN_K)$;
while $(cong - cong^{new}) > 0$ **do**
  $cong = cong^{new}$;
  for $s_i = 0; s_i \le S-1$ **do**
    for $s_j = 0; s_j \le S-1$ **do**
      $ALN_S^{new}[s_i][s_j] = \cos(\vec{\nu_{s_i}^K}, \vec{\nu_{s_j}^K})$;
    end
  end
  $ALN_S = \rho ALN_S + (1-\rho) ALN_S^{new}$;
  $\{C_i^S\} = communitydetection(ALN_S)$;
  for $d_i = 0; d_i \le D-1$ **do**
    for $d_j = 0; d_j \le D-1$ **do**
      $ALN_D^{new}[d_i][d_j] = \cos(\vec{\nu_{d_i}^S}, \vec{\nu_{d_j}^S})$;
    end
  end
  $ALN_D = \rho ALN_D + (1-\rho) ALN_D^{new}$;
  $\{C_i^D\} = communitydetection(ALN_D)$;
  for $k_i = 0; k_i \le K-1$ **do**
    for $k_j = 0; k_j \le K-1$ **do**
      $ALN_K^{new}[k_i][k_j] = \cos(\vec{\nu_{k_i}^D}, \vec{\nu_{k_j}^D})$;
    end
  end
  $ALN_K = \rho ALN_K + (1-\rho) ALN_K^{new}$;
  $\{C_i^K\} = communitydetection(ALN_K)$;
  $cong^{new} = |\{C_i^K\} - \{tC_i^K\}|$;
  $\{tC_i^K\} = \{C_i^K\}$;
end
return $\{C_i^K\}$, $\{C_i^D\}$ and $\{C_i^S\}$;

At first, the communities of keyword layer network, $tC_i^K = C_i^K = \{s\}, s \in ALN_K$, are detected as the initialization. Then, by the mapping relations between keywords with websites, each website is represented as a vector of communities of keyword

layer network, $\vec{v}^{C_S}_{k_i}$,

$$\vec{v}^{C_K}_{s_i} =< \delta_{C^K_0}, \delta_{C^K_1}, \ldots, \delta_{C^K_{nk-1}} > \tag{11}$$

where nk is the number of communities of keyword layer network and $\delta_{C^K_i}$ is indicator function that shows whether this website s_i is in community C^K_i.

The similarity between two websites which are represented as keyword community vectors is computed through the cosine method. This similarity reflects the mapping relations between keywords and websites. Based on these new mapping relations, a new website network, ALN^{new}_S, is constructed and incorporated into ALN_S,

$$ALN_S = (1-\rho)ALN_S + \rho ALN^{new}_S \tag{12}$$

where ρ is the ratio of community-based website network ALN^{new}_S. This combination can revise the communities of this network. It can be seen from Eq. (12) that the communities detected from ALN_S are based on two kinds of relations between websites:

- Association relation
- Mapping relation

Similar to the mapping relation between websites and keywords, there are also mapping relations between websites and webpages. So, the communities of webpage layer network are also influenced by the structure of website layer network. Each webpage can also be represented as a vector of communities of website layer network,

$$\vec{v}^{C_S}_{d_i} =< \delta_{C^S_0}, \delta_{C^S_1}, \ldots, \delta_{C^S_{ns-1}} > \tag{13}$$

where ns is the number of communities of website layer network and $\delta_{C^S_i}$ is indicator function that shows whether this webpage d_i is in community C^S_i. Then, the combination of new webpage network from the similarity of vector representation and original webpage network is,

$$ALN_D = (1-\rho)ALN_D + \rho ALN^{new}_D \tag{14}$$

Similar with websites and webpages, a keyword can also be represented as,

$$\vec{v}^{C_D}_{k_i} =< \delta_{C^D_0}, \delta_{C^D_1}, \ldots, \delta_{C^D_{nk-1}} > \tag{15}$$

where nk is the number of communities of website layer network. Then, the combination of new keyword network from the similarity of vector representation and original keyword network is,

$$ALN_K = (1-\rho)ALN_K + \rho ALN^{new}_K \tag{16}$$

Apparently, this is an iterative procedure. The difference between communities of keyword layer networks of two iterations is computed. When this difference does not reduce, the stopping criteria is reached. The condition can ensure that it reaches the local optimization, which is enough from the experimental results.

Finally, the communities of keyword network are considered as different sub-topics like first strategy: Intuitive Method. The website interests are computed by the mapping between websites and keyword communities. Note that the different values of ρ denote different ways of combination of network structures with mapping relations between them. When $\rho = 0$, Iterative Method (denoted as *nw*) degenerates to Intuitive Method (denoted as *iw*).

6. Website topic model method

The above two methods both explicitly use the website-webpage-keyword structure through the community detection. In this section, we propose a probabilistic model to mine the website interests by implicitly using the website-webpage-keyword structure.

6.1. Model description

As shown in Fig. 3, there is a hierarchical structure between websites, webpages and keywords. In this section, a website topic model (WTM) is proposed to capture this hierarchical structure. Like other topic models [5,6,33], WTM is also a generative model. In this model, different websites have different sub-topic distributions and different sub-topics have different keyword distributions. Each webpage also has its own sub-topic distribution which is impacted by the website it belongs to. The generative process of this model can be described as,

1. draw $\phi_t \sim Dir(\beta)$ for each sub-topic;
2. draw $\theta_s \sim Dir(\alpha)$ for each website;
3. for all webpage of a website:
   (a) draw $\eta_d \sim Dir(\theta_s \cdot \gamma)$;

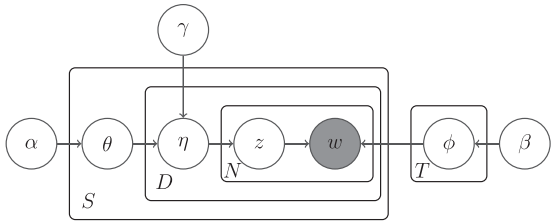

Fig. 5. Graphical models of website topic model.

Table 2
Notations in website topic model.

Symbol	Description
T	the number of sub-topics
D_s	the number of webpages of website s
N_d	a keyword
V	the number of different keywords
θ_s	sub-topic distribution of website s
$\eta_{s,d}$	sub-topic distribution of webpage d of website s
$z_{s,d,n}$	the sub-topic assignment of word n of webpage d of website s
ϕ_t	t-th sub-topic

(b) for all words of a webpage:
(i) draw $z_{s,d,n} \sim Multi(\eta_d)$ for each keyword;
(ii) draw $w_{s,d,n} \sim Multi(\phi_{z_{s,d,n}})$ for each keyword.

In this model, θ_s is the sub-topic distribution of a website s, η_d is the sub-topic distribution of a webpage d, and ϕ_t is the keyword distribution of a sub-topic t. Fig. 5 shows the graphical representation of the model, and the notations are list in Table 2. It can be directly observed from Fig. 5 that there is a (three-layer) hierarchy: the outer layer (i.e., website layer), the middle layer (i.e., webpage layer) and the inner layer (i.e., word layer). Therefore, this model could use the website-webpage-keyword hierarchy. More specially, the sub-topic distribution of webpage η is influenced by the sub-topic distribution of website θ, which is established through $\eta_d \sim Dir(\theta_s \cdot \gamma)$. This kind of dependency is reasonable because the expectation of a webpage is just the sub-topic distribution of website, and at the same time, a variance does exist around the expectation. Apparently, θ_s is just the website interest of website s and ϕ_t describes a sub-topic like the keyword communities in the former two methods.

Although ϕ_t has same meaning with keyword communities in former two methods, there are still differences between these two things. One is that there is no overlap between different communities, but different ϕ_t may have overlap. The other one is that the communities are detected by considering the network structure of keyword layer network or other layer networks which is not considered in WTM.

6.2. Model inference

The posterior distribution of the latent variables in the model is,

$$p(\theta, \eta, z, \phi | \alpha, \beta, \gamma, w) \tag{17}$$

It is difficult to obtain the analytical solution for this high-dimensional and multi-variable distribution, so we need to resort to the Monte Carlo method. In the following, a Gibbs sampling algorithm is designed for the posterior inference and all the conditional distributions are listed.

Sampling θ The prior for this variable is a Dirichlet distribution parameterized by α, but the likelihood for this variable are Dirichlet distributions too. Due to the non-conjugate between two Dirichlet distributions, we cannot obtain a closed-form posterior distribution for this variable,

$$\begin{aligned} p(\theta_s | \cdots) &\propto \prod_d Dir(\eta_d | \theta_s \gamma) \cdot Dir(\theta_s | \alpha) \\ &= \prod_d \left(\frac{\Gamma(\sum_t \theta_s \gamma)}{\prod_t \Gamma(\theta_s \gamma)} \prod \eta_{d,t}^{\theta_{s,t}\gamma - 1} \right) \cdot \prod_t \theta_{s,t}^{\alpha - 1} \end{aligned} \tag{18}$$

Since this conditional distribution is not a standard distribution, Metropolis-Hastings sampling should be adopted to obtain the samples from this distribution.

Sampling η Since the prior for this variable is a Dirichlet distribution and the likelihood is Multinomial distributions, the posterior is still a Dirichlet distribution but with different parameters.

$$\begin{aligned} p(\eta_{s,d}|\cdots) &\propto \prod_{n} Multi(z_{s,d,n}|\eta_{s,d}) \cdot Dir(\eta_{s,d}|\theta_s\gamma) \\ \eta_{s,d}|\cdots &\sim Dir(\theta_{s,1}\gamma + m_{s,d,1}, \theta_{s,2}\gamma + m_{s,d,2}, \cdots, \theta_{s,T}\gamma + m_{s,d,T}) \end{aligned} \tag{19}$$

where $m_{s,\,d,\,t}$ denotes the number of words of webpage d of website s assigned to sub-topic t ($z_{s,d,n} = t$).

Sampling z This variable denotes sub-topic assignment to word n of webpage d of website s, which prior distribution is Multinomial distribution parameterized by the sub-topic distribution $\eta_{s,\,d}$. At the same time, its likelihood is also Multinomial distribution but different parameters.

$$p(z_{s,d,n} = t|\cdots) \propto \eta_{s,d,t} \cdot \phi_{t,w_{s,d,n}} \tag{20}$$

Sampling ϕ This variable is just what the sub-topics are. A sub-topic is a distribution on different words, and in turn, we can see all the sub-topics as the sub-topic distributions of words.

$$\begin{aligned} p(\phi_t|\cdots) &\propto \prod_{n:\{z_{s,d,n}=t\}} Multi(w_{s,d,n}|\phi_t) \cdot Dir(\phi_t|\beta) \\ \phi_t|\cdots &\sim Dir(\beta + m_1, \beta + m_2, \cdots, \beta + m_V) \end{aligned} \tag{21}$$

where m_v is the number of word v in all the webpages.

The whole procedure for the Gibbs sampling is summarized in Algorithm 3 . Each iteration could obtain one sample

Algorithm 3: Gibbs sampling for website topic model.

```
Input: All the webpages of a news event; the sub-topic number K
Output: {θ}, {η}, {z}, {φ}
initialization;
while iter ≤ max_iter do
    for s = 1; s ≤ S do
        Update θ_s by Eq. (18);
        for d = 1; d ≤ D_s do
            Update η_{s,d} by Eq. (19);
            for n = 1; n ≤ N_d do
                Update z_{s,d,n} by Eq. (20);
            end
        end
    end
    Update φ_{1:K} by Eq. (21);
    iter++;
end
```

of the distribution in Eq. (17). After ignoring the first burn-in stage, a number of samples are extracted to compute the expectation of the distribution as the final result.

7. Experiments

In this section, an evaluation metric was designed to facilitate comparisons between three methods based on the nature of the website interest, which is explained first, followed by the performance evaluations of three methods with comparisons using real-world news event data.

7.1. Evaluation metric

To compare the performance of the three methods, an evaluation metric was designed on the assumption that a website's interest for a news event would remain stable as the event unfolded. For example, if a website initially forced on *Economic effect* of *9–11*, there was a high probability they would continue reporting information on *9-11's Economic effect* rather than other sub-topics of the event.

In line with this assumption, the similarity Sim_s between a website's interests at different time stamps of a news event are evaluated to measure their stability of website interest under the chosen method. It appears that the method with largest similarity has the best performance on website interest mining. Similarity is evaluated by

$$Sim_s = \frac{1}{\mathcal{T}_{end} - \mathcal{T}_{start}} \sum_{\tau=\mathcal{T}_{start}+1}^{\mathcal{T}_{end}} \sum_{(i,j)\in T^{\tau-1}\times T^{\tau}} Sim(t_i^{\tau-1}, t_j^{\tau}) \cdot \min\{w_{t_i}^{\tau-1}, w_{t_j}^{\tau}\} \tag{22}$$

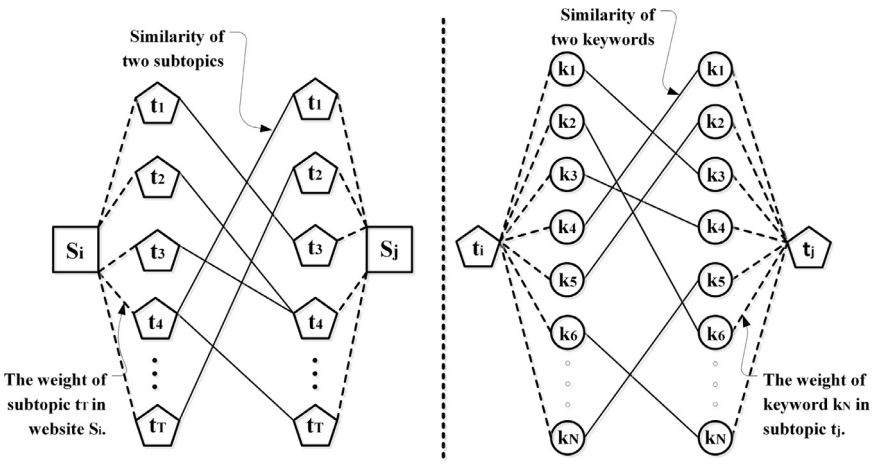

Fig. 6. The evaluation metric used to compare the three methods. The left subfigure shows the similarity calculation between the interests of two websites, S_i and S_j, through their sub-topics. The right subfigure shows the similarity calculation between two sub-topics, t_i and t_j, through their keywords.

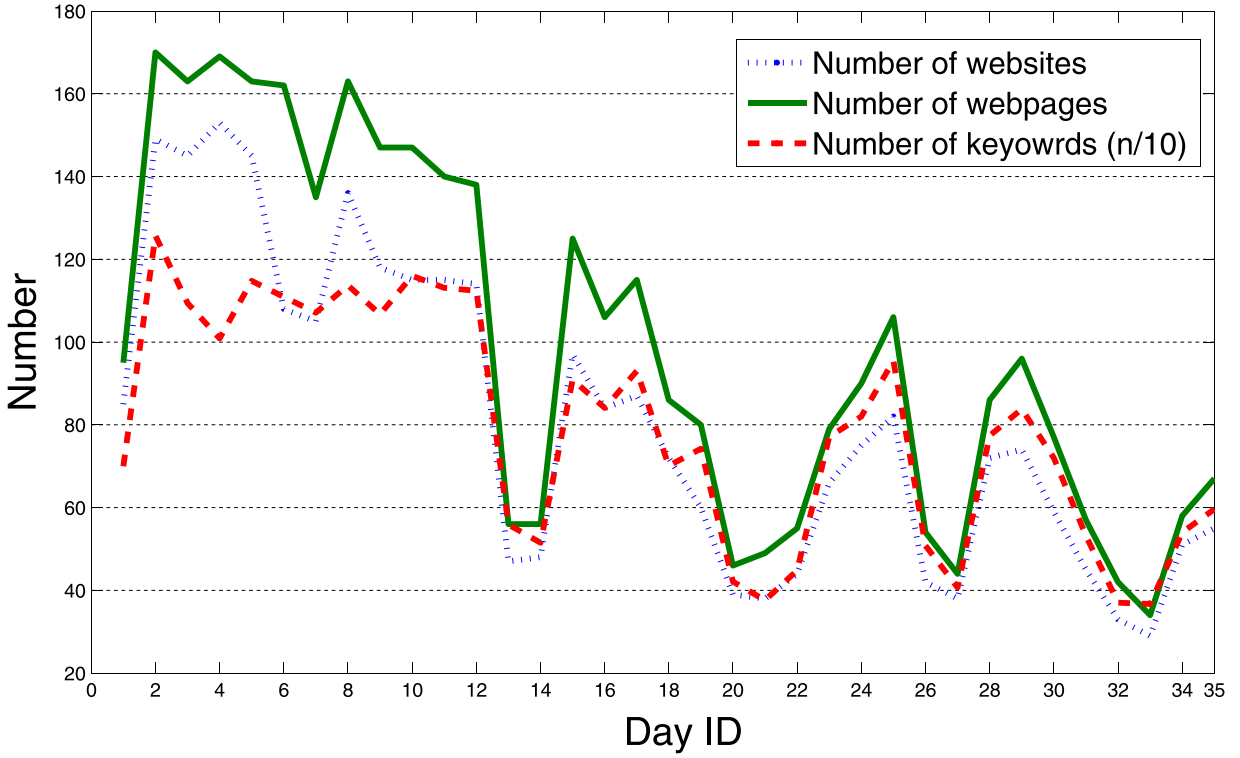

Fig. 7. The number of websites, webpages, and keywords relating to a particular news event per day. Note that the actual number of keywords is ten times the numbers in the figure.

where $\mathcal{T}_{start}$ and $\mathcal{T}_{end}$ are the start and end time stamps of a news event. T^{τ} is the sub-topic number of this news event at time τ and $w^{\tau}_{t_j}$ is the weight of website s on sub-topic t_j at time τ. $Sim(t_i^{\tau-1}, t_j^{\tau})$ is the similarity between two sub-topics of a news event at two different time stamps,

$$Sim(t_i^{\tau-1}, t_j^{\tau}) = \sum_{k_i=k_j, k_i \in N^{\tau-1} \& k_j \in N^{\tau}} \min\{w^{\tau-1}_{k_i}, w^{\tau}_{k_j}\} \tag{23}$$

where N^{τ} is the keyword set at time te and $w^{\tau}_{k_j}$ is the weight of a sub-topic on keyword k_j. The similarity evaluation is also illustrated in Fig. 6.

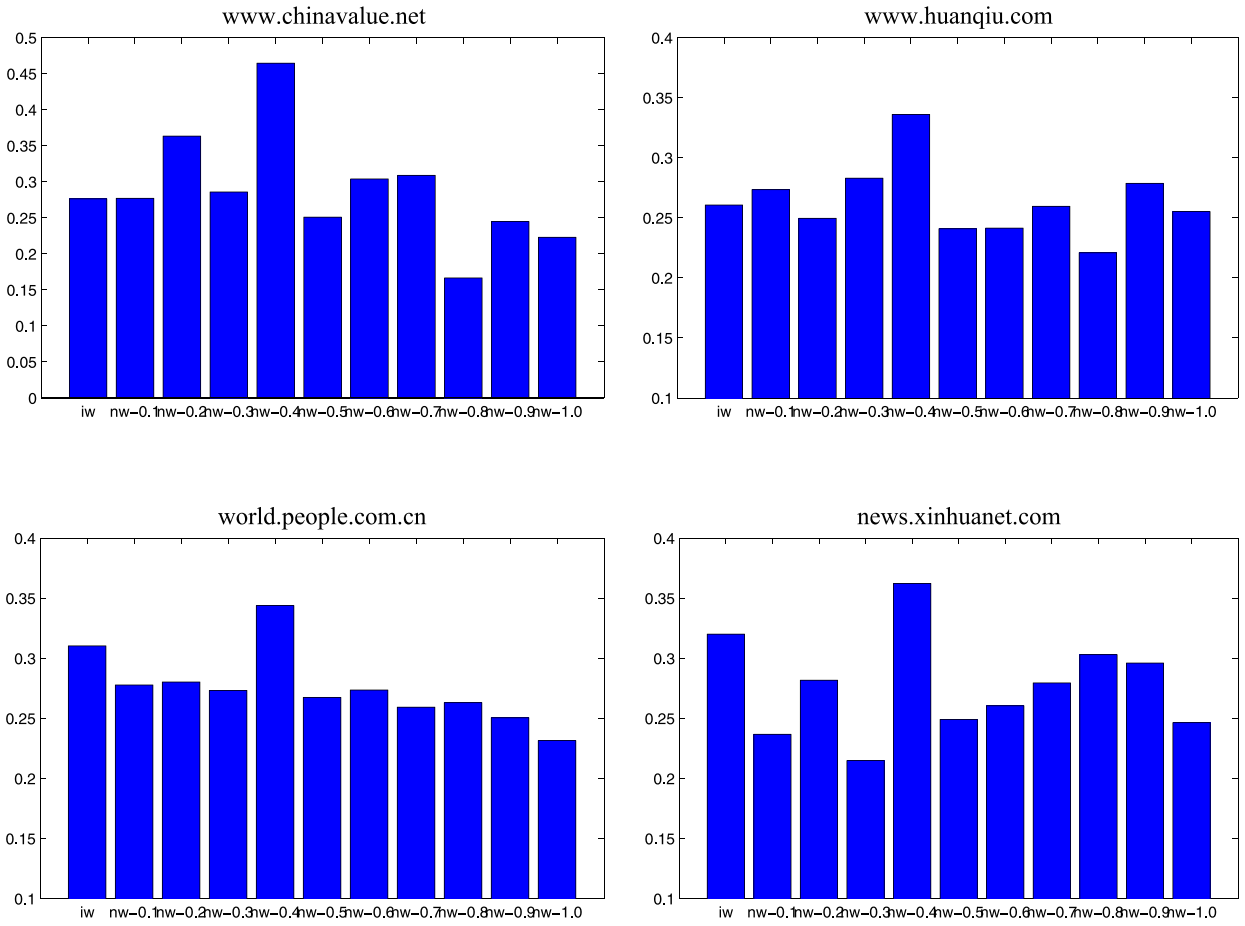

Fig. 8. The Sim_s values of four different websites, i.e., www.chinavalue.net, www.huanquan.com, world.people.com.cn, and news.xinhuanet.com, of news event *Japan Nuclear Leakage* from our proposed methods with different ρ.

7.2. Dataset and setting

The *Japan Nuclear Leakage* showcases the website interest mining performance of the three proposed methods. The data (i.e., webpages) of this news event were collected from the largest Chinese search engine, *Baidu*[4]. Note that although these webpages are published in Chinese, the methods proposed in this paper could be used for any language. The collected data comprised webpages about this news event published over the course of 35 days. The webpages were pre-processed using word segmentation and lemmatization, and stop-words were removed to result in 81 websites, 99 webpages and 790 keywords after merging similar webpages. The statistics of each day are shown in Fig. 7. The experiment settings were as follow: 1) the website interests for this event were mined using three methods at all time stamps; 2) the methods were compared using the metric Sim_s developed in Eq. (22); and 3) the parameter ρ in Iterative Method was adjusted it from {0.1} to {1.0} in steps of {0.1}.

7.3. Results

The intuitive and iterative methods were compared first. The results for four different websites are shown in Fig. 8. Clearly, not all the different combinations of mapping relations and association relations (different ρ values) help to improve website interest mining performance. However, the iterative method reached its peak on four websites with $\rho = 0.4$, which suggests that website interests are preserved across the evolution of news events at this value. Here, one natural question arises: Was ρ's value the result of our website selection? To answer this question, we evaluated all the websites at different values of ρ. The average results are shown in Fig. 9, and, interestingly, the iterative method still reached its peak at $\rho = 0.4$. The Sim_s results for all websites using the iterative method with $\rho = 0.4$ compared to the intuitive method are shown in

[4] http://www.baidu.com.

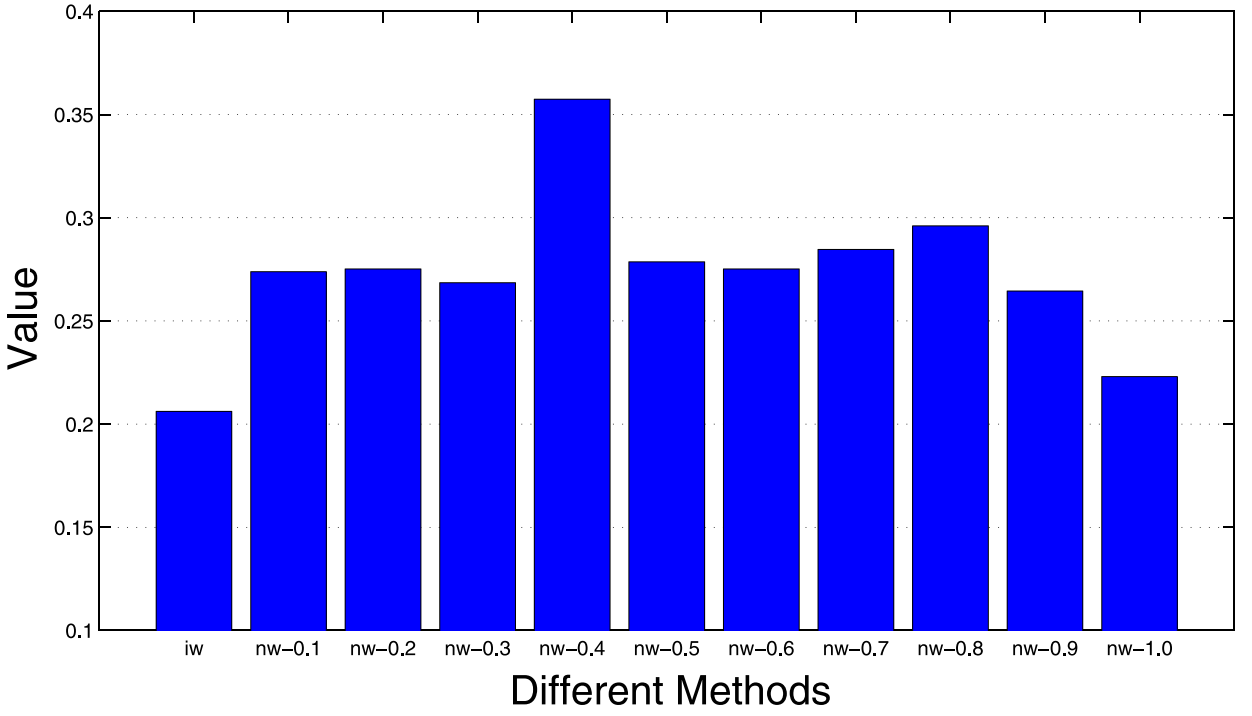

Fig. 9. The Sim_s of each website by different algorithms, including intuitive method and iterative method with different ρ.

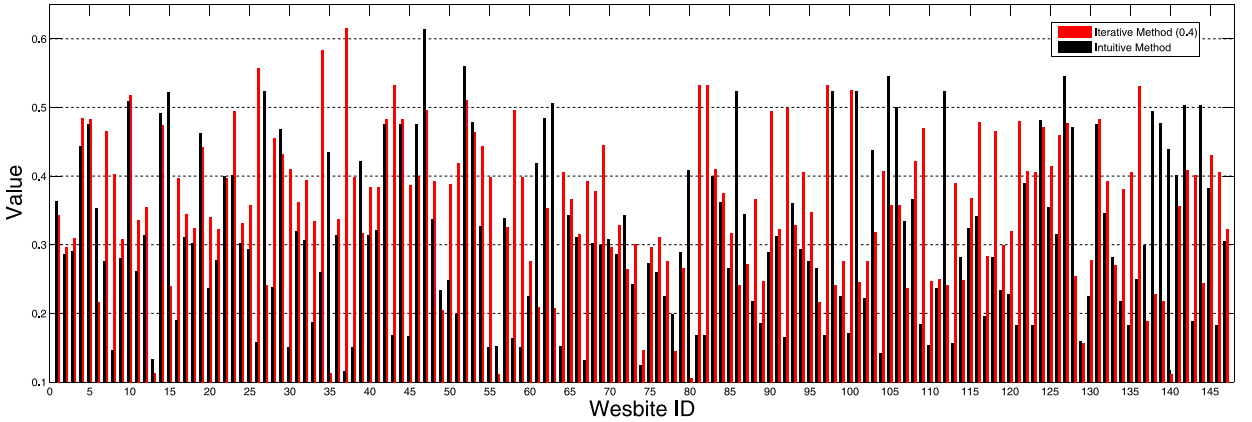

Fig. 10. Comparison between the intuitive method (blue) and iterative method (red) on of all websites in terms of Sim_s. The iterative method outperforms the intuitive method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 10. These results demonstrate that $\rho = 0.4$ is the perfect combination of the three-layered network structures with their mapping relations for this news event. It is worth noting that we believe the value of ρ is influenced by the evolution of the news event, but this kind of influence is not visible here because the two methods were compared using the same news event. For some news events, the evolution between consecutive days can be intense, then the value of ρ tends to be small, and vice versa. Overall, we conclude that the iterative procedure, if appropriately designed, is capable of improving website interest mining performance.

Next, we compare the iterative method and website topic model. Each of these methods is based on a different idea. The iterative method uses association relations implicitly; the website topic model uses them explicitly. The average results from all websites were 0.3575 for the iterative method and 0.4120 for the website topic model. These results demonstrate that the website topic model performs better than the iterative method and, hence, has the best performance of the three methods. However, it should be noted that the iterative method is with the best parameter ($\rho = 0.4$) in this comparison. We attribute the advantage of the website topic model to its soft assignment of words to different subtopics, which is a distinct feature among community-based methods including overlap community detection methods. This feature detects subtopics more accurately and, consequently, the mined website's interests are more accurate as well.

In terms of the computational complexity, the most expensive operation in one iteration of the iterative method is detecting the communities for the three layers. The complexity of community detection for one network is $O(V^2 + E)$ (eigenvector-based algorithm [28]) to $O(V + E)$ (label propagation-based algorithm [31]), where V is the node number and E is the edge number in the network. Considering all three layers, the smallest complexity for one iteration of the iterative method is $O(S + D + N + E_s + E_d + E_n)$. The most expensive operation in the website topic model is sampling. Using Gibbs sampling inference, the complexity for one iteration is $O(S + D + N + K)$, which means sub-topic distributions need to be updated. Since the scale of links is normally greater than the nodes, the complexity of the website topic model is likely to be relatively small. This is mainly because, rather than explicitly computing the links, they are considered through the probabilistic dependencies between latent variables.

8. Conclusions and further study

This paper proposes three different methods to resolve the problem of mining website interests relating to news events. All methods are based on the website-webpage-keyword hierarchy and the association relations between keywords. Two different strategies are used to take advantage of different kinds of information: one explicit, through the community detection; the other implicit, through a probabilistic graphical model. Two alternative methods are proposed within the explicit strategy - an intuitive method and an iterative method. Compared to the intuitive method, the iterative method additionally incorporates the dependencies between websites, webpages, and keywords to iteratively optimize the website interest mining process. The experimental results indicate that this iterative procedure is able to improve upon the performance of the intuitive method. Within the implicit strategy, a website topic model is built from the data to infer a website's interests. Compared to the explicit methods, this model softly assigns keywords to subtopics, which results in the best performance of all three methods. The ability to mine website interests could support personalized news services and malicious website detection.

In future, we will collect, evaluate, and compare more news events using the proposed methods to further our empirical knowledge. It would be also very interesting to design some advertising strategies [17,23] as a practical application for these website interest mining approaches.

Acknowledgments

Research work reported in this paper was partly supported by the Australian Research Council (ARC) under discovery grant DP140101366. This work was jointly supported by the Shanghai Committee of Science and Technology International Cooperation Foundation under grant no. 16550720400.

References

[1] K. Aberer, J. Wu, A framework for decentralized ranking in web information retrieval, in: Web Technologies and Applications, Springer, 2003, pp. 213–226.
[2] M. Andrews, G. Vigliocco, The hidden Markov topic model: a probabilistic model of semantic representation, Top. Cogn. Sci. 2 (1) (2010) 101–113.
[3] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu, Mining social emotions from affective text, IEEE Trans. Knowl. Data Eng. 24 (9) (2012) 1658–1670.
[4] D.M. Blei, J.D. Lafferty, A correlated topic model of science, Ann. Appl. Stat. (2007) 17–35.
[5] D.M. Blei, J.D. McAuliffe, Supervised topic models, arXiv:1003.0783v1 (2010).
[6] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[7] J. Chang, D.M. Blei, Hierarchical relational models for document networks, Ann. Appl. Stat. 4 (1) (2010) 124–150.
[8] M.J. Davern, D. Te'eni, J.Y. Moon, Content versus structure in information environments: a longitudinal analysis of website preferences, in: Proceedings of the Twenty First International Conference on Information Systems, Association for Information Systems, 2000, pp. 564–570.
[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[10] G. Feng, T.-Y. Liu, Y. Wang, Y. Bao, Z. Ma, X.-D. Zhang, W.-Y. Ma, Aggregate rank: bringing order to web sites, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2006, pp. 75–82.
[11] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[12] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.
[13] P.K. GM, K.P. Leela, M. Parsana, S. Garg, Learning website hierarchies for keyword enrichment in contextual advertising, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 425–434.
[14] T.L. Griffiths, M. Steyvers, D.M. Blei, J.B. Tenenbaum, Integrating topics and syntax, Adv. Neural Inf. Process Syst. 17 (2005) 537–544.
[15] H. Guo, Q. Chen, X. Wang, Z. Wang, Y. Wu, Strank: a SiteRank algorithm using semantic relevance and time frequency, in: Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, IEEE, 2009, pp. 4876–4881.
[16] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.
[17] P. Kazienko, M. Adamski, Adros-adaptive personalization of web advertising, Inf. Sci. (Ny) 177 (11) (2007) 2269–2295.
[18] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (5) (2009) 056117.
[19] J. Leskovec, K.J. Lang, M. Mahoney, Empirical comparison of algorithms for network community detection, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 631–640.
[20] C. Lin, R. Xie, X. Guan, L. Li, T. Li, Personalized news recommendation via implicit social experts, Inf. Sci. (Ny) 254 (2014) 1–18.
[21] N. Liu, C.C. Yang, A link classification based approach to website topic hierarchy generation, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 1127–1128.
[22] N. Liu, C.C. Yang, Keyphrase extraction for labeling a website topic hierarchy, in: Proceedings of the 11th International Conference on Electronic Commerce, ACM, 2009, pp. 81–88.
[23] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey, Decis. Support Syst. 74 (2015) 12–32.
[24] X. Luo, Z. Xu, J. Yu, X. Chen, Building association link network for semantic link on web resources, IEEE Trans. Autom. Sci. Eng. 8 (3) (2011) 482–494.
[25] P.N. Mendes, P. Mika, H. Zaragoza, R. Blanco, Measuring website similarity using an entity-aware click graph, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 1697–1701.

[26] M. Mitrović, B. Tadić, Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities, Phys. Rev. E 80 (2) (2009) 026123.
[27] R.M. Nallapati, A. Ahmed, E.P. Xing, W.W. Cohen, Joint latent topic models for text and citations, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 542–550.
[28] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[29] B. Poblete, R. Baeza-Yates, A content and structure website mining model, in: Proceedings of the 15th International Conference on World Wide Web, ACM, 2006, pp. 957–958.
[30] S. Qi, C. Ip, R. Leung, R. Law, A new framework on website evaluation, in: E-Business and E-Government (ICEE), 2010 International Conference on, IEEE, 2010, pp. 78–81.
[31] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.
[32] Y. Rao, Q. Li, X. Mao, L. Wenyin, Sentiment topic models for social emotion mining, Inf. Sci. (Ny) 266 (2014) 90–100.
[33] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Steyvers, Learning author-topic models from text corpora, ACM Trans. Inf. Syst. 28 (1) (2010) 4.
[34] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 424–433.
[35] L. Xie, Y.-L. Yang, Z.-Q. Liu, On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news, Inf. Sci. (Ny) 181 (13) (2011) 2873–2891. Including Special Section on Databases and Software Engineering.
[36] J. Xuan, J. Lu, G. Zhang, X. Luo, Topic model for graph mining, IEEE Trans. Cybern. 45 (12) (2015) 2792–2803, doi:10.1109/TCYB.2014.2386282.
[37] J. Xuan, X. Luo, G. Zhang, J. Lu, Z. Xu, Uncertainty analysis for the keyword system of web events, IEEE Trans. Syst. Man Cybern. Syst. 46 (6) (2016) 829–842, doi:10.1109/TSMC.2015.2470645.
[38] T. Yang, Y. Chi, S. Zhu, Y. Gong, R. Jin, Detecting communities and their evolutions in dynamic social networks: a Bayesian approach, Mach. Learn. 82 (2) (2011) 157–189.
[39] S. Zhao, M.X. Zhou, X. Zhang, Q. Yuan, W. Zheng, R. Fu, Who is doing what and when: social map-based recommendation for content-centric social web sites, ACM Trans. Intell. Syst. Technol. 3 (1) (2011) 5.