Neurocomputing 144 (2014) 463–470

Contents lists available at ScienceDirect

Neurocomputing

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

Predicting the topic influence trends in social media with multiple models ☆

Yi Han [a,b,*], Binxing Fang [b,c], Yan Jia [b]

[a] Peking University, China
[b] National University of Defense Technology, China
[c] Beijing University of Posts and Telecommunications, China

ARTICLE INFO

Article history:
Received 18 October 2013
Received in revised form
18 February 2014
Accepted 11 March 2014
Communicated by M. Wang
Available online 19 June 2014

Keywords:
Social network
Prediction
Time series
Multiple models

ABSTRACT

Online social networks, such as twitter and facebook, are continuously generating the new contents and relationships. To fully understand the spread of topics, there are some essential but remaining open questions. Why are some seemingly ordinary topics attracting? Is it due to the attractiveness of the content itself, or some external factors, such as network structure, time or event location, play a larger role in the dissemination of information? Analyzing the influence and spread of upcoming contents is an interesting and useful research direction, and has brilliant perspective on web advertising and spam detection. In this paper, a novel time series model for predicting the topics social influence has been introduced. In this model, the existing user-generated contents are summarized with a set of valued sequences, and a hybrid model consisting of topical, social and geographic attributes has been adopted for predicting influence trends of newly coming contents. The empirical study conducted on large real data sets indicates that our model is interesting and meaningful, and our methods are effective and efficient in practice.

1. Introduction

With the rapid growth of the social media, the information spreads around the world with surprising speed and intensity. Users of some online social media, such as twitter and facebook, are continuously generating new contents, forming new interactions, and updating their status over time. With the evolving of the network structure, some contents generate and spread very fast. In twitter, some interesting posts (tweets) are usually retweeted thousands of times, and their social influence is also boosted with the reposting activities.

There are some interesting questions people may ask. Why are there some topics obviously more influential than other? What kind of topics could be attractive? Is there a method to predict or estimate the influence of a specified topic? Is the topological substructure of a social network related with the evolution of topics? Answering these questions is essential for understanding the mechanism of evolving social networks.

When a specified topic becomes a hot spot, we consider there are two possible reasons. The topic itself is interesting, or some external factors amplify its influence.

Social media has been studied extensively from variable angles such as degree distribution analysis [1,2], community extraction [3] and pattern discovery [4]. In this paper, we analyze the relationship between hot topics in social media, and formulate the information flows on different topics as a set of sequences. In order to predict the influence trend of a specific topic, an early prediction method with multiple factors is adopted. Moreover, we introduced a novel supervised learning method which considers topical, social and geographic properties of information flows in social networks.

To the best of our knowledge, there is no previous study on social media taking into account the early prediction and geographic properties. We made the following contributions.

First, we introduce a novel time series model to represent the continuously generated content in social media. The fact that user-created contents on a specified topic spread in a specific social network can be modeled by a sequence of vertices which participate the interaction.

Second, we propose a similarity measure among network time series. In a given time period, the continuously generated content about a given topic can be represented as a time series. For different topics, the distance among time series, which can be

☆The research was supported by National Basic Research Program of China (973 Program, No. 2013CB329600), National Natural Science Foundation of China (Nos. 91124002 and 61372191), and China Postdoctoral Science Foundation Program (Nos. 2012M520114 and 2013T60037).

* Corresponding author.
E-mail address: yihan@pku.edu.cn (Y. Han).

http://dx.doi.org/10.1016/j.neucom.2014.03.054
0925-2312/

regarded as the similarity, is measured by content similarity with social and geographic attributes.

Third, we propose a novel prediction model on social media. The influence trends of the topics, as the target value, can be predicted effectively. The influence of upcoming content will be estimated by analyzing the individuals and related existing content.

Last, we conduct systematic experiments on two real data sets. The experimental results indicate that our model is useful and interesting, and our methods are effective.

Our solution can be applied as two different types of applications, topic tracing (Fig. 1(a)) and potential topic discovery (Fig. 1(b)).

In Fig. 1, both are based on the supervised classification model, in which the data before the current time stamp can be used as training data to generate the classifier, and the newly coming content can be used as testing data to tune the classifier. For a given topic, its spreading trace, which can be represented as a series of vertices, keywords and time stamps, can be extracted easily. The classifier which has been retrieved from the network can be applied for predicting the future state of the time series. Fig. 1(a) shows the example. Another important application is the influential topic prediction. The system automatically estimates the future impact of candidate time series, and outputs the topics which are potentially influential (shown in Fig. 1(b)).

The rest of the paper is organized as follows. We review the related work in Section 2, and formulate the problem in Section 3. We discuss the similarity measure and prediction methods in Section 4. A systematic empirical study conducted on real data sets is reported in Section 5. Section 6 concludes the paper.

2. Related work

Our work is highly related to the previous studies on spreading dynamic models, classification and prediction on social media, and social influence estimation. In this section, we review some representative work briefly.

2.1. Spreading dynamic models

Epidemic propagation model is a successful mathematical model for which has a long history, in which, the statues of individuals can be summarized into 3 categories. S (susceptible) indicates the individual is in a healthy state and has a probability of being infected by someone. I (infected) indicates the individual has been inflected and has a probability of infecting others or being recovered. In a network, if two vertices are connected then they are considered to have contact, Thus, if one vertex is infected by a virus and the other is susceptible, then with a certain probability the latter may become infected as time goes on. R (recovered) indicates the individual cannot be infected anymore. Different combinations of above states lead to different models, such as SIR [5] and SIS [6] models. Epidemic propagation model is a straightforward model for describing the information diffusion process. However, the events on social media are affected by lots of external factors, like emergencies, breaking news, social spammers, which cannot be characterized by epidemic propagation model well.

2.2. Classification and prediction on social media

Supervised learning [7] on social networks is a central subject in graph data processing. Some previous studies utilized a certain number of subgraphs as training set. In training set, the target values, which can usually be vertex properties, are available. The goal is to derive the target values of the remaining part of the graph. In some large-scale social networks, a central task is to classify unlabeled nodes given a limited number of labeled nodes. For example, the social service provider manually labels a small number of people who responded to a certain advertisement as positive nodes, and people who did not respond as negative nodes. Based on these labeled nodes, other people's response can be predicted. Graph classification tasks can also be unsupervised. Unsupervised methods classify graphs into a certain number of categories by similarity [8,9]. An interesting direction on evolving social network is the link prediction. The appearance of new links indicates new interactions between vertices. Given a snapshot of a social network at time t and a small number Δt, the objective is to predict the edges that will be added to the network in the time interval $(t,t+\Delta t)$. Murata considered that proximities between nodes can be estimated by using both graph proximity measures and the weights of existing links, and a prediction method based on weighted proximity measure has been introduced in [10]. Leskovec introduced a method of predicting the edge sign. The main goal is to utilize the graph structure and vertex label to infer the hidden sign labeled on edges [11]. A logistic regression classifier has been used to combine the evidence from individual features. Backstrom in facebook developed an algorithm that combines the information from the network structure with node and edge level attributes to guide a random walk on the graph [12]. The scoring function was learned in order to assign weights to edges, and walkers are regarded more likely to visit the nodes to which new links will be created.

2.3. Social influence estimation

Several well-known link-based ranking algorithms, including PageRank [13] and HITS [14], have been designed for ranking entities in social networks in past decades. PageRank [13] measures the importance of a vertex v by considering how collectively others pointing to directly or indirectly. For each vertex, the amount of ranking contribution from a neighbor is decided by

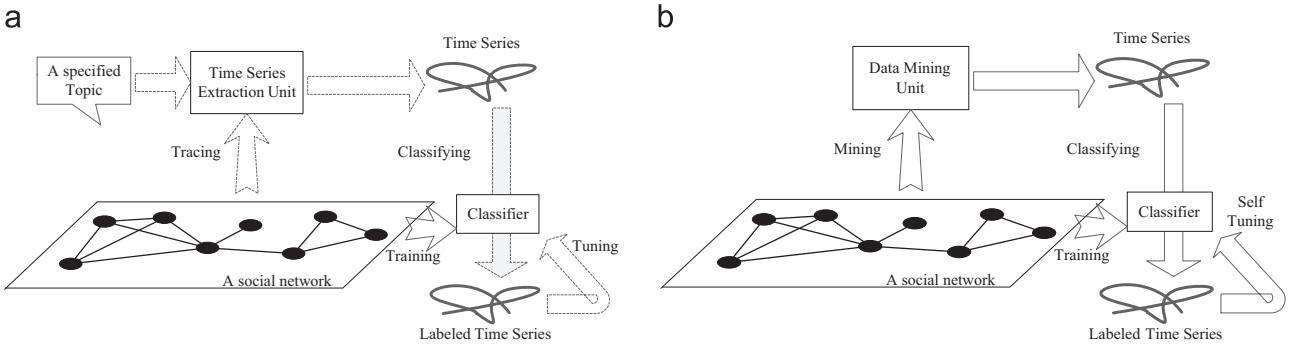


Fig. 1. Two different types of applications. (a) Topic tracing. (b) Potential topic discovery.

the ranking score and the out-degree of the neighbor. Kleinberg proposed Hyperlink-Induced Topic Search (HITS) algorithm for scoring the entities in 1998 [14]. Similarly to PageRank, HITS assigns each entity two scores, its authority and hub value, which measures entities content and linkage, respectively. Some variations of link-based ranking algorithms have been proposed in recent years. Faloutsos [15] proposed an improved random walk model, where a probability of restart has been added to classic random walk model, which can be used for measuring the distance among vertices. In this paper, the social influence value of individuals is measured in a PageRank-like way.

3. Problem formulation

Many social networks, such as online social services, co-authorship networks, web, internet topology map, can be represented by graphs. In order to keep our discussion simple, we assume the network topological structure of social networks does not change, but the content labeled on vertices are evolving. Our solution can be easily extended to the case of evolving structures.

3.1. Topical time series model

In this paper, a topic or a keyword is represented by l and a user's published content is represented by a set of l. Please note that, we focus on analyzing and predicting the topics' influence trends. We assume that the topics have been extracted as sets of keywords in previous steps. Moreover, most of social network sites, such as twitter and facebook, have provided hashtag-like ways for marking the topics, which can also be used as representing the topics. In the information retrieval area, there are many mature topic extraction methods [16,17], so we will not repeat them.

Definition 1. (Occurrence Sequence) In a social network G, $v \in V$ represents the vertices. $\langle t,v,l\rangle$ is an occurrence of l, representing v published content labeled by l at time t.

For a topic l_0, we sort all l_0's triples $\langle t,v,l_0\rangle$ with ascending order in t. l_0's sequence of occurrences, representing l_0's dissemination in the entire network, is

$$ts_{l_0} = \langle t_0, v_0, l_0\rangle,\ \langle t_1, v_1, l_0\rangle, \ldots (t_0 < t_1 < \cdots)$$

A topic may receive different attentions with time going, so we adopt the popularity sequence model representing the phenomenon of topic influence l_0 changing with time. The popularity sequence is thought of as the density according to which a large number of occurrences are distributed. Fig. 2 shows an example of popularity sequence, in which, each point on time axis indicates an occurrence at corresponding time t. The influence sequence model reflects a given topic's influence changing patterns, such as trends, periodical changes, or random changes.

Definition 2. (Popularity) A topic l's popularity function can be represented by $\rho_l: t \to R^+$. $\rho_l(t)$ is the popularity value of l at time t, represents the density of occurrences of l at time t.

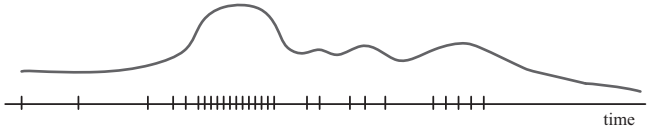


Fig. 2. Popularity sequence.

$$\rho_l(t) = \frac{1}{nw}\sum_{i=1}^{n}\theta\left(\frac{t-t_i}{w}\right) \tag{1}$$

In this paper, a kernel density estimation based method is adopted for calculating the popularity of l. In Eq. (1), $w > 0$ is a smoothing parameter called the bandwidth, and $\theta()$ is the kernel, a symmetric but not necessarily positive function that integrates to one $i = 0, 1, \ldots, n$ represents the n occurrences.

$$\int_{-\infty}^{+\infty}\theta(x)dx = 1$$

$$\theta(-x) = \theta(x) \quad \text{for all values of } u$$

In real applications, a range of implementations of $\theta(x)$ are commonly used: uniform, triangular, triweight [18], Epanechnikov [19], normal, and others. In this case, a smooth curve in t is expected, so a Gaussian basis function is adopted.

$$\theta(x) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)x^2}$$

Actually, the process of density estimation on discrete events can be regarded with the statistics on sliding window on time series. Fig. 3 shows the interaction between bandwidth and time series. The larger the windows on time series we use, the smoother the curve and more information loss we get.

3.2. Social influence sequence

In some social networks, using special network substructures, some topics can be boosted in a very short time. The contents on some bridge nodes and hub nodes usually have larger possibility to be visited than other ordinary vertices. Some interesting topics may be boosted by the activities of influential persons.

Example 1. (Topic influence) In twitter, the information retweeted by "Lady Gaga", who has millions of followers, has a much larger probability to be noticed than an ordinary tweet without any retweet. The participation of such "famous" IDs increases the popularity of contents, so that it helps the information spread faster.

We consider the individual influence should be taken into account when predicting the influence. In social networks, the topological structure usually indicates difference of individual importance. We add an authority weight to each appearance of a specified topic when generating the time series.

A social network can be modeled as a graph $G = \langle V, E, \mathcal{L}\rangle$, where V is the set of entities in the network, an edge $e = \{u, v\} \in E$ is a tuple of 2 vertices, and E is the edge set. The vertices are labeled, $\mathcal{L}$ is a labeling function on vertices. Each vertex $v \in V$ is labeled by $\mathcal{L}(v)$, the attributes that the entity has. In this paper, $\mathcal{L}(v)$ is $\mathcal{L}: V \to 2^{T\times W}$, where $\langle t, l\rangle \in \mathcal{L}(v)$ represents v is labeled by l at time t, and W is the set of all the topics.

In this model, each tuple $\langle t,v,l\rangle$ will be weighted by $\sigma(v)$, where $\sigma: v \to R^+$ is an authority function on v.

Definition 3. (Influence) A topic l's influence function can be represented by $\lambda_l: t \to R^+$. $\lambda_l(t)$ is the popularity value of l at time t, represents the density of occurrences of l at time t.

$$\lambda_l(t) = \frac{1}{nw}\sum_{i=1}^{n}\sigma^{\alpha}(v_i)\theta\left(\frac{t-t_i}{w}\right) \tag{2}$$

In Eq. (2), v_i represents the vertices which generates content l. θ is weighted by $\sigma^{\alpha}(v_i)$. σ is the authority function, and α is a user specified factor.

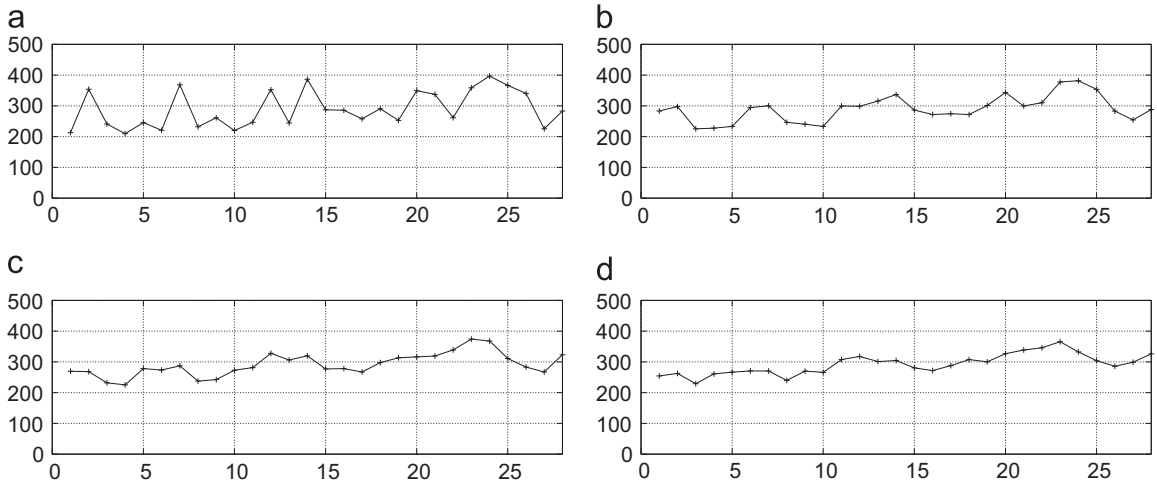


Fig. 3. Popularity sequence in various sliding windows. (a) $w=24$ h. (b) $w=48$ h. (c) $w=72$ h. (d) $w=96$ h.

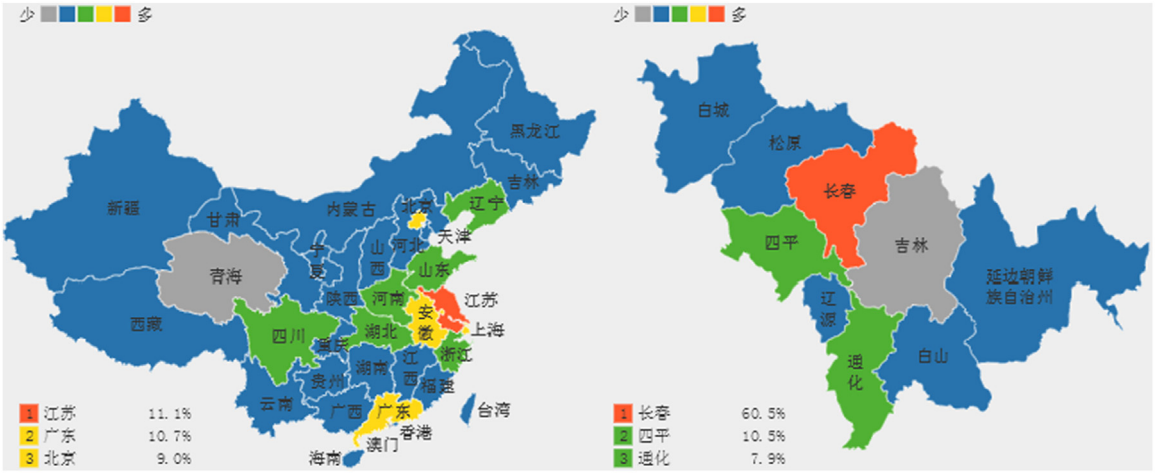


Fig. 4. An example of location popularity in China.

Several well-known link-based ranking algorithms, including PageRank [13] and HITS [14], can be adopted here as $\sigma(v)$ to measure the vertices' importance. Some variations of link-based ranking algorithms have been proposed in recent years, including TrustRank [20], RWRS [15], and so on.

Similar with infection process of spreading dynamic models in Section 2.1, we consider the contents can only be passed through links. Therefore, each time series should be connected and traceable.

Definition 4. (Connected and traceable) In an occurrence sequence ts_l, if $\langle t,v,l\rangle \in ts_l$ ($t_s \le t \le t_e$), in which t_s and t_e represents the beginning and ending time respectively, then there exists another tuple $\langle t',v',l\rangle \in ts_l$ ($t_s \le t' < t \le t_e$) and $\{v',v\} \in E$. In another word, in a time series, for any content after time t, there exists a neighbored source before time t. Therefore, each time series also represents a connected subgraph. In some cases, in order to meet the requirements, a set of tuples on same topic l should be partitioned into several connected subgraphs.

Definition 4 indicates that in a time series, a non-original topic will be able to find its source, and the induced subgraph of a time series must be connected, which represents the connected information propagation map. Thus, extracting the time series for a specific topic can be divided into two steps.

Firstly, we extract the subsets of network vertices based on topics. Each subset represents a topic. Secondly, we select the earliest examples $i_0=\langle t,v,l\rangle$ in a subset. Based on connectivity and traceability requirements, we search the sub-tree rooted by i_0 and remove them from the subset until all nodes are removed.

3.3. Location sequence

In social media, the influence of some events is highly related with their locations.

Example 2. (Location) An important characteristic of twitter is its real-time nature. For example, when sport events, shopping mall sales, discounts or some special events, people nearby usually tweet related to them.

The location factor can be utilized for improving the accuracy of prediction.

Definition 5. (Location popularity) F^d is a two dimensional space which indicates the all possible locations. Given a topic l and its occurrence l_0, and a location $p \in F^d$, the influence of l_0 on p, denoted by $\tau(l_0 \to p)$, is

$$\tau(l_0 \to p) = e^{-(d(l_0,p)^2)/(2s^2)}$$

$d(l_0,p)$ is the distance from l_0 to p, s is a user specified factor indicating the size of window. For a topic l, the location popularity at a location $p \in F^d$ is defined as the sum of influence functions of

all l's occurrences.

$$\tau_l(p) = \sum_{i=i}^{n} e^{-(d(l_i,p)^2)/2s^2}$$

In this paper, we adopted the geographical distance as the distance function d. Fig. 4 shows an example of location popularity of a criminal case occurred in Shanghai and Changchun.

Please note that the distance function d is not suitable in some cases. For example, Pyongyang and Seoul are geographically close, but rarely communicate in social media.

Definition 6. (Influence with location factor) A topic l's influence function can be represented by $\phi_l: t \times F^2 \to R^+$. $\phi_l(t,p)$ is the popularity value of l at time t and location p, represents the density of occurrences of l at time t and a specific location p.

$\phi_l(t,p) = \lambda_l(t)\tau_l(p)$

4. Prediction

The classification is a supervised learning process to assign labels to unknown upcoming tuples. The classic classification process can be divided into two steps, learning and classification.

In the training process, a set of pre-labeled time series, denoted by $T_{training}$, is analyzed for generating a set of classification rules, or a classifier. $T_{training}$, which is also called training set, contains a set of labeled time series. Each time series $ts \in T_{training}$ carries a class label, denoted by $ts.c \in \mathcal{C}$, where $\mathcal{C}$ is the set of labels. A classifier is a mapping $\mathcal{F}$ from a time series space R^{TS} to $\mathcal{C}$. In the second step, a set of test time series, denoted by T_{test}, will be used to estimate the accuracy of the classifier. T_{test} is a set of time series such that each time series $ts \in T_{test}$ also carries a label $ts.c \in \mathcal{C}$. The quality of a classifier $\mathcal{F}$ can be measured by its accuracy on a specified test set. The accuracy of a classifier on a test set is the ratio that the generated class labels match those carried by the time series in the testing set, that is,

$$Accuracy\ (\mathcal{C}, T_{test}) = \frac{|\{ts \in T_{test} : \mathcal{F}(ts) = ts.c\}|}{|T_{test}|}.$$

If the accuracy is considered acceptable, the classifier can be applied to the classification of newly coming time series.

4.1. Early prediction

For a graph $G = \langle V, E, \mathcal{L} \rangle$, and a time series ts, an early classifier $\mathcal{F}_e$ can conduct the classification based on the prefix of ts. An early classifier is serial if $\mathcal{F}_e(s_{t_0}) = \mathcal{F}_e(s_{t_0+\Delta t})$ for any $\Delta t > 0$, where $\mathcal{F}_e(s_{t_0})$ is the subsequence before time stamp t_0, and Δt is a positive number representing a small period.

Therefore, for a time series ts, the influence of ts can be predicted before its ending. In this paper, our goal is to design a method to generate a classifier. The target value is Boolean value, which represents the trend of the time series.

The nearest neighbor (NN for short) based prediction has been widely used for time series classification. The main idea of the NN based classification is as follows. All the training time series are put into a pool. For a time series ts_u whose label is unknown, the NN classifier finds the closest match in the pool, and assigns ts_u with its nearest neighbor's label. For two time series ts_1 and ts_2, the closeness/distance can be measured, and we denote the distance by $dist(ts_1,ts_2)$.

Some popularly used distance measures include Euclidean distance [21], dynamic time warping distance (DTW) [22], and alignment based distances [23] (for symbolic sequences).

4.2. Similarity measure

We assume the entire topic set is certain, and topics can be organized into a knowledge hierarchy. The similarity of two specified topics can be measured by the tree distance. However, in some real applications, the hierarchical tree is not sufficient to describe the real world. For example, "apple" can be categorized as fruits, or a popular electronic brand. Therefore, some scholars introduced an OLAP based knowledge hierarchical structure [24]. A Topic-Concept cube has been used for organizing the keywords, the distance between topics can be retrieved by calculating the shortest path in the cube.

In this paper, all the possible topics can be found in a large knowledge network, and the length of the path is adopted as the similarity measure. That is,

$$\mathcal{S}_t(l_1, l_2) = \frac{1}{length_{path}(l_1 \to l_2)}$$

4.3. A nearest neighbor based solution

In order to assign the labels to newly coming time series, the NN-based classifier searches the previous labeled time series, and uses the closest match's label for pending time series.

The NN-classifier of a graph $G = \langle V, E, \mathcal{L} \rangle$, denoted by $\mathcal{F}_G$, contains a set Ω of time series with labels. For a new time series ts with unknown labels, $\mathcal{F}_G(ts) = influence(nn_\Omega(ts))$.

Definition 7. (Nearest neighbor) In a graph $G = \langle V, E, \mathcal{L} \rangle$, a time series set Ω, and a time series $ts_{l_1} \notin \Omega$, the set of nearest neighbors of ts in Ω, denoted by $nn_\Omega(ts)$, is

$$nn_\Omega(ts) = \{ts' \in \Omega | \nexists ts'' \in \Omega : \mathcal{S}(ts'', ts) > \mathcal{S}(ts', ts)\}$$

Ω carries a minimum prediction length, which is a numeric attribute. The value of minimum prediction length will be learned during the training process. For a time series ts, if its prefix with length of k matches a time series in Ω, and its length meets the constraint of minimum prediction length, the influence of ts can be measured. Formally, if $nn(ts(k)) = ts'$, where $ts' \in \Omega$, and $k \le MPL(ts')$, $\mathcal{F}_\Omega ts = influence(ts')$. For a time series $ts \in \Omega$, minimum prediction length of ts, denoted by $MPL(ts)$, can be calculated as follows. For a time series $ts \in \Omega$, minimum prediction length of ts, denoted by $MPL(ts)$, can be calculated as follows. For a time series $ts' \in T_{test}$, if $ts = nn_\Omega(ts') = nn_\Omega(ts'(m))$ $(m = k, k+1, \ldots)$, but $nn_\Omega(ts') \neq nn_\Omega(ts'(m-1))$, $MPL(ts) = k$.

5. Experimental result

In this section, we report a systematic experimental study on large real data sets. The effectiveness of nearest neighbor based predictor has been demonstrated. We also discuss the effectiveness of the social influence and location factor. All programs were implemented in Java using eclipse platform.

5.1. Data sets

We verify our models and algorithms in YHPODS system. YHPODS, which is an open-source project based on the Apache Software Foundation's unstructured information management architecture (UIMA) platform, is designed by National University of Defense Technology (NUDT for short), and run on a super-computer cluster with 64 nodes. It is keeping collecting labeled social network data from the micro-blogging sites.

We collected 18,012,823 user IDs and 33,237,699 who follows whom (WFW) relations from twitter, and their tweets in a period of 1

month. We call the data set the **TW data set**. A stop list with is used for The Porter Stemming Algorithm [25] is adopted here for removing the commoner morphological and inflexional endings from labels. We train the data set by using 300 carefully selected topics.

Fig. 5 shows 4 examples of time series on TW data, where X axis represents time. We sampled value of the time series in units of days. Y axis represents the influence function value. The size of windows is 6 h, and σ is PageRank.

In Fig. 5, ts_{apple} is relatively stable in whole month, but ts_{sport} shows some irregular movements. Since the stock markets are closed on the weekends, ts_{stock} reflects the cyclical characteristics. ts_{UEFA} goes with obvious irregular jitter. We guess there are more discussions since Feb. 17th and 22nd are the first and second round of the UEFA Champions League.

TW data does not contain the geographic tags. For verifying the effectiveness of location factor in Section 3.2, we crawled a small portion of Sina Weibo (the largest Chinese microblogging network) in 2013. We call this data set WB data. The WB data contains 10,427 bloggers and 24,313 reposts, all contents carry location tags. In this data set, each node represents a Sina Weibo blogger ID; if a blogger reposts other's contents, there exists a directed edge between them.

5.2. Training and prediction

For verifying the effectiveness of the algorithm, the minimal prediction length (MPL for short) is introduced in Section 4.1. We use over 1400 manually selected topics of twitter data before 2008 as training set for time series in Ω. For each $ts \in \Omega$, the minimal prediction length $\in(ts)$ is calculated. Fig. 6 shows Ω the distribution of $\in(ts)$.

In Fig. 6(a), the x axis shows the 300 manually selected topics, and the y axis shows the value of $\in(ts)$. We can find that, for a small number of ts, the $\in(ts)$ is about 50%. That means for these topics, only half of an expected period of data needs being learned for a success prediction. For example, $\in(ts_{stock})$ in Fig. 6 is 14 days.

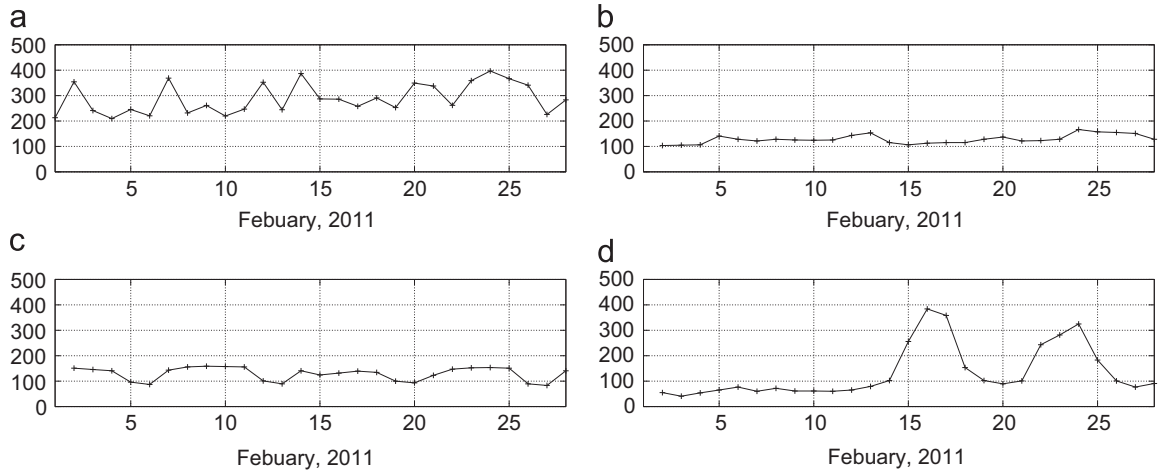


Fig. 5. Four examples of time series. (a) Sport. (b) apple. (c) stock. (d) UEFA.

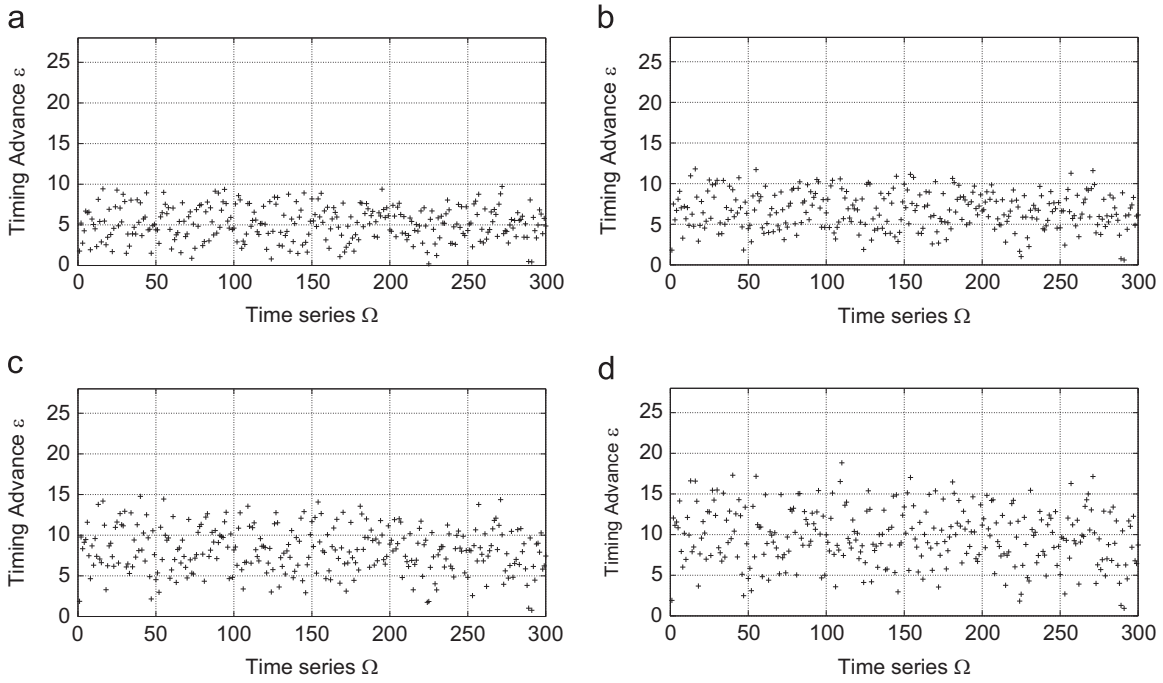


Fig. 6. Ω and $\varepsilon(ts)$. (a) $w=12$ h. (b) $w=24$ h. (c) $w=36$ h. (d) $w=48$ h.

The results demonstrate that by using case based reasoning, the most important increase of prediction accuracy occurs on the prefixes through about 50% of the full length.

Moreover, we also analyzed how ε changes with sliding window size. In Fig. 6, we can see if w is smaller than 12 h, some time series in Ω can get large ε. When w goes up, ε also goes up, but slower. When w goes to 48 h, for some time series, $\in$ is even greater than 50%.

5.3. Precision

In our model, the size of sliding window is an important parameter. As shown in Fig. 2, the width of the sliding window determines the smoothness of the time series, and plays an important effect on the accuracy of the classification. If the sliding window is large, the curve is relatively smooth, easier to find the best match. However, if the sliding window is small, the curve becomes sharp, the burst topics can be shown clearer, but the learning process is also easier to be over-fitting.

We divided all the time series into 4000 small pieces, and got 4000 subsequences. We use average trend of each subsequence as a target attribute for testing the topic classification accuracy of our solution. In Fig. 7, the bars labeled by "select" represents the overall accuracy result on selected 300 topics, and "random" represents the random topics.

In Fig. 7, for random topics, the accuracy result is irrelevant with the size of sliding window. However, obviously, the overall accuracy of select topics is much higher than that of random topics.

5.4. Effectiveness of social influence

In TW data, the follow relationships are kinds of information pipes which work in "subscribe–distribute" ways. The follow activities can be considered as building connections, and they cannot form the information spreading or influence amplification directly. The repost activities are secondary dissemination of information. The social influence measured on repost network is more feasible, so we choose the repost network to verify the social influence and location factor.

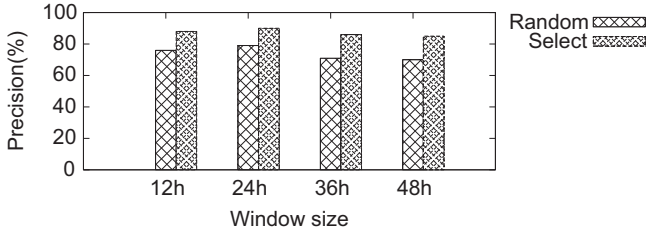


Fig. 7. Sliding windows and precision.

Fig. 8 shows two real information propagation maps extracted from WB data. We can easily find that the map in Fig. 8(a) looks more likely artificial. Fig. 8(a) shows a very simple repost structure, and the depth of tree is small. The propagation tree in Fig. 8(b) is more natural. We can infer that the individuals shown in Fig. 8(a) are social spammers.

We run a series of experiments to determine the value of α (Section 3.2, a manually selected parameter for balancing the authority function in λ_l). In order to get the optimal MPL value, for each topic l, the different values of α are tested until the MPL value cannot be better. $\alpha=0$ indicates $\sigma^{\alpha}=1$, represents the social influence factor α is useless for improving the prediction. For our manually selected 300 topics, only 19 topics meet such condition.

5.5. Effectiveness of location factor

Different with the social influence factor, the location factor upgrades the time series into two dimensions. $\phi_l(t,p)$ indicates the time series are time-related as well as location-related. We have the function $\phi_l(t,p)$ for integral operation on different ranges (entire country, provinces, cities and counties) in order to get the optimal MPL value. For each topic l, the different values of α are tested until the MPL value cannot be better.

For 300 manually selected topics, we calculate the number of topics in different ranges on which the MPL performs optimal. Table 1 listed the result. Since the most users of Sina Weibo locate in China, we only calculated the results in Mainland China. Please note that, Beijing, Shanghai, Chongqing and Tianjin are four direct-controlled municipalities Mainland China administratively. We count them as provinces.

In Table 1, the topics in "Entire country" class mean that the MPL does not perform better in any sub-regions. We can find that over 60% topics perform better in province level. The location factor significantly helped improving the prediction time.

Table 1
Optimize the MPL with location factor.

Range	Entire country	Province	City	County
Percentage (%)	12	65	23	0

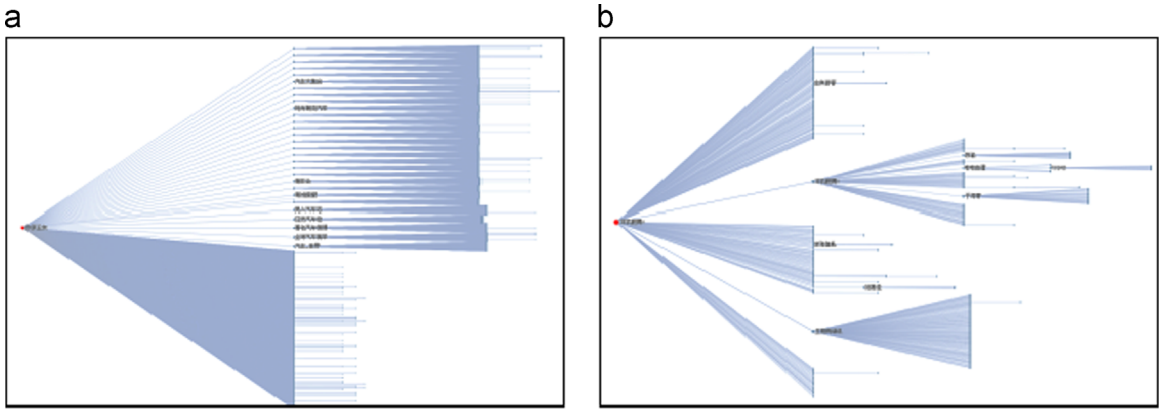


Fig. 8. Sliding windows and precision. (a) uid: 1276821910. (b) uid: 2145494291.

6. Conclusion

In this paper, a novel time series model has been proposed. In this model, the existing user-generated contents are summarized with a set of valued sequences. An early predictor based on Nearest Neighbor is adopted for analyzing the properties of series. The influence of newly coming contents is estimated with the predictor. The experimental results demonstrate that by using case based reasoning. The predictors are able to affirm an early time of reliable predictions.

Analyzing and predicting the influence and spread of upcoming contents is an interesting and useful research direction, and has brilliant perspective on web advertising and spam detection.

References

[1] J. Kleinberg, Small-world phenomena and the dynamics of information, in: Proceedings of Advances in Neural Information Processing Systems 14, MIT Press, 2001, p. 2001.

[2] H. Tong, S. Papadimitriou, P.S. Yu, C. Faloutsos, Proximity tracking on time-evolving bipartite graphs, in: Proceedings of the 2008 SIAM International Conference on Data Mining (SDM' 08), 2008.

[3] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 26113.

[4] N. Vanetik, E. Gudes, S.E. Shimony, Computing frequent graph patterns from semistructured data, in: ICDM, 2002, pp. 458–465.

[5] Boris Shulgin, Lewi Stone, Zvia Agur, Pulse vaccination strategy in the SIR epidemic model, Bull. Math. Biol. 60 (6) (1998) 1123–1148.

[6] Romualdo Pastor-Satorras, Alessandro Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (14) (2001) 3200.

[7] U. Lee, Z. Liu, J. Cho, Automatic identification of user goals in web search, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005, pp. 391–400.

[8] K. Tsuda, T. Kudo, Clustering graphs by weighted substructure mining, in: ICML, 2006.

[9] K. Tsuda, K. Kurihara, Graph mining with variational dirichlet process mixture models, in: SDM, 2008, pp. 432–442.

[10] T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measures, in: Proceedings of the International Conference on Web Intelligence, IEEE/WIC/ACM, IEEE, 2008, pp. 85–88.

[11] Jure Leskovec, Daniel Huttenlocher, Jon Kleinberg, Predicting positive and negative links in online social networks, in: Proceedings of the 19th International Conference on World wide web, ACM, 2010.

[12] L. Backstrom, J. Leskovec, Supervised random walks: predicting and recommending links in social networks, in: WSDM, 2011.

[13] Taher H. Haveliwala, Topic-sensitive pagerank, in: Proceedings of the 11th International Conference on World Wide Web, ACM, 2002.

[14] Jon M. Kleinberg, et al., The Web as a Graph: Measurements, Models, and Methods. Computing and Combinatorics, Springer, Berlin, Heidelberg (1999) 1–17.

[15] Tong Hanghang, Christos Faloutsos, Jia-Yu Pan, Fast random walk with restart and its applications, 2006.

[16] Kathleen McKeown, Dragomir R. Radev, Generating summaries of multiple news articles, in: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1995.

[17] Regina Barzilay, Michael Elhadad, Using lexical chains for text summarization, in: Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, vol. 17(1), 1997.

[18] Philippe Delsarte, Jean-Marie Goethals, Tri-weight codes and generalized Hadamard matrices, Inf. Control 15 (2) (1969) 196–206.

[19] Vassiliy A. Epanechnikov, Non-parametric estimation of a multivariate probability density, Theory Probab. Appl. 14 (1) (1969) 153–158.

[20] Baoning Wu, Vinay Goel, Brian D. Davison, Topical trustrank: using topicality to combat web spam, in: Proceedings of the 15th International Conference on World Wide Web, ACM, 2006, pp. 63–72.

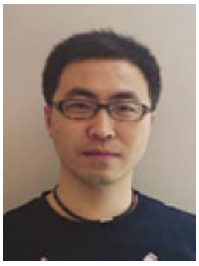
[21] E.J. Keogh, S. Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration, in: KDD, 2002, pp. 102–111.

[22] E.J. Keogh, M.J. Pazzani, Scaling up dynamic time warping for datamining applications, in: KDD, 2000, pp. 285–289.

[23] L. Kajan, A. Kertesz-Farkas, D. Franklin, N. Ivanova, A. Kocsor, S. Pongor, Application of a simple likelihood ratio approximant to protein sequence classification, Bioinformatics 22 (23) (2006) 2865–2869.

[24] D. Kang, D. Jiang, J. Pei, Z. Liao, X. Sun, H.-J. Choi, Multidimensional mining of large-scale search logs: a topic-concept cube approach, in: WSDM, 2011.

[25] M.F. Porter, An Algorithm for Suffix Stripping, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA (1997) 313–316.



Yi Han is a postdoctoral research fellow at School of Electronics Engineering and Computer Science (EECS), Peking University. He received a BEng degree, a MEng degree and a Ph.D. degree from National University of Defense Technology, in 2004, 2006 and 2011, respectively. His research interests can be summarized as developing effective and efficient data analysis techniques for novel data intensive applications. Particularly, He is currently interested in various techniques of data mining, web search, information retrieval, data warehousing, online analytical processing, and database systems, as well as their applications in social networks and business. His research has been supported in part by the National Natural Science Foundation of China, National High-tech R&D Program of China (863 Program), National Basic Research Program of China (973 Program), and Postdoctoral Science Foundation of China.

Binxing Fang is a full professor at Beijing University of Posts and Telecommunications. His research interests include information network analysis, information security, and network security.

Yan Jia is a full professor at National University of Defense Technology. Her research interests include information network analysis, data mining, information security.