Decision Support Systems 48 (2010) 354–368

Contents lists available at ScienceDirect

Decision Support Systems

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

Using text mining and sentiment analysis for online forums hotspot detection and forecast

Nan Li [a], Desheng Dash Wu [b,c,*]

[a] Department of Computer Science, University of California, Santa Barbara, USA
[b] Reykjavík University, Iceland
[c] RiskLab, University of Toronto, Canada

ARTICLE INFO

Article history:
Received 15 July 2008
Received in revised form 8 September 2009
Accepted 17 September 2009
Available online 24 September 2009

Keywords:
Text mining
Sentiment analysis
Cluster analysis
Online sports forums
Dynamic interacting network analysis
Hotspot detection
Machine learning
Support vector machine

ABSTRACT

Text sentiment analysis, also referred to as emotional polarity computation, has become a flourishing frontier in the text mining community. This paper studies online forums hotspot detection and forecast using sentiment analysis and text mining approaches. First, we create an algorithm to automatically analyze the emotional polarity of a text and to obtain a value for each piece of text. Second, this algorithm is combined with K-means clustering and support vector machine (SVM) to develop unsupervised text mining approach. We use the proposed text mining approach to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span. The data sets used in our empirical studies are acquired and formatted from Sina sports forums, which spans a range of 31 different topic forums and 220,053 posts. Experimental results demonstrate that SVM forecasting achieves highly consistent results with K-means clustering. The top 10 hotspot forums listed by SVM forecasting resembles 80% of K-means clustering results. Both SVM and K-means achieve the same results for the top 4 hotspot forums of the year.
© 2009 Elsevier B.V. All rights reserved.

1. Introduction

In the Internet and information Age, online data usually grows in an exponential explosive fashion. The majority of these web data is in unstructured text format that is difficult to decipher automatically. Other than static WebPages, unstructured or loosely formatted texts often appears at a variety of tangible or intangible dynamic interacting networks [2,4,16,34]. A variety of heterogeneous online communities, societies and forums embody the interacting networks nowadays. When faced with tremendous amounts of online information from various online forums, information seekers usually find it very difficult to yield accurate information that is useful to them. This has motivated the research on identification of online forum hotspots, where useful information are quickly exposed to those seekers. Our research is to provide a comprehensive and timely description of the interacting structural natural groupings of various forums, which will dynamically enable efficient detection of hotspot forums, thus benefit Internet social network members in the decision making process.

As efficient business intelligence methods, data mining and machine learning provide alternative tools to dynamically process large amounts of data available online. Another most recent technique called sentiment analysis, also referred to as emotional polarity computation, has always been simultaneously employed when conducting online text mining. The purpose of text sentiment analysis is to determine the attitude of a speaker or a writer with respect to some specific topic. The attitude can be any forms of judgment or evaluation, the emotional state of the author when writing, or the intended emotional communication. It is recognized that the performance of sentiment classifiers are dependent on domains or topics [22].

In this paper, online forums hotspot detection and forecast are studied using sentiment analysis and text mining approaches. We develop this approach in two stages: emotional polarity computation and integrated sentiment analysis based on K-means clustering and support vector machine (SVM).The proposed unsupervised text mining approach is used to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span. Data are collected from Sina sports forums (webite: http://bbs.sports.sina.com.cn/treeforum/App/list.php?bbsid=33&subid=0), which include a range of 31 different topic forums and 220,053 posts. Computation indicates that within the same time window, SVM forecasting achieves highly consistent results with K-means clustering.

The rest of the paper is organized as follows. Section 2 discusses related work of our study. Section 3 presents models and methodology. Empirical results and discussion are given in Section 4. Finally, Section 4 concludes the paper.

* Corresponding author. RiskLab, University of Toronto, Canada.
E-mail addresses: dash@risklab.ca, dash@ru.is (D.D. Wu).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.09.003

2. Related work

This section investigates three streams of related work: dynamic cluster analysis of online forums, sentiment analysis of web documents and web text mining using machine learning.

2.1. Dynamic cluster analysis of online forums

Online forums are usually related to each other due to two reasons. First, strong commonalities are shared by forums with similar topics or themes. For example, within an entertainment society, the Academy Awards forum might be highly correlated to the Golden Globes Award forum. Secondly, emerging events will trigger a temporary correlation between certain forums. For example, the movie "No Country for Old Men" won "Best Motion Picture of the Year" in the 2008 Academy Awards, which rendered noticeable connections between the corresponding forums during the Oscar season. We aim to study the second inter-forum correlation and propose a mathematical approach to dynamically capture, describe and predict these time-varying correlations.

Extensive research work has been conducted upon various types of interacting social networks such as dynamic networks upon individuals, industrial manufacturers, listed companies, and online virtual communities [2,4,16,34,41–43]. One pioneer work from the Doctoral Thesis of Asavathiratham at MIT in 1996 [2] created an influence model as a tractable representation for the dynamics of networked Markov chains. This work has been utilized by several scholars, e.g., [4], where tools are developed to automatically and unobtrusively learn the social network structure that arises within a human group based on wearable sensors. [34] chose 662 main ceramic manufacturers in Guangdong Foshan ceramic industry cluster to construct a Competition Relationship Network (CRN) and proved that the network defined by competition relationship is a highly clustered scale-free network. Besides, correlated listed company network in stock markets constitutes another important research area in both academia and industry [42,43]. Regarding network dynamics of online virtual societies and communities, [16] proposed a relationship algebra used for various interesting computations on a social network weaved in the virtual communities.

It is observed that limited work was done to depict timely dynamics of online sports communities. Online sports forums within a virtual society are the focus of our study, where machine learning is used to dynamically depict the interacting structure and to cluster the forums according to their emotional polarity.

2.2. Sentiment analysis of web documents

There are a variety of metrics to classify web documents, including topics, structures, authors, time and so forth. Text classification based on its emotional polarity has become a newly-emerged frontier appealing to the web mining community. To illustrate how it works, suppose you are considering a vacation in city C, you might use a search engine online such as Google, and shoot the query "C". It would be handy to know what fraction of the matches Google returns recommends C as a travel destination [18]. Incorporating sentiment analysis into search engine and text retrieval technologies enables a more efficient and functional service for users [45]. Sentiment analysis has been utilized in applications such as news tracking and summarizing, online forums, file sharing, chatting rooms, blogging etc. Youtube introduced sentiment classification technology early this year to categorize all its comments into "Poor" or "Good" [44].

As a promising research area, text sentiment analysis has been extensively studied [1,3,26–28,33,35], where sentiment analysis is used for text classification tasks [8,13,14,40]. Existing sentiment calculation approaches fall into two types: machine learning based approach [3,33] and semantic orientation based approach [1,26–28,33,35]. Languages that have been studied include English [3,13,26–28], Chinese [33,35] and Arabic [1]. Our research aims to further extend the application of text sentiment analysis into cluster analysis for network dynamics of online communities, preliminarily Chinese sports forums.

2.3. Web text mining using machine learning

To conduct clustering and forecasting of online forum hotspots, we use two machine learning approaches: K-means and SVM. K-means has been studied and applied in a wide range of domains, e.g., bioinformatics [10–12,39], information security [36], pattern recognition [6,7,19], text classification [22]. In addition, various derivatives of conventional K-means algorithm have been developed [5,9,31]. Based on statistical learning, SVM is able to overcome problems such as over-fitting and local minimum to achieve high generalization [21,29,30,37,38]. Application of SVM includes text classification [15], image processing [24], and time series analysis [25,32]. In our study, machine learning is the key bridge between emotional polarity data and network dynamics. All the forums of Sina sports community form our research targets, each one of which will be converted into a vector representing its user attention within the current time window, in forms of number of posts and average value of sentiment. Those vectors acquired after feature extraction will be fed into the machine learning models for both clustering and forecasting.

3. Models and methodology

Our approach is mainly composed of the following steps: data collection and cleansing, text sentiment calculation and marking, hotspot detection based on K-means clustering and hotspot forecast based on SVM classification. Fig. 1 depicts the conceptual diagram of our approach, where three modules are defined to integrating text sentiment calculation, K-means and SVM for analyzing forum hotspots.

Module 1 is to convert Chinese texts into value based data through text sentiment computation and analysis. In this module, a new key word based approach is introduced to calculate the sentiment value for each piece of text by use of the commercial Java library developed by Lietu Enterprise Search and the HowNet lexicons. Our approach will yield an integer value for each post, with the sign showing its emotional polarity and the absolute value its emotional intensity.

Based on the sentimental values from Module 1, Module 2 applies K-means into all the forums of Sina sports community to calculate cluster values in each period, i.e., t1, t2,…tT, where T is the length of time cycle under consideration. In our K-means module, there are five inputs: The number of topic posts, the average number of responses of topic posts, the average text sentiment value of topic posts, the proportion of positive posts among all the topic posts and the proportion of negative posts among all the topic posts. Hotspot forums are identified by K-means as those closest to the theoretical centers of those clusters. This route generally follows previous work [30,31]. Module 2 is SVM-based classification module, which utilizes forum performance-related data and yielded cluster values to train machine learning model and apply the trained machine learning model to new forums for hotspot identification. K-means clustering results are fed into SVM model as the supervised learning outputs. As can be seen, our integrated approach differs from existing sentiment calculation work, which is either based on machine learning [3,33] or semantic orientation [1,26–28,33,35]. In fact, we aggregate both semantic orientation and machine learning tools and further extend the application of text sentiment analysis into cluster

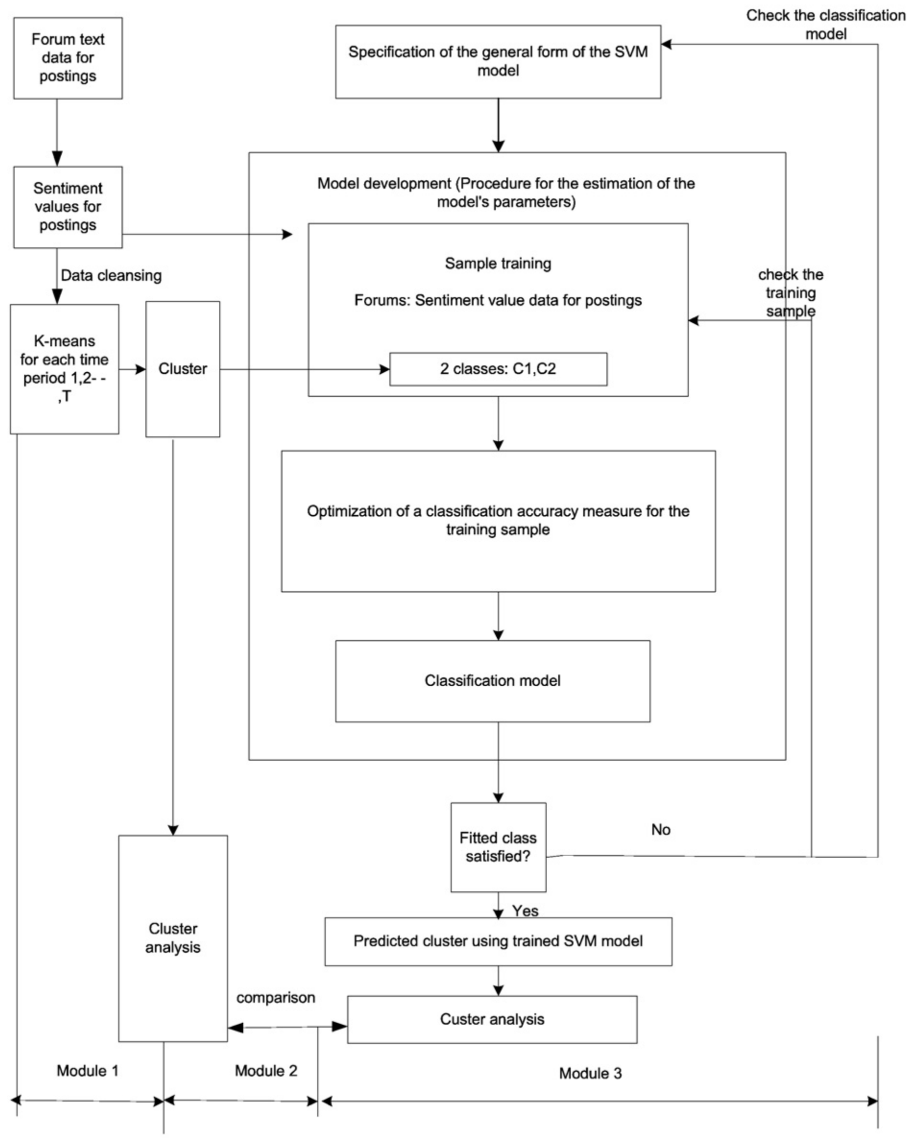

Fig. 1. Conceptual diagram of our approach.

analysis for network dynamics of online communities. This unique approach also combines the Lietu Enterprise Search Java library and the HowNet lexicons, which are applied to a unique problem of Chinese sports forums detection and prediction.

3.1. Data collection

Before data crawling and cleansing process are initiated, a comprehensive view of the structure of Sina sports community is necessary. Online Sina sports community exhibits a tree-like structure with root forums, branch forums and a nonseparable bottom layer of leaf forums. There are in total 49 leaf forums for this community. Fig. 2 illustrates the tree-like structure of the Sina sports community, where the root node, red circle node and yellow rectangular node represent the whole community, the first layer forums and the leaf forums respectively.

We proceed with the data crawling and cleansing process in the following four steps:

Step 1. *Manually create table SINA_LEAFORUM_URLLIST*

In this step, we manually store the information for all the 49 forums into a table named SINA_LEAFORUM_URLLIST in the database, where their names and URL links are contained.

Step 2. *Create table SINA_FORUM_URL based on SINA_LEAFORUM_URLLIST*

After the acquisition of the links for all the leaf forums, we parse the first pages of them in depth and generate a list of URLs of web pages that contain all the topic posts and the comment posts. The list will be written into the SINA_FORUM_URL table in the database.

Step 3. *Traverse the links in the SINA_FORUM_URL table and crawl down all the posts*

This step is to traverse through all the links that are in the SINA_FORUM_URL table, to parse out all the topic and comment posts contained on the corresponding web pages, and to store them into two tables of SINA_FORUM_TOPIC_POST and SINA_FORUM_COMMENT_POST. Two parsing templates are designed in XML format to parse the posts, which are SinaSportForumReplyPostParseTemplate.xml and SinaSportForumTopicPostParseTemplate.xml. Fig. 3 demonstrates the crawling procedure and the structures of the relational tables and the XML templates, where the green highlighted item in the tables are the primary keys.

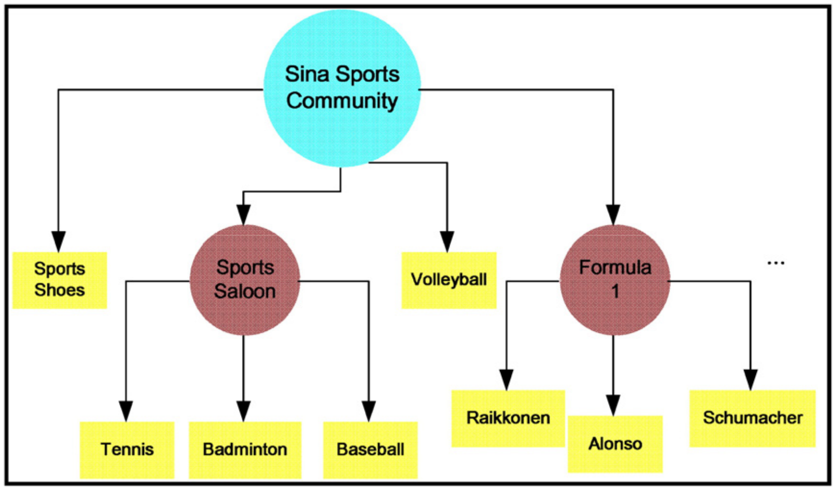

Fig. 2. The tree-like structure of Sina sports community.

Step 4. *Data cleansing*
When the crawling process is accomplished, data cleansing process is applied to the downloaded post sets. In this phase, we manually remove noise data and irrelevant data. Noise data include forums with strange picture/video postings that are not clearly shown online. Irrelevant data are from forums where there are not enough postings or posting contents that are not related to the forum topics at all. After removing noisy data and outliers, the set of leaf forums is narrowed down to 31, with a time span of 52 time windows across the year of 2007 and each time window is of a week length.

3.2. Text sentiment computation of forum posts

In this section, semantic orientation based approach will be developed using a new algorithm by adding up the sentiment values for all key words to achieve the sentiment value for the whole article. Text sentiment analysis is aimed at calculating an integer value for each piece of text, the absolute value of which represents the influential power and the sign of which denotes its emotional polarity.

Suppose the current post is p, since it is written in Chinese, we first utilize computer-based automatic word segmentation tool to decompose p into an array of key words $\{w_1, w_2, w_3, \ldots, w_n\}$, where there are n of them in total. Each key word $w_i(i=1,2,3,\ldots,n)$ will be assigned a sentiment

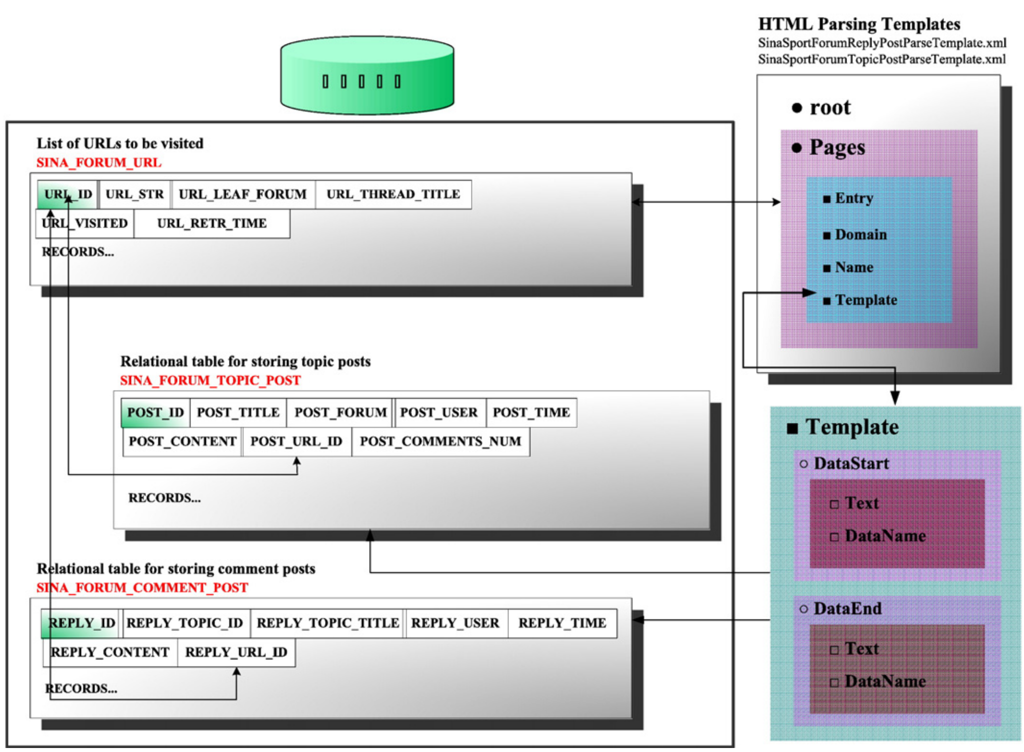

Fig. 3. Parsing links in table SINA_FORUM_URL to generate tables SINA_FORUM_TOPIC_POST and SINA_FORUM_COMMENT_POST.

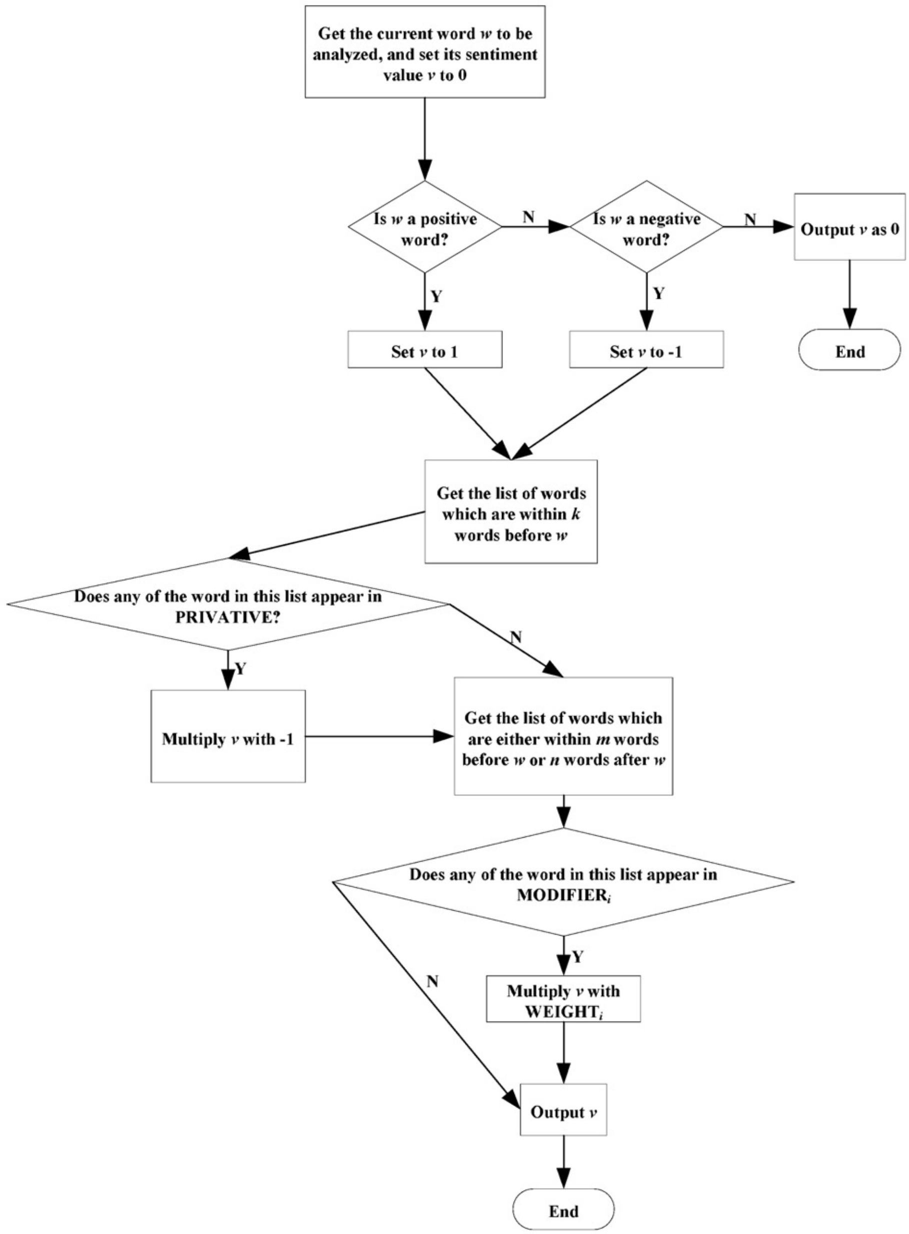

Fig. 4. Calculation of the sentiment value v for key word w based on word lists derived from HowNet.

value v_i by our proposed algorithm, while the sentiment value for p is the sum of the sentiment values for all the key words. Let V_p denote the sentiment value for p and we have

$$V_p = \sum_{i=1}^{n} v_i. \tag{1}$$

Calculation of the sentiment value array $\{v_1, v_2, v_3, \ldots, v_n\}$ is based on key words comparison and matching. In order to calculate the sentiment value for each key word contained in p, a comprehensive Chinese dictionary consisting of a complete list of sentiment-labeled words and phrases is entailed. In our work, the beta version of Chinese word sets with sentiment labels released by HowNet[1] on October 22, 2007 is utilized, and from which we derive eight word lists in Chinese. These eight lists are: positive Chinese words (POSITIVE), negative Chinese words (NEGATIVE), Chinese privatives (PRIVATIVE) and five lists of Chinese modifiers, with different emotional intensities. These five modifier lists are named as MODIFIER$_i$ ($i = 1,2,3,4,5$), each of which is assigned a value WEIGHT$_i$ ($i = 1,2,3,4,5$) denoting its sentimental intensity. The procedure of the calculation for the sentiment value of a key word is described in Fig. 4.

[1] http://www.keenage.com/html/e_index.html.

3.3. Hotspot detection using K-means clustering

As aforementioned, the 31 selected leaf forums will undergo feature extraction process. Each forum can be treated as a data point in a vector space. K-means clustering is applied to these 31 data points to obtain a cluster natural groupings description for all time windows in the year 2007. Again, each time window has a length of a week.

Suppose that the current time window is W_i and the 31 leaf forums are denoted as $\{F_1, F_2, F_3,\ldots, F_{31}\}$. During the feature extraction process, we use a vector $V^i(j)$ to represent the emotional polarity or quantification of user attention of any forum $F_j(j=1,2,3\ldots,31)$ within the time span W_i. The data set used as the input of the K-means clustering in W_i is denoted as $\{V^i(1), V^i(2), V^i(3),\ldots,V^i(31)\}$, which will be clustered into k groups. $V^i(j)$ is composed of five elements: the number of the topic posts in F_j within W_i, the average number of responses of topic posts, the average sentiment value of topic posts, the fraction of positive posts among all the topic posts, and the fraction of negative posts among all the topic posts. We denote these five elements by $\mathrm{NUM}^i(j)$, $\overline{RESPONSE}^i(j)$, $\overline{SENTIMENT}^i(j)$, $\mathrm{POS\_PERC}^i(j)$ and $\mathrm{NEG\_PERC}^i(j)$. Mathematically, we can express $V^i(j)$ as:

$$V^i(j) = \begin{pmatrix} NUM^i(j) \\ \overline{RESPONSE}^i(j) \\ \overline{SENTIMENT}^i(j) \\ POS\_PERC^i(j) \\ NEG\_PERC^i(j) \end{pmatrix}. \tag{2}$$

Eq. (2) displays the structure of the representation vector for leaf forums after feature extraction. The transformed vectors are used as the inputs to K-means model. For each W_i, with a given k, a clustering view of all the 31 leaf forums is obtained by the K-means algorithm, with a center forum for each cluster. The hotspot forums are those closest to the theoretical centers of the clusters. For each time window, the clustering result by K-means is presented in a vector containing 31 elements, and each of which is an integer value of either 1 or 0, with 1 denoting a hotspot while 0 a non-hotspot.

3.4. Hotspot detection using SVM

Apart from K-means clustering, SVM is utilized in this section to realize hotspot forecasting. SVM forecasts the clustering view of the leaf forums in a sliding time window manner, whose results will be compared to those from K-means.

A sliding time window that goes through the whole experiment time span distinguishes the SVM-based approach from the K-means-based one. In order to forecast the hotspot distribution within the current time window, we fed into the SVM model with the historical data we obtain from the last time window. As for the output of the SVM, which serves as the supervised learning tool in our work, the clustering result by the K-means approach within the current time window is used. A well-trained SVM is utilized to carry out prediction for the next time window, by inputting the data obtained from the current one. Suppose there are T time windows, $\{W_1, W_2, W_3,\ldots,W_T\}$, and the current one is W_i. If a forecast for W_{i+1} is expected, we first train a SVM by inputting forums' representation vectors of W_{i-1} and setting the output as the clustering result for W_i by K-means. Then the trained SVM generates classification outputs for data of W_i. Finally, SVM result is compared to the K-means clustering result for data of W_{i+1}.

For each SVM, the input is a matrix containing 31 leaf forums' representation vectors, and the output is a vector containing 31 integer values either 1 or 0 with 1 representing a hotspot and 0 a non-hotspot. Each training and testing sample corresponds to a leaf forum. Computation based on SVM involves both the training process and test process. In the training process, we use a matrix tuple $<I, O>$, where I and O denotes input and output training sample data to SVM. Mathematically, we have

$$I = \begin{pmatrix} V^{i-1}(1) \\ V^{i-1}(2) \\ \cdots \\ V^{i-1}(j) \\ \cdots \\ V^{i-1}(31) \end{pmatrix} = \begin{pmatrix} NUM^{i-1}(1), \overline{RESPONSE}^{i-1}(1), \overline{SENTIMENT}^{i-1}(1), POS\_PERC^{i-1}(1), NEG\_PERC^{i-1}(1) \\ NUM^{i-1}(2), \overline{RESPONSE}^{i-1}(2), \overline{SENTIMENT}^{i-1}(2), POS\_PERC^{i-1}(2), NEG\_PERC^{i-1}(2) \\ \cdots \\ NUM^{i-1}(j), \overline{RESPONSE}^{i-1}(j), \overline{SENTIMENT}^{i-1}(j), POS\_PERC^{i-1}(j), NEG\_PERC^{i-1}(j) \\ \cdots \\ NUM^{i-1}(31), \overline{RESPONSE}^{i-1}(31), \overline{SENTIMENT}^{i-1}(31), POS\_PERC^{i-1}(31), NEG\_PERC^{i-1}(31) \end{pmatrix} \tag{3}$$

and

$$O = \begin{pmatrix} L^i(1) \\ L^i(2) \\ \cdots \\ L^i(j) \\ \cdots \\ L^i(31) \end{pmatrix}, \tag{4}$$

where $L^i(j)$ in O denotes the clustering result for F_j in W_i by K-means. If $L^i(j)=1$, K-means labels F_j is a hotspot, while if $L^i(j)=0$ K-means labels F_j is a non-hotspot. Similarly, let $<I', O'>$ denote the input and output matrix of SVM in the test process and we have

$$I' = \begin{pmatrix} V^i(1) \\ V^i(2) \\ \cdots \\ V^i(j) \\ \cdots \\ V^i(31) \end{pmatrix} = \begin{pmatrix} NUM^i(1), \overline{RESPONSE}^i(1), \overline{SENTIMENT}^i(1), POS\_PERC^i(1), NEG\_PERC^i(1) \\ NUM^i(2), \overline{RESPONSE}^i(2), \overline{SENTIMENT}^i(2), POS\_PERC^i(2), NEG\_PERC^i(2) \\ \cdots \\ \mathrm{NUM}^i(j), \overline{RESPONSE}^i(j), \overline{SENTIMENT}^i(j), POS\_PERC^i(j), NEG\_PERC^i(j) \\ \cdots \\ NUM^i(31), \overline{RESPONSE}^i(31), \overline{SENTIMENT}^i(31), POS\_PERC^i(31), NEG\_PERC^i(31) \end{pmatrix} \tag{5}$$

and

$$O' = \begin{pmatrix} L^{i+1}(1)' \\ L^{i+1}(2)' \\ \cdots \\ L^{i+1}(j)' \\ \cdots \\ L^{i+1}(31)' \end{pmatrix}, \tag{6}$$

where $L^{i+1}(j)'$ in O' represents the binary classification result for F_j in W_i by SVM. If $L^{i+1}(j)'=1$, SVM classifies F_j as a hotspot, while F_j is classified as non-hotspot if $L^{i+1}(j)'=0$. Comparative study is carried out between O' and the clustering result by K-means in W_{i+1}.

4. Empirical results and discussion

4.1. Data preparation

The data preparation for the empirical studies primarily includes three tasks: data downloading, data cleansing and data statistics. The data sets used in our experiments are crawled down and compiled from the Internet by an automatic crawling Java program, which consists of two major modules: the target URL list generating module and the HTML page parsing module. We choose to conduct our experiments on Sina sports community because this is the most popular and prestigious online sports community in China. The aforementioned crawler crawled down a complete set of posts in the form of both topics and responses from Sina sports community. This was done within the time span from the time this community was founded until February 2008. The data view before any cleansing and filtering process is demonstrated in Table 1.

When the crawling is done, noticeable inconsistency and noise of post data entail cleansing and filtering process. A common time span T is expected, in which the vast majority of the forums have sufficient data distribution. The cleansing process includes the following six steps.

Step 1. Segment the continuous time line for data into time windows.
Step 2. Determine the optimal value for T.
Step 3. Get a subset F of the 49 forums which have dense data distribution within T.
Step 4. Generate a new post set P that falls within the range defined by both T and F.
Step 5. Calculate the text sentiment for all the posts in P.
Step 6. Create the new view for cleansed data sets.

The data view after cleansing and filtering phase is also demonstrated in Table 1, where 200701 and 200752 stand for the 1st and 52nd week of the year 2007. The reason that the size of data increases after cleansing is that the results of word segmentation are written back to the database. Because the posts to be analyzed are written in Chinese, word segmentation constitutes a significant prerequisite step in text sentiment computation. The word segmentation software tool used in our experiment is the commercial Java library developed by Lietu Enterprise Search[2]. It is demonstrated in Table 1, where we show a whole experiment time span from the first week of 2007 till the last week of 2007. 31 out of the 49 leaf forums are selected as the final set of leaf forums under study. Only topic posts are taken into consideration during the preliminary experiment.

We also calculate aggregated statistics in order to get a preliminary intuitive view for user attention of the selected 31 forums within the year 2007. Table 2 shows the average number of topic posts and the average number of responses for the 31 forums spanning across the 52 time windows of the year 2007[3]. According to the number of topic posts, the most popular forums among users include "Basketball—Yao Ming", Soccer Tycoons—AC Milan", "Basketball—NBA", "Soccer Tycoons—Milan International", etc. The most popular forums based on the average response number include "Soccer Tycoons—Real Madrid", "Soccer Tycoons—Juventas", "Soccer Tycoons—Milan International", "Soccer Tycoons—FC Barcelona", etc.

4.2. Text sentiment calculation for topic posts using HowNet dictionaries

As previously mentioned, text sentiment computation is a key step in our empirical studies. We use the Chinese lexicons resourced from HowNet,[4] an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents, to form up eight key word lists. Results are described in Table 3. These eight lists correspond to those introduced in Section 3.

Based on the algorithm depicted in Section 3, we conduct sentiment calculation for the 220,053 posts by the eight word lists in Table 3. Since the posts to be analyzed are written in Chinese, word segmentation constitutes a significant prerequisite step in text sentiment computation. The word segmentation software tool used in our experiment is the commercial Java library developed by Lietu Enterprise Search. The Lietu tool not only converts the text into an array of words, but also tags each of the words with its part of speech.

[2] http://www.lietu.com/demo/index.htm.
[3] The original language for all the forum names is Chinese.
[4] http://www.keenage.com/html/e_index.html.

Table 1
The data view of collected post data from Sina sports community.

	Post type	Number of posts	Size of data/KB	Starting time	Ending time	Number of forums
Before data cleansing	Topic post	510,218	616,640	1999-03-08 11:13:20	2008-01-02 00:22:29	49
	Comment post	5,565,216	1,978,632	2003-08-01 22:33:52	2008-02-16 14:36:25	49
After data cleansing	Topic posts	220,053	1,210,112	2007-01-01 00:00:48	2007-12-31 23:59:59	31

4.3. Computation using K-means clustering

In this section, we conduct K-means clustering among the 31 selected leaf forums for each time window in 2007, based on their emotional polarity. Text sentiment analysis is employed to calculate the emotional polarities for all the posts. We will use K-means to achieve a clustering view for all the 31 forums within each time window over the year 2007, which generates in total 52 clustering results. One deficiency of K-means is that a predetermined value of k is required. To overcome this drawback, K-means cluster analysis is conducted for a set of k values ranged from 5 to 20. The forums yielding the smallest Euclidean distances to the centers of clusters are considered as hotspot forums within the current time window. Multiple metrics are employed to analyze the clustering results from a wider spectrum of perspectives.

We will examine the clustering results by K-means from the following two perspectives. First, we present the clustering natural groupings for each time window. Second, we show the results on a forum basis by presenting the emotional polarity each forum gets over the year 2007.

4.3.1. Clustering results shown on a time window basis

Table 4 demonstrates part of the clustering results by K-means in the year 2007, when k is set from 5 to 7. As before, "200701" stands for the first time window of 2007. The forums listed in the table are those closest to the theoretical cluster centers and denote the hotspot forums selected by K-means. The naming rules for those forums are the same as in Table 2.

Table 2
Post data statistics upon selected 31 forums of Sina sports community over the year 2007.

Forum ID	Forum name	Average # of posts	Average # of Responses
1	Chinese Soccer—Care About Chinese Football	120	4.65336
2	Sports shoes	365	16.36289
3	Soccer Tycoons—Arsenal	59	13.97825
4	Soccer Tycoons—Juventas	213	28.2192
5	Basketball—Guangzhou Hongyuan	17	7.041931
6	International Soccer—Spanish Football League	22	11.2351
7	Soccer Tycoons—Liverpool	48	7.64624
8	Sports Saloon—Billiard	17	6.819567
9	Basketball—Chinese Basketball Association	67	15.03061
10	Sports Saloon—Tennis	21	9.869563
11	International Soccer—Italian Football League	41	14.39272
12	Soccer Tycoons—Chelsea	127	15.26819
13	The Game of Go	17	12.62573
14	Chinese Soccer—Dalian Shide	23	10.97456
15	International Soccer—German Football League	5	8.454958
16	Soccer Tycoons—AC Milan	474	19.30753
17	International Soccer—English Football League	47	10.31931
18	Chinese Soccer—Shandong Luneng	192	10.8492
19	Outdoor activities	21	1.878042
20	Soccer Tycoons—Milan International	375	27.68831
21	Football lottery	198	4.880246
22	Soccer Tycoons—Manchester United	189	22.66756
23	Basketball—NBA	456	16.47136
24	Basketball—Yao Ming	603	18.24459
25	Soccer Tycoons—A.S. Roma	46	10.11552
26	Sports Saloon—Table Tennis	14	11.62929
27	Soccer Tycoons—FC Barcelona	61	23.70714
28	Volleyball	145	10.55084
29	Soccer Tycoons—Real Madrid	100	40.08629
30	Soccer Tycoons—Bayern Munchen	16	7.178974
31	Chinese Soccer—China Super League of Football	120	6.833243

Note: the "—" in a forum name separates the leaf forum name from the root forum name.

4.3.2. Clustering results shown on a forum basis

In addition to observing hotspot distribution on a timely basis, we further inspect the hotspot distribution among forums, i.e. which forums tend to get more user attention over the year than the others. We propose a method to measure the degree H_j to which the forum F_j gets attention, which counts the number of times F_j is considered as a hotspot, over all the 52 time windows as well as over all the values of k. Usually a larger value of H_j indicates a higher popularity for F_j in year 2007. Fig. 5 is a visualization of the hotspot distribution of the 31 forums in the year 2007 achieved by K-means clustering, with the vertical axis showing their degree values and a higher value of the degree implying a higher user attention.

As shown by Fig. 5, the hotspot degree values for forums span from 226 to 469. Based on the statistics visualized in Fig. 5, we list in Table 5 the top 10 most popular forums in Sina sports community by K-means clustering over the year 2007.

It is observed that the hot forum set determined by K-means clustering is highly consistent with the set obtained in Section 4.1 after intuitive statistics is calculated. The most popular sports topics among Chinese users include basketball, soccer etc. Forums such as Basketball—Yao Ming, Soccer Tycoons—AC Milan, Soccer Tycoons—Chelsea are substantiated to be hotspot forums over the year 2007 by both approaches. Therefore, our approach incorporating K-means clustering and text sentiment analysis is sufficient to provide helpful information for users to get a good mastery of the hotspot ranking and distribution of Sina sports community.

4.4. Computation using SVM classification

During this phase of experiment, we apply SVM-based binary classification to forecast the hotspot distribution among the selected 31 forums of Sina sports community. The analysis is similar to Section 4.3. SVM-based approach forecasts the clustering natural groupings for the future time window by using the data from the past time window. SVM achieves a clustering result by classifying each forum as either hotspot forum or non-hotspot forum, thus converting the clustering task into a binary classification task.

Table 3
Eight Chinese key word lists based on HowNet online knowledge base.

Created Chinese key word lists	Description
POSITIVE	4566 Chinese words with positive sentimental polarity
NEGATIVE	4370 Chinese words with negative sentimental polarity
PRIVATIVE	14 Chinese privatives, manually collected
The following are five lists of Chinese modifiers, with a decrement in their intensities	
MODIFIER_1	85 modifiers, with a weight value 10
MODIFIER_2	42 modifiers, with a weight value 8
MODIFIER_3	37 modifiers, with a weight value 6
MODIFIER_4	29 modifiers, with a weight value 4
MODIFIER_5	12 modifiers, with a weight value 2

Table 4
Clustering results by K-means for the three time windows.

Time window	$k=5$	$k=6$	$k=7$
200701	1. Soccer Tycoons—Juventas 2. Basketball—Guangzhou Hongyuan 3. Outdoor activities 4. Soccer Tycoons—Milan International 5. Basketball—Yao Ming	1. Soccer Tycoons—Liverpool 2. Chinese Soccer—Care About Chinese Football 3. Sports Saloon—Tennis 4. The Game of Go 5. Chinese Soccer—Shandong Luneng 6. Basketball—Yao Ming	1. Sports shoes 2. Soccer Tycoons—Juventas 3. Sports Saloon—Tennis 4. International Soccer—Italian Football League 5. The Game of Go 6. International Soccer—English Football League 7. Basketball—Yao Ming
200702	1. Sports shoes 2. Soccer Tycoons—Liverpool 3. Soccer Tycoons—Chelsea 4. International Soccer—English Football League 5. Chinese Soccer—Shandong Luneng	1. Sports shoes 2. International Soccer—Italian Football League 3. Chinese Soccer—Shandong Luneng 4. Outdoor activities 5. Basketball—Yao Ming 6. Chinese Soccer—China Super League of Football	1. Sports shoes 2. International Soccer—German Football League 3. Chinese Soccer—Shandong Luneng 4. Outdoor activities 5. Soccer Tycoons—Milan International 6. Basketball—Yao Ming 7. Chinese Soccer—China Super League of Football
200703	1. Sports shoes 2. Soccer Tycoons—Chelsea 3. Soccer Tycoons—Manchester United 4. Soccer Tycoons—FC Barcelona 5. Chinese Soccer—China Super League of Football	1. Sports shoes 2. Soccer Tycoons—Chelsea 3. The Game of Go 4. Chinese Soccer—Dalian Shide 5. Soccer Tycoons—Manchester United 6. Soccer Tycoons—FC Barcelona	1. Soccer Tycoons—Juventas 2. International Soccer—Spanish Football League 3. Sports Saloon—Billiard 4. Soccer Tycoons—Milan International 5. Soccer Tycoons—Manchester United 6. Basketball—Yao Ming 7. Chinese Soccer—China Super League of Football

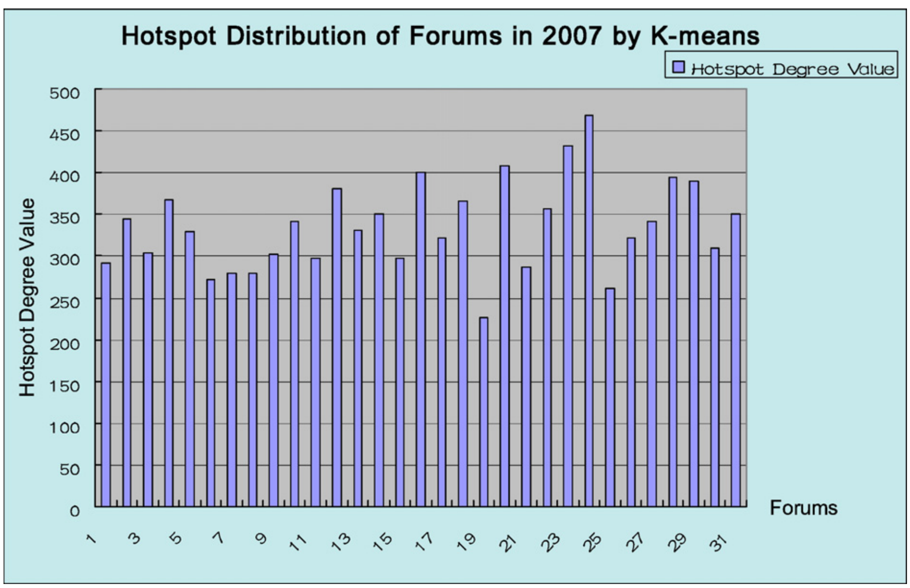

Fig. 5. The hotspot distribution of the 31 leaf forums using K-means.

As mentioned previously, each training and forecasting cycle of SVM classification strides over three time windows, rendering a total time span for the SVM forecasting starting from the third week of 2007 till the last week of 2007. For each time window, the forecasting result achieved by SVM is compared to that by K-means in the next section. The SVM tool used in this experiment is the open source LIBSVM library[5] written in Java [20,23].

Similar to Section 4.3, we examine the forecasting results by SVM from two perspectives: a time window basis forecasting and a forum basis forecasting. Table 6 demonstrates part of the forecasting results for three time windows in the year 2007. The forums listed in the table are those forecasted as the hotspot forums by SVM. The k value in the first row is a parameter of the K-means method, which is used to enable supervised learning for SVM training. Note that there exists disparity between the k value of K-means clustering and the actual number of hotspots that are labeled by SVM.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Similar forum-based analysis to Section 4.3 is employed here to acquire a view from a forum perspective. The hotspot degree value is defined the same as before. Fig. 6 shows the hotspot distribution of the 31 forums in the year 2007 achieved by SVM forecasting, with the

Table 5
Top 10 popular forums in 2007 Sina sports community by K-means.

Forum ID	Forum name	Hotspot degree
24	Basketball—Yao Ming	469
23	Basketball—NBA	432
20	Soccer Tycoons—Milan International	408
16	Soccer Tycoons—AC Milan	401
28	Volleyball	395
29	Soccer Tycoons—Real Madrid	389
12	Soccer Tycoons—Chelsea	380
4	Soccer Tycoons—Juventas	367
18	Chinese Soccer—Shandong Luneng	366
22	Soccer Tycoons—Manchester United	356

Table 6
Forecasting results by SVM for the 27th till the 29th time window of the year 2007.

Time Window	$k=7$	$k=8$	$k=9$
200727	1. Sports shoes 2. Basketball—Chinese Basketball Association	1. Soccer Tycoons—Juventas 2. Soccer Tycoons—Liverpool 3. Sports Saloon—Tennis 4. Soccer Tycoons—Chelsea 5. The Game of Go 6. Soccer Tycoons—AC Milan 7. International Soccer—English Football League 8. Soccer Tycoons—Milan International 9. Soccer Tycoons—Manchester United 10. Basketball—Yao Ming 11. Soccer Tycoons—Bayern Munchen	1. Soccer Tycoons—AC Milan 2. International Soccer—English Football League 3. Basketball—NBA 4. Basketball—Yao Ming
200728	1. Soccer Tycoons—AC Milan 2. Basketball—NBA 3. Basketball—Yao Ming	1. Soccer Tycoons—AC Milan 2. Soccer Tycoons—Milan International 3. Basketball—NBA 4. Basketball—Yao Ming	1. Soccer Tycoons—AC Milan 2. Basketball—NBA 3. Basketball—Yao Ming
200729	1. International Soccer—Spanish Football League 2. Basketball—NBA 3. Soccer Tycoons—Bayern Munchen	1. Chinese Soccer—Care About Chinese Football 2. Soccer Tycoons—Arsenal 3. International Soccer—Spanish Football League 4. International Soccer—English Football League 5. Soccer Tycoons—Manchester United 6. Soccer Tycoons—Bayern Munchen	1. Chinese Soccer—Care About Chinese Football 2. Soccer Tycoons—Liverpool 3. Chinese Soccer—Dalian Shide 4. International Soccer—German Football League 5. Football lottery 6. Sports Saloon—Table Tennis 7. Soccer Tycoons—Bayern Munchen 8. Chinese Soccer—China Super League of Football

vertical axis showing their degree values. Again, a higher value of the degree implies a higher user attention.

As shown by Fig. 6, the hotspot degree values for forums span from 224 to 441. Based on the statistics visualized in Fig. 6, we list in Table 7 the top 10 most popular forums in Sina sports community by SVM forecasting over the year 2007. The results shown in this section further present a noticeable consistency with the results achieved by K-means clustering. It is clearly demonstrated, by both Table 5 and Table 7, that the two lists of top 10 most popular forums resemble each other to as much as 80%. On top of this, K-means and SVM provide the same results for the top 4 most popular forums in the year 2007, which are Basketball—Yao Ming, Basketball—NBA, Soccer Tycoons—Milan International and Soccer Tycoons—AC Milan. Therefore, a strong connection between text sentiment and hotspot distribution for online sports forums is confirmed by both techniques. This has verified the feasibility of detecting and forecasting hotspot forums with the aid of text sentiment analysis. Besides, SVM-based approach realizes a forecast for the next time window. Finally, in order to see a detailed SVM computation, we depict in Fig. 7 the distribution of hotspots with respect to time. In Fig. 7, the X and Y axis represent respective weeks in the total time horizon and corresponding hotspot forums identified by SVM. Note that a k value of 8 is used in K-means model. Fig. 7 generates the same result as in Table 6. Both Table 6 and Fig. 7 suggest during the 27th, 28th and 29th week, the number of hotspot forums identified by SVM is 11, 4 and 6 respectively.

4.5. Comparative Study between K-means and SVM

This section carries out a formal comparative study between K-means and SVM to validate model consistency using five widely used metrics: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) [17]. For a certain value of k, a comparative study is exerted for each one of the 50 time windows in 2007, which are the 50 time windows in the SVM-based

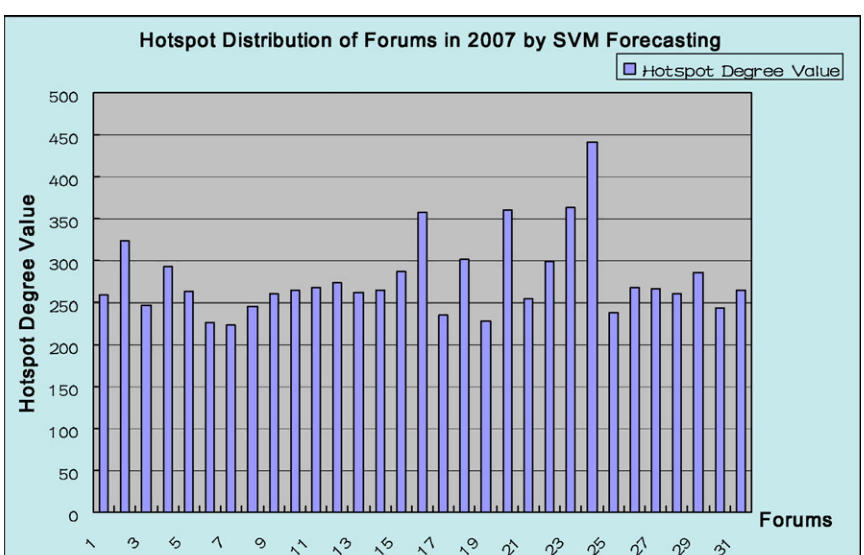

Fig. 6. The hotspot distribution of the 31 leaf forums in 2007 based on SVM.

Table 7
The top 10 most popular forums in Sina sports community by SVM forecasting over the year 2007.

Forum ID	Forum name	Hotspot degree value
24	Basketball—Yao Ming	441
23	Basketball—NBA	363
20	Soccer Tycoons—Milan International	361
16	Soccer Tycoons—AC Milan	358
28	Sports shoes	323
29	Chinese Soccer—Shandong Luneng	302
12	Soccer Tycoons—Manchester United	298
4	Soccer Tycoons—Juventas	293
18	International Soccer—German Football League	287
22	Soccer Tycoons—Real Madrid	285

experiment. Each time window corresponds to a set of these five metrics, which are defined as follows.

Definition 1. Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP denotes the number of forums that are estimated by both K-means and SVM as hotspots; TN denotes the number of forums that are estimated by both K-means and SVM as hotspots; FP denotes the number of forums that are estimated by SVM as hotspots whereas non-hotspots by K-means; FN denotes the number of forums that are estimated by SVM as non-hotspots whereas hotspots by K-means. Accuracy shows the fraction of forums that are classified into the same category by both K-means and SVM among all the forums.

Definition 2. Sensitivity

$$Sensitivity = \frac{TP}{TP + FN}. \tag{8}$$

Sensitivity shows the fraction of forums which are classified by SVM as hotspots among all forums that are labeled by K-means as hotspots.

Definition 3. Specificity

$$Specificity = \frac{TN}{TN + FP}. \tag{9}$$

Specificity shows the fraction of forums which are classified by SVM as non-hotspots among all forums that are labeled by K-means as non-hotspots.

Definition 4. PPV

$$PPV = \frac{TP}{TP + FP}. \tag{10}$$

PPV shows the fraction of forums which are labeled by K-means as hotspots among all the forums that are classified by SVM as hotspots.

Definition 5. NPV

$$NPV = \frac{TN}{TN + FN}. \tag{11}$$

NPV shows the fraction of forums which are labeled by K-means as non-hotspots among all the forums that are classified by SVM as non-hotspots.

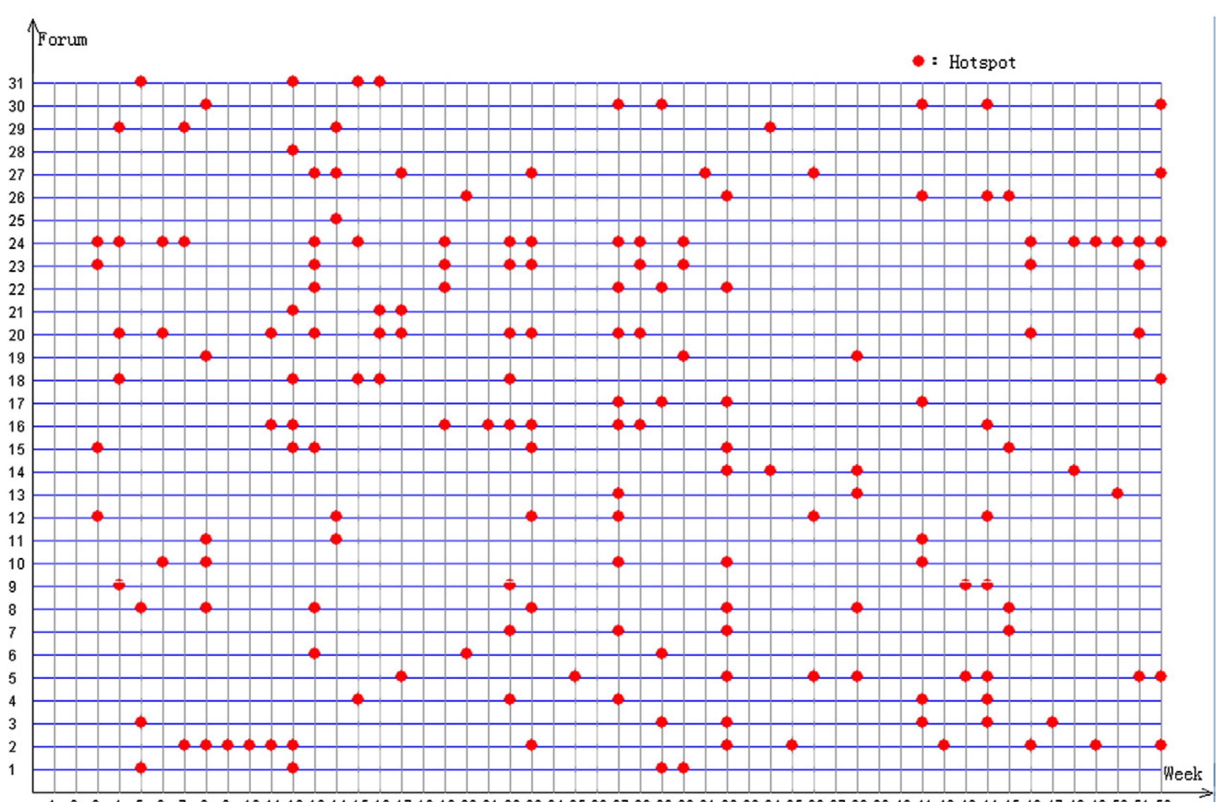

Fig. 7. Distribution of hotspots with respect to time.

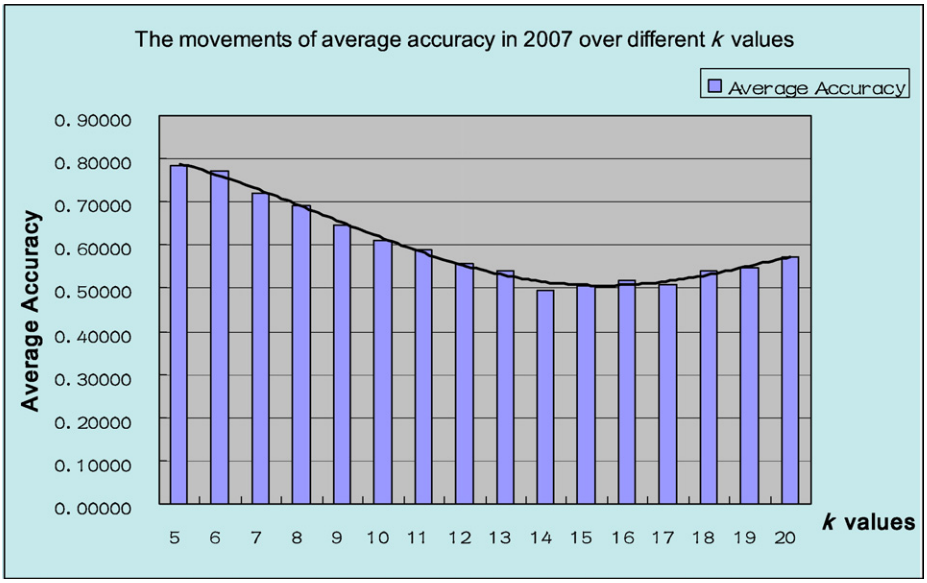

Fig. 8. The movements of average accuracy over different k values in the year 2007.

Using Formulae (7)–(11), we calculate the five metrics over different k values ranged from 5 to 20. To get a better visualization of the movements of the five metrics over different k values, we depict five metrics from Figs. 8–12. These figures visualize the movements of the average values of the five metrics over different k values.

It is clearly indicated through Figs. 8–Fig. 12 that, four measurements are monotonic functions of k except average accuracy. Fig. 8 suggests that our method is generally sufficient to achieve a satisfying result for accuracy, especially when k is set to a rather small value. During the subsequent experiments, larger values of k (from 20 to 25) are employed, and it is proved that average accuracy increases with k when k has reached a certain value. Therefore, the assumption of facilitating hotspot detection and prediction by machine learning techniques and sentiment analysis is justified. The rest of the four metrics provide evaluation from four other perspectives. Sensitivity is a critical evaluation measurement, which denotes the fraction of forums which are classified by SVM as hotspots among all forums that are labeled by K-means as hotspots. It is visualized in Fig. 9 that the average sensitivity for all time windows displays a monotonic increment over different k values, which is reasonable considering that a larger possibility is enabled for consistency when k is designated a relatively larger value. The fact that when k is larger than 17, a good result is obtained on a sensitivity level proves that under this setting, our SVM-based method is sufficient in capturing the majority hotspots, which are approved by K-means, for the immediate future. PPV constitutes another important measurement in our experiment, which denotes the fraction of forums labeled by K-means as hotspots among all the forums classified by SVM as hotspots. It is shown that when k is set smaller than 13, a good result is achieved on a PPV level, which indicates that the SVM forecasting results are more reliable when k is relatively small.

5. Conclusions and discussions

We have developed an algorithm to automatically analyze the emotional polarity of a text, based on which a value for each piece

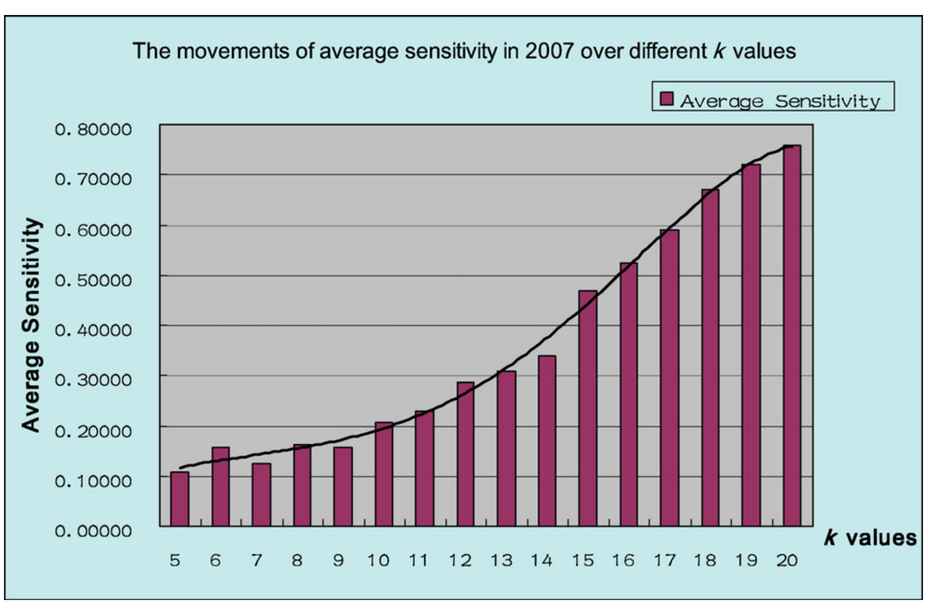

Fig. 9. The movements of average sensitivity over different k values in the year 2007.

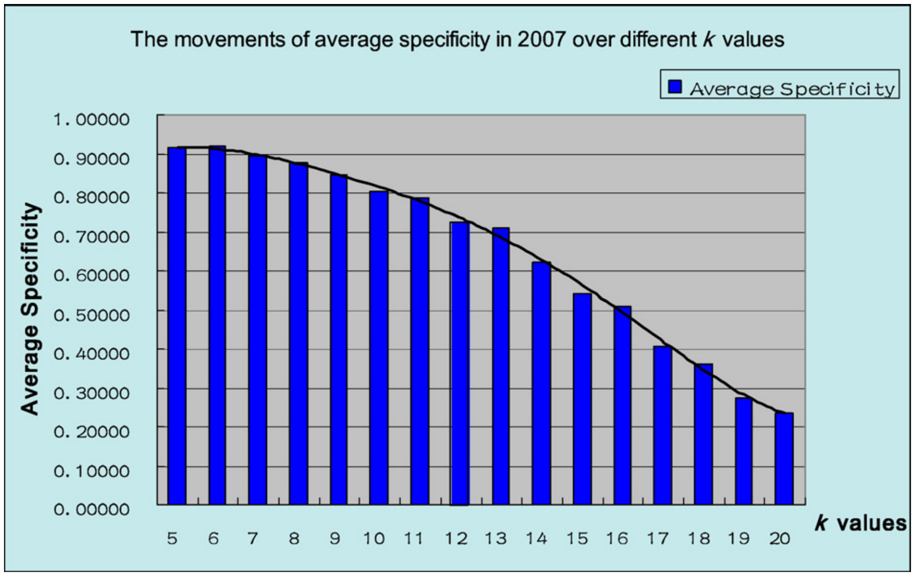

Fig. 10. The movements of average specificity over different *k* values in the year 2007.

of text is obtained. The absolute value of the text represents the influential power and the sign of the text denotes its emotional polarity. This algorithm is combined with K-means clustering and SVM classification to develop integrated approach for online sports forums cluster analysis. We apply unsupervised clustering algorithm to group the forums into various clusters, with the center of each cluster representing a hotspot forum within the current time span. In addition to clustering the forums based on data from the current time window, we also conduct forecast for the next time window. Empirical studies present strong proof of the existence of correlations between post text sentiment and hotspot distribution. Computation indicates both SVM and K-means produce consistent natural groupings results.

Companies, as information seekers can benefit from our hotspot predicting approaches in several ways. For example, marketing objectives at the marketing department of big retail stores such as Walmart should follow the same rules as the sales objectives, and be measurable, quantifiable, and time specific. However, in practice customers' behavior are always hard to be explored and captured. Using our hotspot predicting approaches can help the marketing department understand what their specific customers' timely concerns regarding goods and services information. Results generated from our approach can be also combined to market basket analysis to yield comprehensive decision support information.

A firm in financial sector or the financial department of a giant company may profit from such a sentimental and text mining process. In financial market, right before a security market opens and trading begins, analysts people on sales and trading desks usually try to get an overall fix on market sentiment and for particular investments. To get a feel for what will take place, decision makers used to make phone calls to trusted contacts, browse through news, morning reports and use other more quantitative tools. Our hotspot based semantic engine can aggregate the content in the forum and media feeds to determine whether stories on a particular company are positive, negative or neutral, and then generate simple data displays and charts that enable one to get a grasp on likely market sentiment for a company's security very quickly [17].Further work can be done based on this research. First, predicting hotspot using past data may not be accurate since

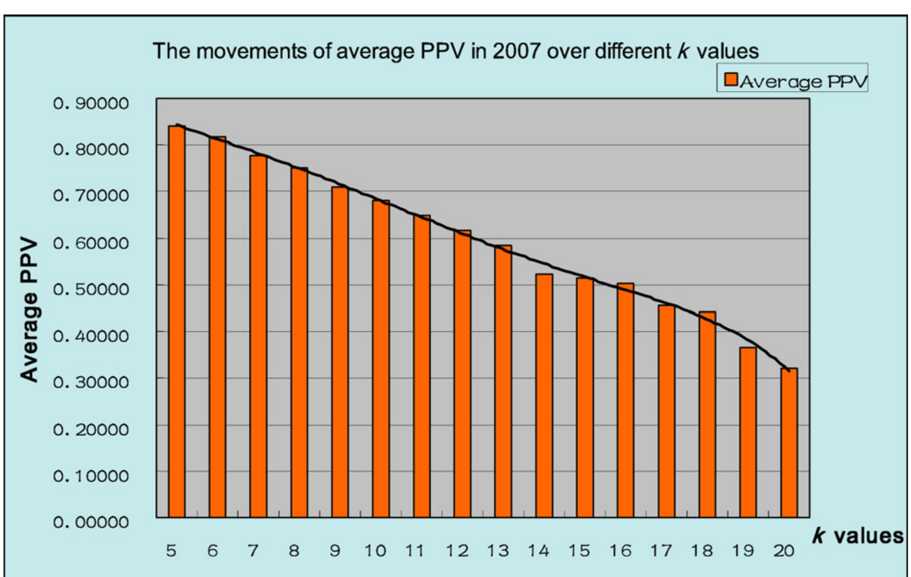

Fig. 11. The movements of average PPV over different *k* values in the year 2007.

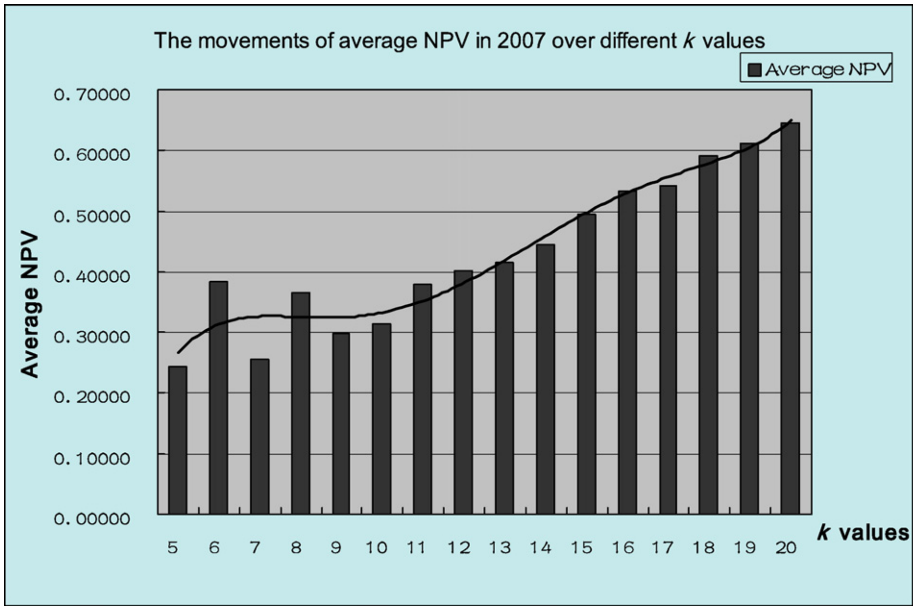

Fig. 12. The movements of average NPV over different *k* values in the year 2007.

many of the hotspots are emergent events that has no correlation with past hotspot history. Therefore, algorithm design can be improved to treat this problem and yield a more accurate calculation of sentiment. Regarding supervised learning, algorithms other than SVM, or variations of SVM, can be incorporated as well. Second, we can incorporate topic extraction. It is very natural to pop the question what event or topic triggered the user attention after a hotspot is detected. Currently our model is not able to provide analysis in this aspect, which entails a thorough exploration of topic extraction for hotspots in the future. Third, a practical system, in the form of a website portal, is desired as our major future work. The system is expected to possess the following functions.

a) Users are able to observe the hotspot forum distribution and its natural groupings by inputting a time span.
b) Users are able to forecast the hotspot forum distribution and its natural groupings for the immediate future time windows.
c) Users are able to choose the ways hotspot detection results are visualized.
d) Users are able to choose among different clustering or forecasting algorithms.
e) Users are able to further inspect the posts and their sentiments for any detected hotspot forum.
f) Users are able to extract the topics of hotspot forums based on their posts.
g) Users are able to calculate the sentiment value for any post they choose.

References

[1] K. Ahmad, Y. Almas, Visualising sentiments in financial texts? Proceedings of the Ninth International Conference on Information Visualisation (2005) 363–368.
[2] C. Asavathiratham, The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains, *Dept. of EECS*. 2000, MIT, Cambridge, 2000, p. 188.
[3] P. Chaovalit, L. Zhou, Movie review mining: a comparison between supervised and unsupervised classification approaches, Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
[4] K.W. Cheung, J.T. Kwok, M.H. Law, K.C. Tsui, Mining customer product ratings for personalized marketing, Decision Support Systems 35 (2) (2003) 231–243.
[5] J. Coble, D. Cook, R. Rathi, L. Holder, Iterative structure discovery in graph-based data, International Journal of Artificial Intelligence Techniques 1–2 (14) (2005) 101–124.
[6] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis 1 (3) (1997) 131–156.
[7] C.C. Freifeld, K.D. Mandl, B.Y. Reis, J.S. Brownstein, HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports, Journal of the American Medical Informatics Association 15 (2008) 150–157.
[8] J. Gaurav, A. Ginwala, Y.A. Aslandogan, An approach to text classification using dimensionality reduction and combination of classifiers, Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration (2004) 564–569.
[9] A. Goswami, R.M. Jin, G. Agrawal, Fast and exact out-of-core k-means clustering, Fourth IEEE International Conference on Data Mining (2004) 83–90.
[10] V. Guralnik, G. Karypis, A scalable algorithm for clustering protein sequences, Proc. Workshop Data Mining in Bioinformatics (BIOKDD), 2001, pp. 73–80.
[11] K.F. Han, D. Baker, Recurring local sequence motifs in proteins, Journal of Molecular Biology 251 (1) (1995) 176–187.
[12] K.F. Han, D. Baker, Global properties of the mapping between local amino acid sequence and local structure in proteins, Proceedings of the National Academy of Sciences of the United States of America (1996) 5814–5818.
[13] V. Hatzivassiloglou, K.R. McKeown, Predicting the semantic orientation of adjectives, Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, New Brunswick, NJ, 1997, pp. 174–181.
[14] R.Q. Huang, J.H.L. Hansen, Dialect classification on printed text using perplexity measure and conditional random fields, IEEE International Conference on Acoustics, Speech and Signal Processing (2007) 993–996.
[15] T. Joachims, Text categorization with SVM: learning with many relevant features, Proceedings of ECM, 10th European Conference on Machine Learning, 1998.
[16] J.I. Khan, S. Shaikh, Relationship algebra for computing in social networks and social network based applications, 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 113–116.
[17] N. Li, X. Liang, X. Li, C. Wang, D. Wu, Network environment and financial risk using machine learning and sentiment analysis, Human and Ecological Risk Assessment 15 (2) (2009) 227–252.
[18] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.
[19] T. Saegusa, T. Maruyama, Real-time segmentation of color images based on the K-means CLUSTERING on FPGA, International Conference on Field-Programmable Technology, 2007, pp. 329–332.
[20] S. Schauland, A. Kummert, P. Su-Birm, I. Uri, Y. Zhang, Vision-based pedestrian detection—improvement and verification of feature extraction methods and SVM-based classification, IEEE Intelligent Transportation Systems Conference (2006) 97–102.
[21] Z.H. Sun, Y.X. Sun, Fuzzy support vector machine for regression estimation, IEEE International Conference on Systems, Man and Cybernetics, vol. 4, 2003, pp. 3336–3341.
[22] S. Tan, J. Zhang, An empirical study of sentiment analysis for chinese documents, Expert Systems with Applications 34 (4) (2008) 2622–2629.
[23] D. Thanh-Nghi, J.D. Fekete, Large scale classification with support vector machine algorithms, ICMLA 2007, Sixth International Conference on Machine Learning and Applications, 2007, pp. 7–12.

[24] S. Tong, E. Chang, Support vector machine active learning for image retrieval, Proceedings of ACM International Conference on Multimedia, 2001, pp. 107–118.
[25] T.B. Trafalis, H. Ince, Support vector machine for regression and applications to financial forecasting, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks 6 (2000) 348–353.
[26] P.D. Turney, Mining the web for synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the Twelfth European Conference on Machine Learning, Springer-Verlag, Berlin, 2001, pp. 491–502.
[27] P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., (2002).
[28] P.D. Turney, M.L. Littman, 315–346, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 21 (2003) 315–346.
[29] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[30] R. Vahidov, R. Elrod, Incorporating critique and argumentation in DSS, Decision Support Systems 26 (3) (1999) 249–258.
[31] D. Wu, Performance evaluation: an integrated method using data envelopment analysis and fuzzy preference relations, European Journal of Operational Research 194 (1) (2009) 227–235.
[32] D. Wu, Z. Yang, L. Liang, Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank, Expert Systems with Applications 31 (1) (2006) 108–115.
[33] D. Xu, S. Liao, Q. Li, Combining empirical experimentation and modeling techniques: a design research approach for personalized mobile advertising applications, Decision Support Systems 44 (3) (2008).
[34] J.M. Yang, X.Z. Huang, D. Zhuang, S.T. Zhang, The complex network analysis of competitive relationships between manufacturers in Foshan Ceramic Industry Cluster, 2006 International Conference on Management Science and Engineering, 2006, pp. 1020–1023.
[35] S. Yuan, A personalized and integrative comparison-shopping engine and its applications, Decision Support Systems 34 (2) (2003).
[36] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, C. Larson, Automatic online news monitoring and classification for syndromic surveillance, Decision Support Systems 47 (4) (2009) 508–517.
[37] X.H. Zhang, Z.B. Lu, C.Y. Kang, Underwater acoustic targets classification using support vector machine, Proceedings of the International Conference on Neural Networks and Signal Processing 2 (2003) 932–935.
[38] Y.G. Zhao, Q.M. He, An unbalanced dataset classification approach based on v-support vector machine, The Sixth World Congress on Intelligent Control and Automation, vol. 2, 2006, pp. 10496–10501.
[39] W. Zhong, G. Altun, R. Harrison, P.C. Tai, Y. Pan, Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property, IEEE Transactions on NanoBioscience 4 (3) (2005) 255–265.
[40] S. Zhou, T.W. Ling, J. Guan, J.T. Hu, A. Zhou, Fast text classification: a training-corpus pruning based approach, Proceedings of the Eighth International Conference on Database Systems for Advanced Applications, 2003, p. 127-13.
[41] http://domino.research.ibm.com/comm/research_projects.nsf/pages/cni.index.html.
[42] http://finance.google.com/finance?q=NYSE%3AIBM.
[43] http://finance.yahoo.com/q/co?s=INTC.
[44] http://www.youtube.com
[45] http://zp.isoche.com/.

Nan Li is a Ph.D. candidate at the Department of Computer Science, University of California, Santa Barbara. Her research mainly focuses on business data mining, text mining and Sentiment Analysis. Her work has been published/accepted at such journals as Human and Ecological Risk Assessment.

Desheng Dash Wu is the affiliated Professor in RiskLab at the University of Toronto and the Director of RiskChina Research Center at the University of Toronto. His research interests focus on enterprise risk management, business data mining, and performance evaluation in financial industry. He is the coauthor of *Enterprise Risk Management* book. He is co-editor in chief of International Journal of Services Sciences. His work has appeared in several journals as *International Journal of Production Research, European J. of Operational Research, IEEE Transactions on Knowledge and Data Engineering, Annals of Operations Research, J. of OR Society, International J. of Production Economics, Expert Systems with Applications, Computers and Operations Research, Human and Ecological Risk Assessmen*t, *International Journal of System Science, etc.* He has more than forty journal papers and coauthored 2 books. He has served as Editor/Guest Editor/Chair for several journals/conferences. The special issues he edited include those for *Annals of Operations Research*, *Human and Ecological Risk Assessmen*t, and *Production Planning and Control*. He is a Member of the Professional Risk Managers' International Association (PRMIA) Academic Advisory Committee.