KSS'2013 NINGBO

Proceedings

Knowledge Creation Towards Emergency Management

Shouyang Wang, Yoshiteru Nakamori and Weiliang Jin (eds.)

JAIST Press

ISBN: 978-4-903092-36-2

Risk Classification of Baidu Hot Word Based On Support Vector Machine

Yang Hu **Xijin Tang**

Institute of Systems Science, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, 100190 P.R. China
huyang11@mails.gucas.ac.cn, xjtang@amss.ac.cn

Abstract

Web text mining has provided a new approach to analyzing problems such as earthquake warning and terrorism monitoring, etc. In Web text mining, text classification is a widely leveraged method. In this paper, we combine text classification and search records from search engine together to facilitate detection of societal risk with referring to Tecent Weibo's volume at the same time. Here SVM is utilized to automatically identify risk category of Baidu hot word after extracting the news content linked by the first page of Baidu hot word search results. By implementing the process of SVM (Support Vector Machine) to text classification, we report the results of multi-classification experiment on Baidu hot word with some discussions. Finally, future research field is given.

Keywords: SVM, text classification, text extraction, Baidu hot word, Tencent Weibo

1 Instructions

Internet speeds up the information exchanging among people. People use the Internet to get the information they want easily. On the other hand, the record of people's Web behavior has stimulated the research on social problems. For example, Web text mining enables earthquake warning and terrorism monitoring by applying text classification to twitter with additional information such as hyperlink, geography tag [1-3]. Moreover, records from search engine are analyzed to forecast the outbreak of flu before official statistics, which is often 2 weeks later after aggregating the diagnosing cases from clinics [4]. In this paper, we combine text classification in Web text mining and records form Baidu, the biggest domestic Chinese search engine to provide an instantly reflection of societal risk in China.

This paper is organized as follows. Section 2 provides the detail of collecting Baidu Hot word and corresponding news, also Tencent Weibo's volume are employed in this part. Section 3 presents the process of SVM applied to text classification for Baidu hot word. Section 4 discusses the risk classification result of 15 experiments based on SVM. Conclusions and future work are given in Section 5.

2 Collecting Baidu Hot Words and Their Corresponding News

Baidu is now the biggest domestic Chinese search engine. The contents of high searching volume reflect focus of search engine's users. In another word, Baidu serves as an instantaneous corpus to maintain a view of people's empathic feedback for social hotspots, etc. In such way, we can utilize Baidu as a perspective to analyzing societal risk.

2.1 Baidu hot word and text extraction

The portal of Baidu hot words provides hot words every five minutes. The more users search, the higher rank hot words have. A Java Web crawler is implemented to grab hot words in each hour. At the end of each day, we accumulate the hot words in the past 24 hours and use htmlparser to parse out the hyperlink and corresponding title [5]. Then we store the hot words in xml files and save the first page of the corresponding Baidu search results. Figure 1 shows the process of collecting Baidu hot word.

Based on the 1st page of search results, news

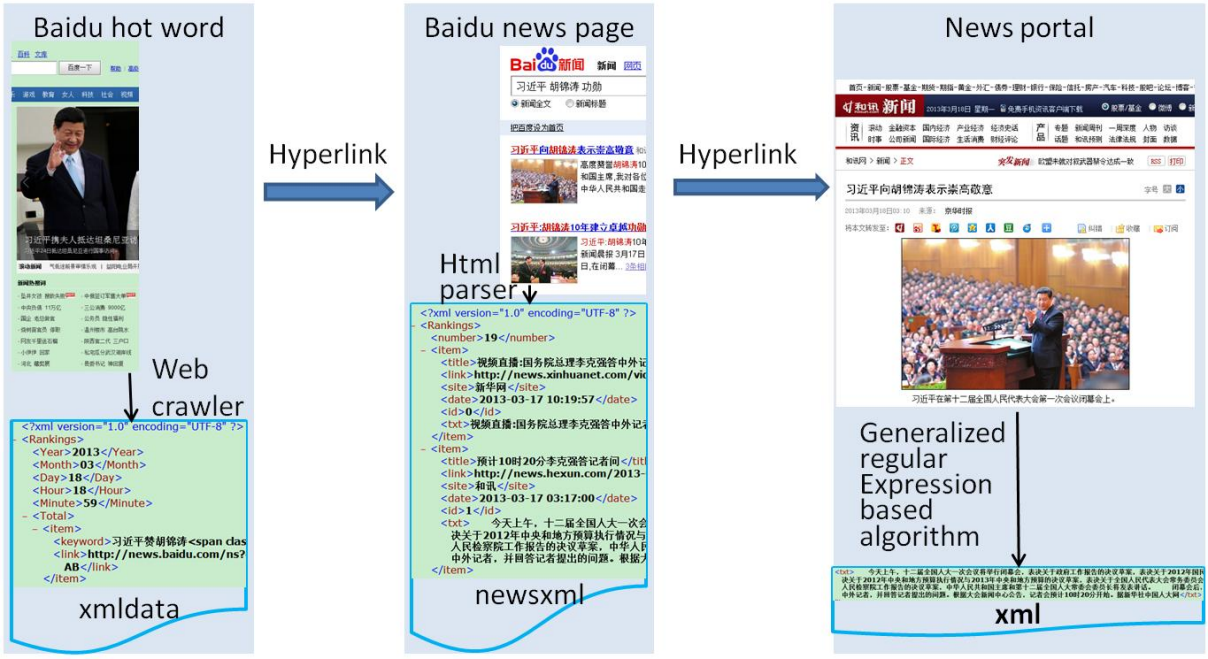


Figure 1. The portal of Baidu news redirects to word search page consisting of news page URLs

text from the news portals whose links are contained in the 1st page is extracted. As the news portals linked by Baidu news is varied, we firstly make a statistics of the news portals usually appeared at the first page of Baidu hot word search results. We find that top 10 news portals in June of 2012 are identical to those in September of 2012. Moreover, top 20% news portals release almost 80% of the news listed at the 1st page of the hot word search results [6]. By combing the top 20 percents news portals, we determine 138 news portals as our target seeds for text extraction.

By leveraging the computational method called generalized regular expression-based algorithm, we obtain the plain texts of diverse Web pages. The procedure of text extraction is as below.

Input:

S_1: html files of news pages concerning Baidu hot word

Loop:

For each html file x in S_1

Get all div blocks of x;

Abandon div blocks containing less than given threshold Chinese characters;

Select div block w with highest share of Chinese character;

Filter HTML tag of w using regular expression

End for

In the final, the news item is stored as an xml file consisting of six sub items including news *title*, *link* to news portal, *site* of news portal, publishing *date*, *id* which is the rank of this news item in word search page and the plain *text* of news page.

```
<item>
  <title>预计10时20分李克强答记者问</title>
  <link>http://news.hexun.com/2013-03-17/152155807.html</link>
  <site>和讯</site>
  <date>2013-03-17 03:17:00</date>
  <id>1</id>
  <txt>今天上午,十二届全国人大一次会议将举行闭幕会，表决关于政府工作报告的决议草案，表决关于2012年国民经济和社会发展计划执行情况与2013年国民经济和社会发展计划的决议草案，表决关于2012年中央和地方预算执行情况与2013年中央和地方预算的决议草案,表决关于全国人民代表大会常务委员会工作报告的决议草案,表决关于最高人民法院工作报告的决议草案,表决关于最高人民检察院工作报告的决议草案,中华人民共和国主席和第十二届全国人大常委会委员长将发表讲话。　　闭幕会后，国务院总理李克强将在人民大会堂金色大厅会见采访十二届全国人大一次会议的中外记者，并回答记者提出的问题。根据大会新闻中心公告，记者会预计10时20分开始。据新华社中国人大网</txt>
</item>
```

Table 1. Risk levels of 7 categories based on Baidu hot word (November, 2011 to October, 2012)

	national security	economy & finance	public morals	daily life	social stability	government management	resources & environment	total risk
2011. 11	0. 048	**0. 066**	0. 084	0. 124	0. 097	**0. 136**	0. 047	0. 601
2011. 12	0. 034	0. 058	0. 082	**0. 189**	**0. 117**	0. 099	0. 038	0. 616
2012. 01	0. 059	0. 040	0. 077	0. 187	0. 089	**0. 062**	0. 061	0. 575
2012. 02	0. 056	**0. 016**	0. 085	0. 167	0. 100	0. 074	0. 042	0. 540
2012. 03	**0. 028**	0. 025	0. 106	0. 116	0. 093	0. 105	0. 033	0. 505
2012. 04	0. 048	**0. 016**	**0. 049**	0. 145	0. 096	0. 120	0. 038	0. 511
2012. 05	0. 060	0. 018	0. 062	0. 145	0. 071	0. 088	0. 041	0. 485
2012. 06	0. 048	0. 037	0. 131	0. 154	0. 082	0. 112	0. 043	0. 607
2012. 07	0. 056	0. 065	0. 108	0. 103	**0. 056**	0. 095	**0. 067**	**0. 551**
2012. 08	0. 036	0. 051	0. 081	**0. 092**	0. 100	0. 081	0. 055	**0. 495**
2012. 09	**0. 109**	0. 034	0. 112	0. 162	0. 106	0. 083	0. 039	**0. 645**
2012. 10	0. 052	0. 028	**0. 162**	0. 173	0. 099	0. 078	**0. 011**	0. 604

*Bold number is the highest risk level and red number is the lowest risk level in each risk category.

2.2 Manual labeling for risk category of Baidu hot word

Based on the results of a study of risk cognition taken before 2008 Beijing Olympic Games [7, 8], we attribute each Baidu hot word into one certain risk category manually from November 1, 2011. The risk index compendium sorts out societal risk into 7 categories, national security, economy & finance, public morals, daily life, social stability, government management and resources & environment with 30 sub categories. Table 1 shows the risk levels of 7 categories between November, 2011 and October, 2012 based on manual labeling of risk category for Baidu hot word. Here the risk level denotes the proportion of total frequency of hot words labeled as one of the 7 risk categories to the total frequency of Baidu hot search words.

As shown in Table 1, risk level of each category is less than 0.2 and risk level of daily life is higher than those of the other risk categories. The risk level is highest in December, 2011 when risk level in daily life occupies one third of the total risk level. Rising price and income gap attract most of people's attention at the end of year. On the contrary, the risk level falls down sharply during August of 2012 when London Olympic Games takes place. The majority of hot words are relevant to sports which are risk-free during that period.

In order to monitor the risk level, it is necessary to map each hot word into a risk category and calculate the daily level of each category. Manual labeling of Baidu hot search words is a heavy burden, and then machine learning by SVM is explored to automatic labeling.

2.3 Risk level based on Tencent Weibo

There exists an obvious decline of risk level between July and September, 2012 in Table 1. To validate the risk level based on the Frequency that Baidu hot words occur, we make a comparison between risk levels based on frequency and Tencent Weibo's volume. We use Tencent API to search for Baidu hot word of each day and get the volume of Tencent Weibo about the Baidu hot word from 00:00:00 to 23:59:59 in that day. Similarly, we compute the risk level as the proportion of total Tencent Weibo's volume for hot words labeled as one of the 7 risk categories to total Weibo's volume for Baidu hot search words. Figure 2 show the risk level based Tencent Weibo's volume of Baidu hot word.

As shown in Figure 2, the red line represents the risk level based on Tencent Weibo's volume and the blue dashed line represents the risk level based on frequency that Baidu hot words occur. The correlation coefficient between the two risk levels is 0.73. In that sense, Baidu hot words indeed also reflect the focus of Weibo users. As search engine users and Weibo users are in fact

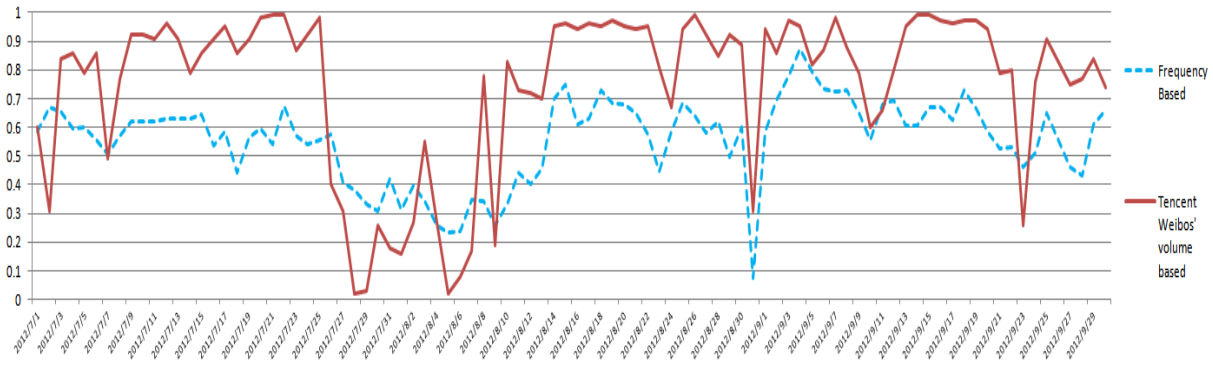


Figure 2. Risk levels based on Frequency and Tencent Weibo's volume of Baidu hot word (July 1, 2012 to September 30, 2012)

two main streams of Web users, the approach that realizing detection of societal risk through Baidu hot words could be more convinced.

3 Process of Support Vector Machine to Text Classification

SVM and others methods such as logistic regression, KNN, decision tree, artificial neural network, naive Bayes, etc are widely applied in text classification [9-11]. As for SVM, the unbalance of samples in each category perturbs the classification precision obviously. When applying SVM to risk classification of Baidu hot word, the varied categories of words at different time and different contexts are also negative for precision. For example, "Liu Xiang", a proper noun as the name of a well-known Chinese athlete, who withdrew from 2012 London Olympic dramatically because of his bruised right foot, is labeled as risk category of medical care from August 7, 2012. As more information about operation team exposed, risk category about "Liu Xiang" changes to morals and integrity after August 9, 2012. To justify the prediction results, category membership score is leveraged with using text content of Baidu hot word as samples while repeating each title twice.

In the process of SVM to text classification, we firstly segment plain text into Chinese terms to construct the initial dictionary by MMSeg [12]. At this step, stop words from HIT (Harbin Institute of Technology) are eliminated. The stop words list contains 767 functional words in Chinese.[1] Then terms within top given ratio on Chi score in each category are filtered into the final dictionary. As for methods such as information gain, mutual information and chi-square, chi-square out-performed other two methods in test [13]. So chi-square is tried here.

According to the dictionary obtained, plain text is transformed into text vectors of terms. Then weights are assigned to feature words in the text vectors. In this research, $tf \cdot idf$ is applied. In addition to $tf \cdot idf$, a pile of methods such as $tf \cdot rf$ (relevance frequency), $tf \cdot \chi^2$, $tf \cdot ig$ (information gain) can be adopted. Here rf measures the unbalance of documents containing the observed term between different categories. The formula of rf is $\log(2+a/b)$, where a is the number of positive documents in bipartition containing the term and b is the number of negative documents containing the term [14]. Afterwards text vectors of feature weights are inputted as samples into SVM training process.

In the prediction part, we leverage the category membership score to enhance the classification precision. The formula of category membership score is given as Equ.(1).

$$score = \frac{\sum S_i}{2*k} + \frac{k}{2*n} \quad (1)$$

where k is the number of voters supporting a certain category, n is the number of categories, S_i is the score of each supporting voter. As n-category classification problem in SVM can

[1] Obtained from http://ir.hit.edu.cn/bbs/viewthread.php?tid=20

be treated as multiple binary-classification problem, C_n^2 voters as classifiers in bipartition are computed. The philosophy of the category membership score is: for one test sample, the bigger the scores of voters as the first item in Equ.(1) are and the more the supporting voters as the second item are, the more convinced that the sample belongs to this category is. In the end, classification results of the biggest category membership score are chosen as the predicted category.

4 Experiment Results

According to the process addressed in Section 3, we carry out experiments using libSVM [15].

4.1 Data Source

Based on the manual labeling of each hot search word, we map the corresponding news text into the same risk category as that of the hot search word. The news text with double-repeated news title and corresponding risk category constitute the sample. Source data are from January 5, 2013 to February 28, 2013. Table 2 shows the number of Baidu hot search words and samples of each risk category in during that period. From Table 2, we see that three categories of risk, daily life, social stability and government management, contribute main risks. Obviously the samples among different risk categories are unbalanced. Figure 3 shows the sample size of each day from January 5, 2013 to February 28, 2013. The Average sample size during this period is 570.

Table 2. The number of Baidu hot words and samples of each risk category (January 5, 2013 to February 28, 2013)

Risk Category	Num. of Baidu hot words	Num. of samples
national security	124 (8%)	2443 (8%)
economy & finance	57 (4%)	1092 (4%)
public morals	131 (8%)	2492 (8%)
daily life	260 (16%)	4983 (16%)
social stability	137 (9%)	2635 (9%)
government management	356 (22%)	6769 (22%)
resources & environment	81 (5%)	1555 (5%)
risk-free	459 (29%)	8863 (29%)
Total	1605	30832

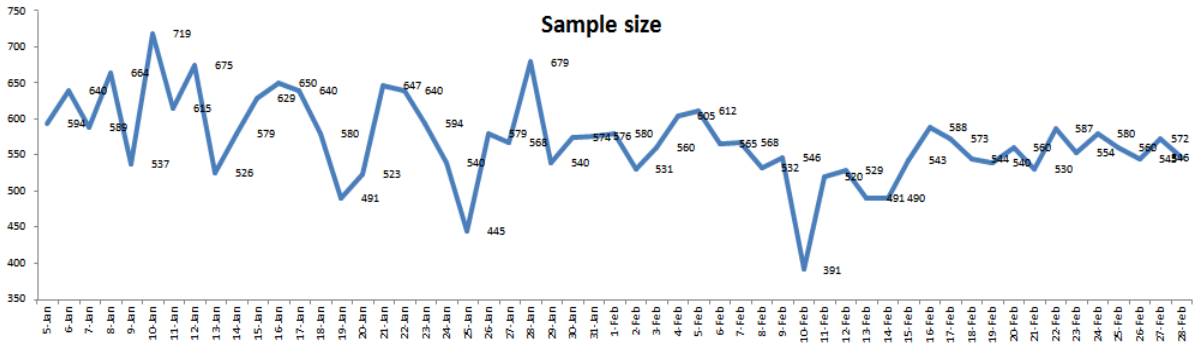


Figure 3. The number of Baidu hot words corresponding news of each day (January 5, 2013 to February 28, 2013).

4.2 Experiment Design

In our experiments, we classify Baidu hot search words of each day using previous N-day's data (N=1, 2, 3…) as training sets. The design of the experiments is based on the occurrence analysis of Baidu hot search words from January 5, 2013 to February 28, 2013 as

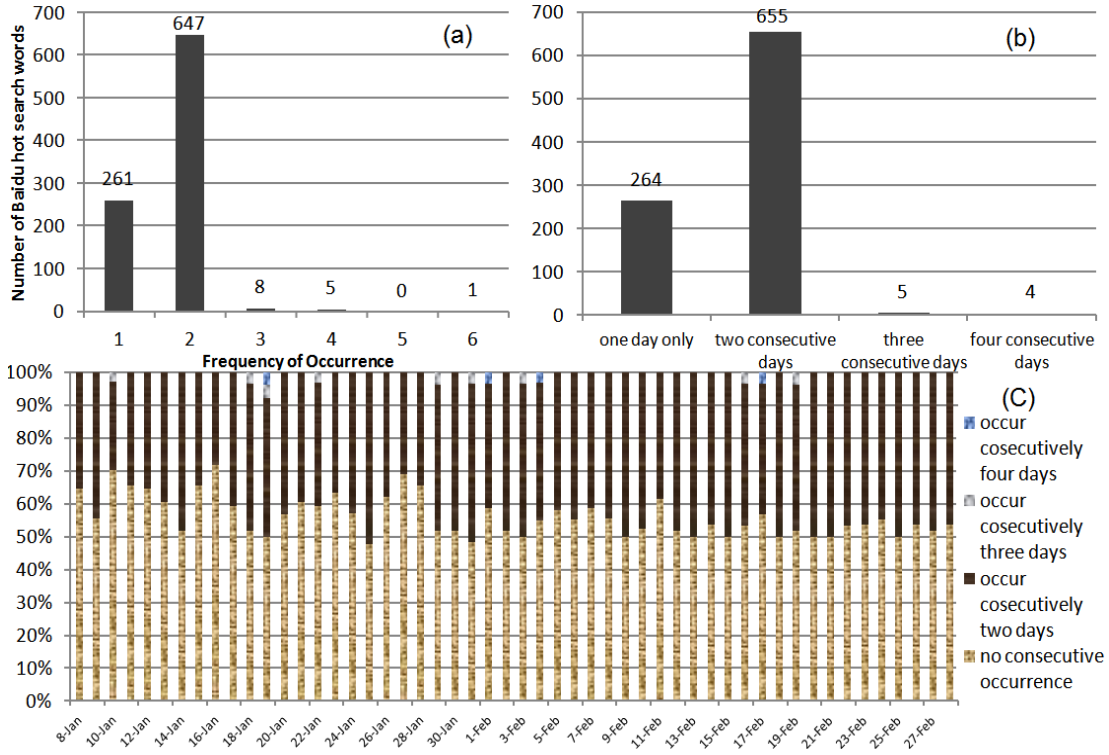
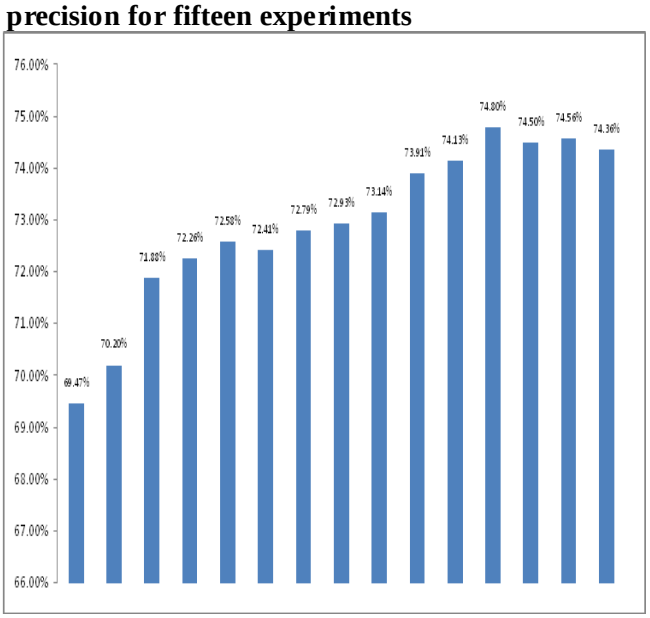


Figure 4. Statistic of occurrence for Baidu hot search words (January 5, 2013 to February 28, 2013)

shown in Figure 4. Figure 4(a) is the distribution of frequency that Baidu hot search words occur. In total, 261 hot words occur only once, 647 hot words occur twice, 8 occur three times, 5 occur four times and 1 occurs six times. Moreover, more than 70% of hot words consecutively occur two days or more as shown in Figure 4(b). In Figure 4(c), we find that nearly 40% of Baidu hot search words of each day already occur consecutively in previous days. Then it is quite

Table 3. Classification precision for fifteen experiments

	Average classification precision
Experiment 1	69.47%
Experiment 2	70.20%
Experiment 3	71.88%
Experiment 4	72.26%
Experiment 5	72.58%
Experiment 6	72.41%
Experiment 7	72.79%
Experiment 8	72.93%
Experiment 9	73.14%
Experiment 10	73.91%
Experiment 11	74.13%
Experiment 12	74.80%
Experiment 13	74.50%
Experiment 14	74.56%
Experiment 15	74.36%



*Parameter setting: RBF (radial basis function) is employed as kernel function, for feature selection top 40% terms in chi score are chosen, TF*IDF is adopted as feature weights.

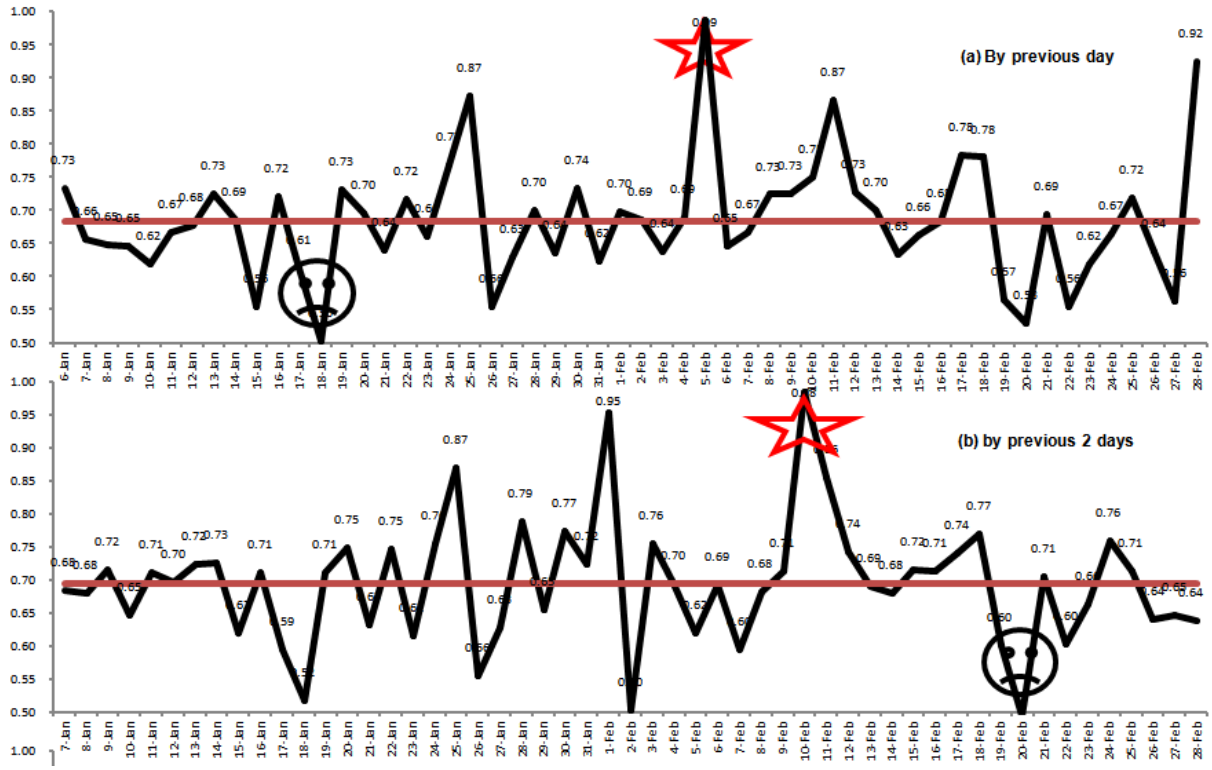
natural to use previous several days' data as training sets to classify Baidu hot search words. For comparison, we take 15 tests to find the fittest model. Experiment 1 uses previous day's data for training and today's data for testing, Experiment 2 uses previous 2-day's data for training, Experiment 3 uses previous 3-day's data for training and others go on like this. In fifteen experiments, RBF (Radial Basis Function) is employed as kernel function of SVM. Terms of top 40% on chi score are chosen to constitute the dictionary. TF-IDF is adopted as the feature weights.

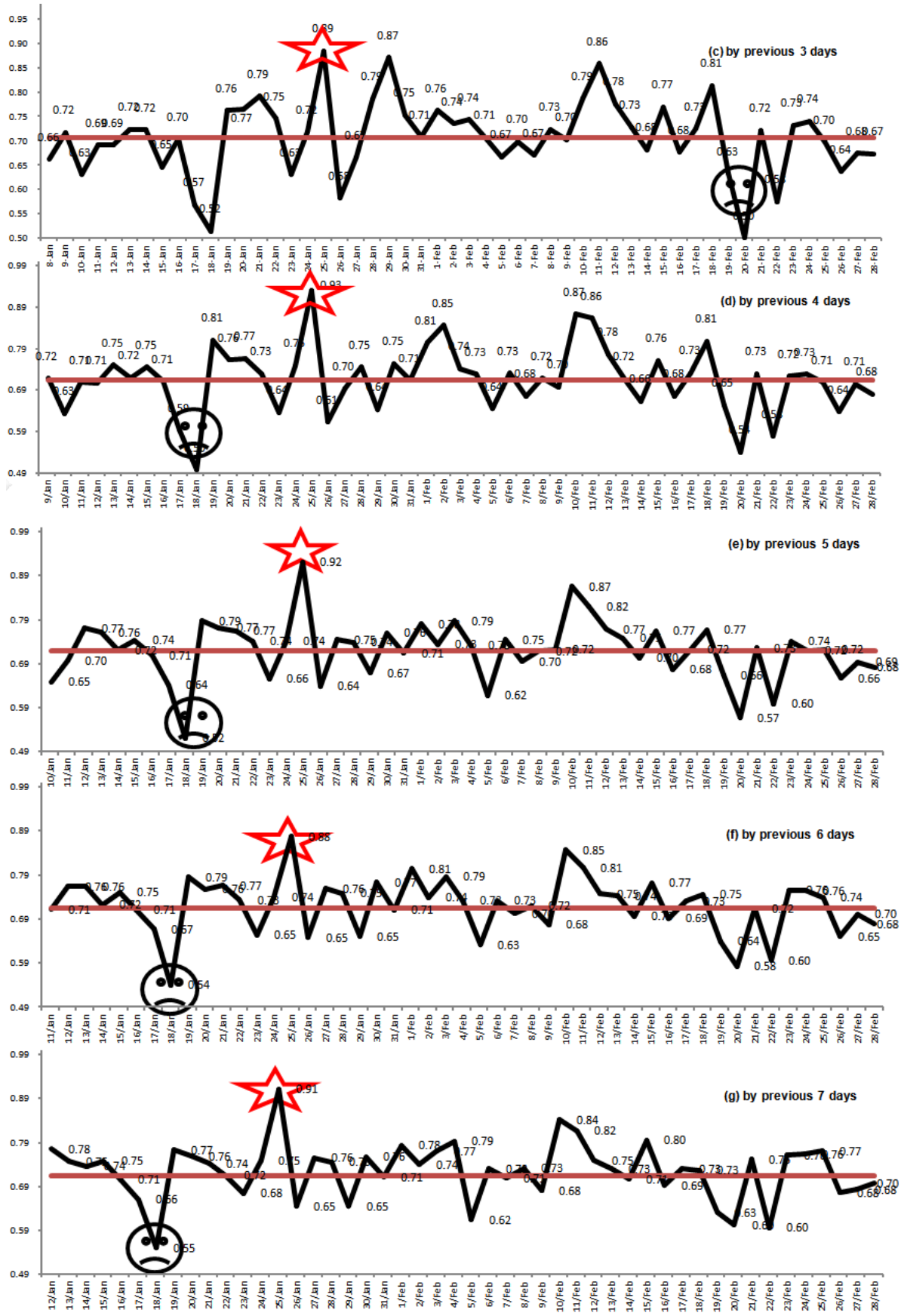
4.3 Experiment Results

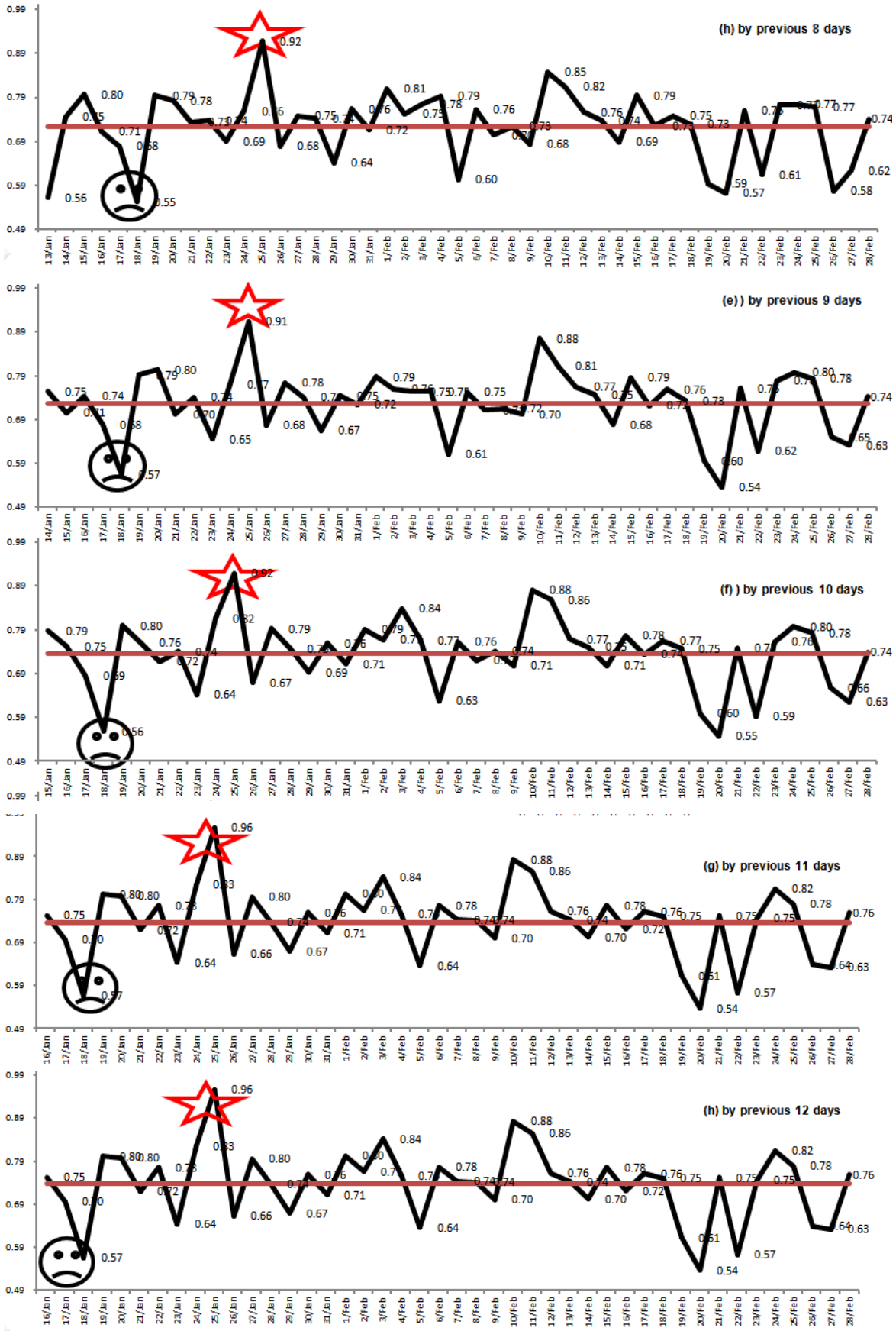
To measures the performance of SVM classification, we use the standard definition of precision as shown in Equ.(2) in this research.

$$precision = \frac{\left| S_{L(SVM)=L(Manual)} \right|}{\left| S_{sample} \right|} \qquad (2)$$

where $S_{L(SVM)=L(Manual)}$ is defined as the set of those news that SVM gives the same label as manual labeling. S_{sample} is defined as the set of news in the test samples. For each experiment, the precision values of each day are averaged to get a single-number measure of classification performance. The average classification precision of fifteen experiments is given in Table 3. Among fifteen experiments, Experiment 12 gets the highest precision and when using more than 10 days' data as training data, the precisions are hold around 14%. In such way, using 12 days' data would be the fittest model. We see that from Experiment 1 to Experiment 12, the classification keep an uprising trend. The reason is that as more data accumulated, the feature words from enriched corpus would be more representative so that for newly occurring hot words and text content, the SVM could give out a correct risk category with parsing out the singular feature word. However, when after 12 days, the classification precision settles out. Because of the emergence of novel hot words, the feature words of the past days cannot cover them, so it is hard to improve the classification precision any more. Figure 5 shows the detail of classification precision for each day of the 15 experiments.




(c) by previous 3 days
(d) by previous 4 days
(e) by previous 5 days
(f) by previous 6 days
(g) by previous 7 days


(h) by previous 8 days
(e)) by previous 9 days
(f)) by previous 10 days
(g) by previous 11 days
(h) by previous 12 days

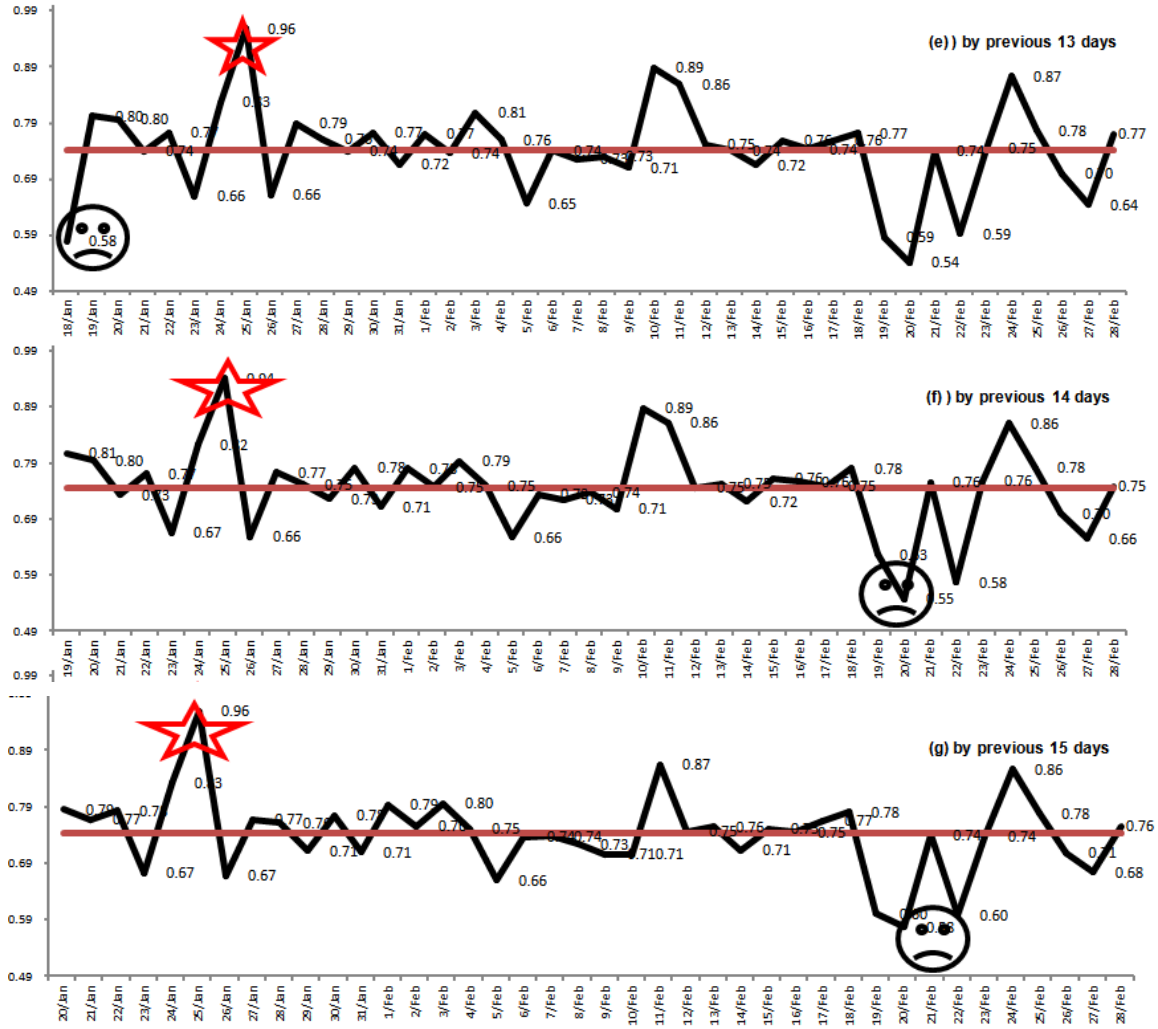


Figure 5. Classification precision for each day of 15 experiments, respectively.

As shown in Figure 5, the classification results on January 25 get the highest precision in 13 experiments. On the contrary, January 18 gets the lowest precision in 11 experiments and February 20 gets the lowest precision in 5 experiments.

In 15 experiments, classification precision for one day falls behind the other days when Baidu hot search words in that day that neither occur before nor are evident in risk classification. Hot words on January 18 in Experiment 1 are one typical example. The classification results on January18 get the worst precision in Experiment 1. Among Baidu hot search words on January18, we find one hot search word that does not appear before refers to one pop singer and actress talking about a political issue, which is labeled “, while manual label is national security, instead. Other hot words like GDP and a city mayor, who go to office by bicycle, which do not occur before, are indeed ambiguous in risk classification.

For 15 experiments, the best occurs when Baidu hot words occur before. The most typical example is the period of Chinese Spring Festival. Majority of Baidu hot search words concentrate on people’s holiday life including traffic problem, rising price and air pollution caused by fireworks, etc.

5 Conclusion

In this paper, we develop one Java program to collect Baidu hot words and their corresponding news, then we leverage the process of SVM to text classification and category membership scores to automatically identify risk category of Baidu hot words adopting the risk index com-

pendium form IOP, CAS. Based on the risk classification, we may have a vision of societal risk through Baidu hot words with referring to Tecent Weibo's volume of Baidu hot words in each day at the same time.

Fifteen experiments using different previous days' data are tested to find the fittest model to label the risk of today's hot words. The results show that classification for today by previous 12 days' data gets the highest precision. The uprising classification precision shows that as more data accumulated, the feature words from enriched corpus would be more representative, then the possibility that the SVM give out the correct risk category could be higher with parsing out the singular feature word for newly occurring hot words and text content. However, the classification precision settles out when more than 12 days because the feature words of the past days cannot cover the emergent novel hot words any further.

A lot of work needs to be done in the future. As a classification problem, empirical comparison with other methods is needed. And the effect of different removal percentage of feature words and different kernel types on classification precision is promising [16]. We also need to apply SVM to Baidu hot word using longer longitudinal data as more data are accumulated.

Acknowledgment

This research is supported by National Basic Research Program of China under Grant No. 2010CB731405 and National Natural Science Foundation of China under Grant No.71171187.

References

[1] Cheong M, Lee V C S. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter[J]. Information Systems Frontiers, 2011, 13(1): 45-59.

[2] Shen, Yang, et al. "Emotion mining research on micro-blog." Web Society, 2009. SWS'09. 1st IEEE Symposium on. IEEE, 2009.

[3] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[4] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2008, 457(7232): 1012-1014.

[5] Wu, D., Tang, X. J.: Preliminary analysis of Baidu hot words. Proceedings of the 11th workshop of systems science and management science of youth and 7th conference of logistic systems technology. Wuhan University of Science and Engineering Press, 478-483(2011) (in Chinese)

[6] Yang, H., Tang, X. J.: Using Support Vector Machine for Classification of Baidu Hot word. Proceedings of the 6th International Conference on Knowledge Science, Engineering and Management. Springer Berlin Heidelberg, LNAI 8041, 580-590(2013)

[7] Tang, X.J.: Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study. Proceedings of the First International Conference on Complex Sciences: Theory and Applications. Springer, LNICST5, 2338-2353(2009)

[8] Zheng, R., Shi, K., Li, S.: The Influence Factors and Mechanism of Societal Risk Perception. Proceedings of the First International Conference on Complex Sciences: Theory and Applications. Springer, LNICST5, 2266-2275(2009)

[9] Yang, Y.M., Liu, X.: A re-examination of text categorization methods. Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 42-49(1999)

[10] Tong, S., Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research, 2, 45–66(2001)

[11] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34, 1-47(2002)

[12] Tsai, C. H.: MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm. Available: http://www.geocities.com/hao510/mmseg/

[13] Yang, Y.M., Pedersen, J. O.: A comparative study on feature selection in text categorization. Proceeding of the 14th Interna-

tional Learning Conference on Machine Learning. Morgan Kaufmann Publishers, 412-420(1997)

[14] Lan, M., Tan, C. L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, 721-735(2009)

[15] Chang, C.C., Lin, C. J.: libsvm2.8.3. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[16] Zhang, W., Yoshida, T., Tang, X.J.: Multi-word extraction from Chinese text collection, Proceedings of APWeb'2008 Workshops (Y. Ishikawa et al. eds.). Springer-Verlag, LNCS 4977, 42-53(2008)