Using Support Vector Machine for Classification of Baidu Hot Word

Yang Hu and 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. Support vector machine (SVM) provides embarkation for solving multi-classification problem toward Web content. In this paper, we firstly introduce the workflow of Support Vector Machine. And we utilize SVM to automatically identifying risk category of Baidu hot word. Thirdly, we report the results with some dicsussions. Finally, future research topics are given.

Keywords: SVM, text classification, text extraction, Baidu hot word.

1 Introduction

Text classification is a popular research topic. Methods in text classification include logistic regression, KNN, decision tree, artificial neural network, naive Bayes, etc. As a well-known tool among those, support vector machine is widely utilized [1-3]. In this paper, SVM to text classification is conducted to automatically identify risk category of Baidu hot word.

This paper is organized as follows. Section 2 presents the process of support vector machine applied to Baidu hot word. Section 3 provides the detail of collecting Baidu Hot word and corresponding news. Section 4 discusses the risk classification result of 4 experiments based on SVM. Conclusions and future work are given in Section 5.

2 Process of Support Vector Machine to Text Classification

In the process of SVM to text classification as shown in Figure 1, construction of dictionary is the first step. At first, plain text is segmented into Chinese terms by MMSeg [4], while stop words are eliminated. In this research, stop words form HIT (Harbin Institute of Technology) are obtained. Their stop words contain 767 functional words in Chinese.[1] Finally, the remaining terms constitute the initial dictionary.

Next is feature selection as the way to generate the dictionary of salient terms. Among methods such as information gain, mutual information and chi-square, chi-square out-performed other two methods in test [5]. Here chi-square is tried. Terms within top given ratio on Chi score in each category are filtered into the dictionary.

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

M. Wang (Ed.): KSEM 2013, LNAI 8041, pp. 580–590, 2013.

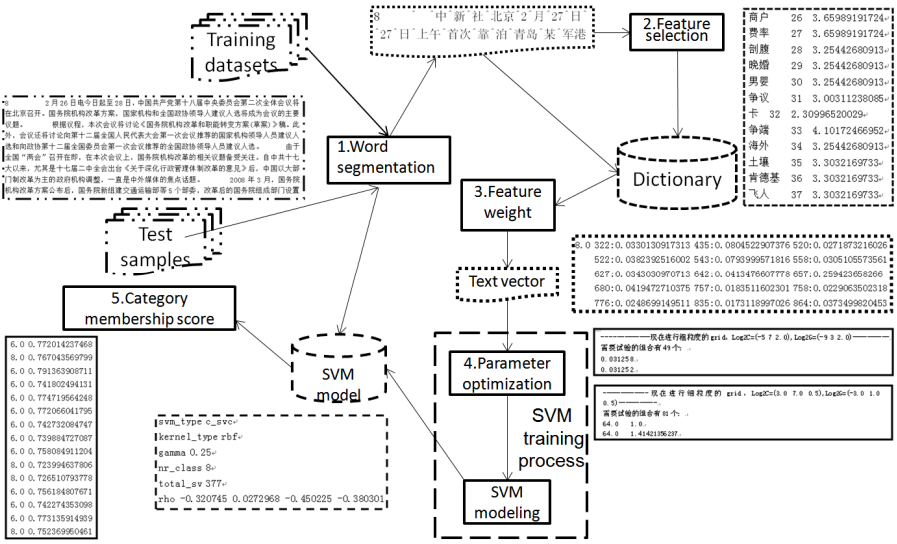

Fig. 1. A process of applying SVM to text classification

According to the dictionary obtained, plain text is transformed into text vectors of terms in the dictionary. Then weights are assigned to feature words in the text vectors. In this research, $tf \cdot idf$ is tried. 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. 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 [6].

Afterwards text vectors of feature weights are inputted as samples into SVM training process. In SVM modeling, the unbalance existing among separate categories of samples is irritating. Moreover, categories of words at different time and different context are varied. 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 bruised foot, is labeled as risk category of medical care from August 7, 2012. As more information about operating team exposed, risk category about "Liu Xiang" changes to morals and integrity from August 9, 2012. Therefore, classification precision by SVM is perturbed. So the category membership score is leveraged to enhance the classification precision. The category membership score is computed by Equ.(1).

$$score = \frac{\sum S_i}{2 * k} + \frac{k}{2 * n} \tag{1}$$

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 be treated as multiple binary-classification problem, C_n^2 voters

as classifiers in bipartition is computed. The philosophy of the category membership score is: for one test sample, the bigger the score of voter as the first item in Equ.(1) is and the more the supporting voters as the second item are, the more convinced the judgment that the sample belongs to this category is. With category membership score, classification result whose score is under the chosen best-fit threshold is ignored.

3 Collecting Baidu Hot Words and Their Corresponding News

Baidu is the biggest Chinese search engine in the world now. People search for information of their concerns and the content of high searching volume reflects focus of people. That's to say, Baidu serves as an instantaneous corpus to maintain a view of people's empathic feedback for community affairs, etc. Thus Baidu is utilized as a perspective to analyze societal risk with application of SVM addressed in Section 2.

3.1 Baidu Hot Word

The portal of Baidu news (http://news.baidu.com) provides routinely 10~20 hot search words which are updated every 5 minutes. Each hot news word corresponds to an individual URL of the word search page which consists of links to news from diverse news portals as shown in Figure 2.

Fig. 2. The portal of Baidu news redirects to word search page consisting of news page URLs

To collect Baidu hot search words, a Web crawler is customized to grab Baidu news page every hour. Then open source package htmlparser is leveraged to extract hot words from these html pages. Afterwards the hot search words are stored as an xml file. Meanwhile, a score is given to hot search word based on its rank in Baidu news page. If there are ten hot words, the 1st one is given 20 points, the 2nd one is 18 points and the last one is 2 points. If there are 20 hot words, the score of 1st hot word

is 20, the score of 2nd hot word is 19 and the last one's score is 1. Here score reflects the degree of people's attention. At 23:59 of each day, hot words gathered in the past 24 hours are accumulated and the scores in each hour are added up as day score of each hot word, then hot words are stored in descending order according to their day scores as an xml file. The 1st page among corresponding search pages is also stored. Based on those data, Baidu Hot Word Vision system adopting JSP was developed [7]. Users can get the daily, weekly and monthly rank list of Baidu hot word with frequency distribution, use iView and CorMap to analyze social events and search Baidu hot words containing key words of users' interest.

3.2 News Text Extraction

Based on the 1st page of search results, news text from the news portals whose links are contained in the 1st page is extracted. For big news portal such as Sina (www.sina.com.cn), there exist branch sites such as finance.sina.com.cn, sports.sina.com.cn, etc. Besides, those news portals often adopt new design of Web pages, an obstacle to customization mode of text extraction for each news portal. Hence statistics of news portals is utilized for text extracting. Then target seeds of text extraction are chosen by accumulating counts of hot search word-related news and figuring out affiliated news sites.

3.2.1 Statistics of News Sites

The target seeds for text extraction are determined by combining the set of web sites which account for 80% of hot words-related news. Table 1 lists the top 10 news portals.

Table 1. Top 10 news portals in June and September of 2012

News portal	Rank in June	Rank in September
人民网(www.people.com.cn)	1	5
搜狐(www.sohu.com)	2	1
21CN(www.21cn.com)	3	9
凤凰网(www.ifeng.com)	4	2
第一视频(www.v1.cn)	5	6
财讯(www.caixun.com)	6	10
新浪(www.sina.com.cn)	7	4
和讯网(www.hexun.com)	8	3
山东新闻网(www.sdnews.com.cn)	9	7
网易(www.163.com)	10	8

Figure 3 shows the distribution of frequency that major news portals contribute Baidu hot words in June and September of 2012. The result in Figure 3 shows that top 20% news portals release almost 80% of the news listed at the 1st page of the hot word search results. Furthermore, combination of target seeds in 2 months increases

the proportion. Web sites not in the scope of target seeds are also included by employing titles of their news as the origin of corpus for latter processing. Then, 138 news portals are chosen as target seeds for text extraction.

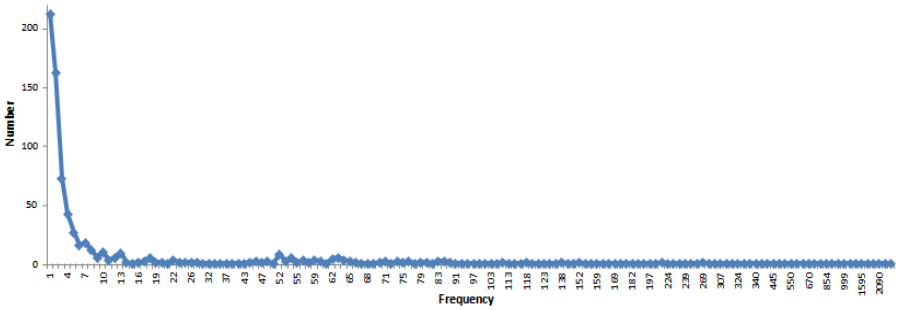

Fig. 3. The distribution of frequency that major news portals contribute Baidu hot words in June and September of 2012

3.2.2 News Text Extraction

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

```
Input:
     S1: html files of news pages concerning Baidu hot word
Loop:
    For each html file x in S1
       Get all div blocks of x
       Abandon div blocks containing less than given threshold Chinese
       characters
       Select div block w of highest share of Chinese character
       Filter HTML tag of w using regular expression
     End for
```

The threshold of the algorithm is set carefully according to the practical application. In this research, the value of threshold is 300 by default.

After the text extraction, the news text is stored as an xml file in directory YYYY-MM-DD/id.xml, where id is the rank of hot words in daily list. Each item consists of 6 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>传金正恩下令导弹部队待命攻击美军基地</title>
  <link>http://news.hexun.com/2013-03-29/152626268.html?from=rss</link>
  <site>和讯</site>
  <date>2013-03-29 07:00:00</date>
  <id>13</id>
  <txt>环球外汇3月29日讯在美国连续两次向出动B2隐形轰炸机飞往韩国，向朝
       鲜当局展现军力后，据《路透社》报导，朝鲜领导人金正恩周五(3月29
       日)在紧急会议中命令国内导弹部队随时待命，准备攻击韩国和太平洋
       (601099,股吧)的美军基地。</txt>
</item>
```

3.3 Manual Labeling of Risk for Baidu Hot Word

We map Baidu hot search word into a certain risk category. The risk category adopts the results of a study of risk cognition taken before 2008 Beijing Olympic Games [8, 9]. The risk index compendium sorts out societal risk into 7 categories, national security, economy & finance, public morals, daily life, social stability, government management, resources & environment with 30 sub categories. Figure 4 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 Figure 4, risk level of each category is less than 0.2 and risk level in daily life is higher than those of other 6 risk categories. The risk level is highest in December, 2011 and risk level in daily life takes up one third of the total risk level. We find that 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, 2012 when London Olympic Games takes place. The majority of hot words are relevant to sports which are risk-free during that period.

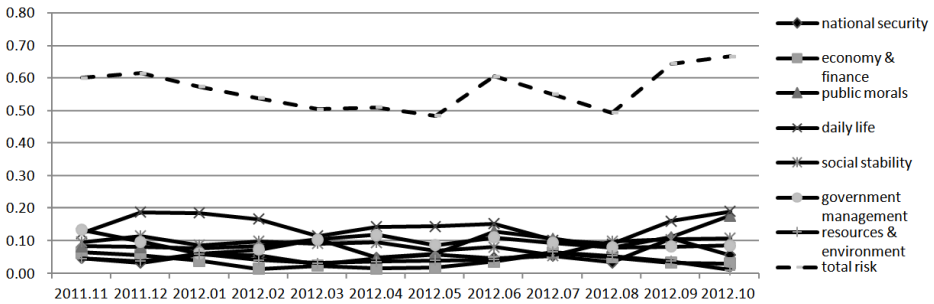

Fig. 4. Risk levels of 7 categories based on Baidu hot word (November, 2011 to October, 2012)

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

4 Experiment Results

According to the process addressed in Section 2, we carry out experiments using libSVM [10]. Source data are from January 5, 2013 to February 28, 2013. 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 titles and corresponding risk category constitute the sample. Table 2 shows the number of Baidu hot search words and samples of each risk category during that period. From Table 2, we see that the three categories of risk, daily life, social stability and government management, contribute main risks. So the samples among different risk categories are unbalanced.

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

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 shown in Figure 5. Figure 5(a) is the distribution of frequency that Baidu hot search words occur. In total, 261 hot words occur 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 5(b). In Figure 5(c), we find that nearly 40% of Baidu hot search words of each day already occur consecutively in previous days. Then it is quite natural to use previous several days' data as training sets to classify Baidu hot search words. In our experiment, we take 4 tests to find the fittest model.

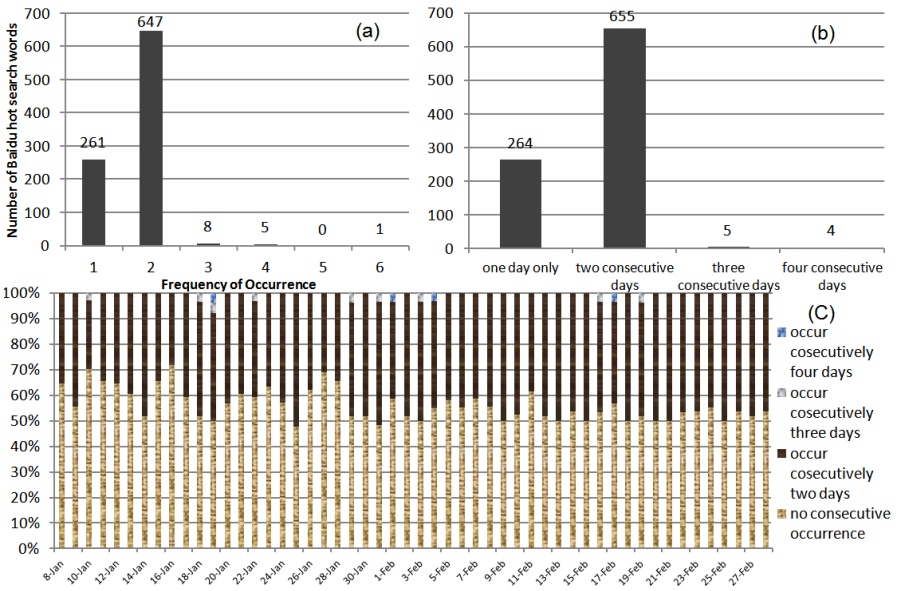

Fig. 5. Statistics of occurrence for Baidu hot search words (January 5, 2013 to February 28, 2013)

Four experiments are designed to test classification in our research. Experiment 1 uses previous day's data for training and today's data for testing, Experiment 2 uses previous 2-day's data as training sets, Experiment 3 uses previous 3-day's data as training sets and Experiment 4 uses previous 4-day's data for training. Figure 6 shows the sample size of each day from January 5, 2013 to February 28, 2013.

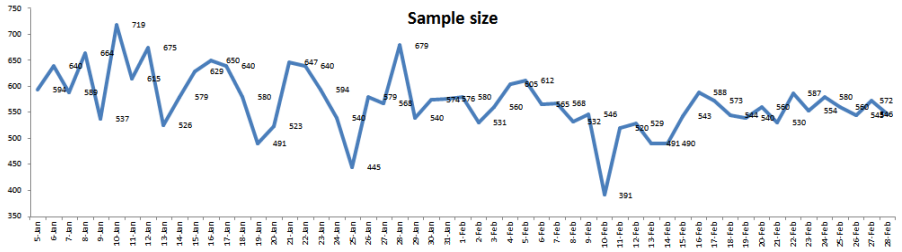

Fig. 6. The number of Baidu hot words corresponding news of each day (January 5, 2013 to February 28, 2013)

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

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

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 obtain a single-number measure of classification performance. The average classification precision of four experiments is given in Table 3. Among four experiments, Experiment 4 gets the highest precision. Figure 7 shows the detail of classification precision for each day of the four experiments.

Table 3. Classification precision for four experiments

	Average classification precision
Experiment 1	68.4%
Experiment 2	69.4%
Experiment 3	70.7%
Experiment 4	71.5%

**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.*

As shown in Figure 7, the classification results on February 4 in Experiment 1, February 9 in Experiment 2, January 24 in Experiment 3 and Experiment 4 get the highest precision in each experiment. On the contrary, January 17 in Experiment 1

and Experiment 4, February 19 in Experiment 1 and Experiment 2 get the lowest precision.

In four experiments, classification precision for one day falls behind the other days when the day contains Baidu hot search words that neither appear before nor are evident in risk classification. Hot words on January 17 in Experiment 1 are one typical example. The classification on January17 gets the worst precision in Experiment 1. Among Baidu hot search words on January17, we find one hot search word that does not appear before refers to one singer and actress talking about a political issue. The machine's label is risk-free, 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 happen before, are indeed ambiguous in risk classification.

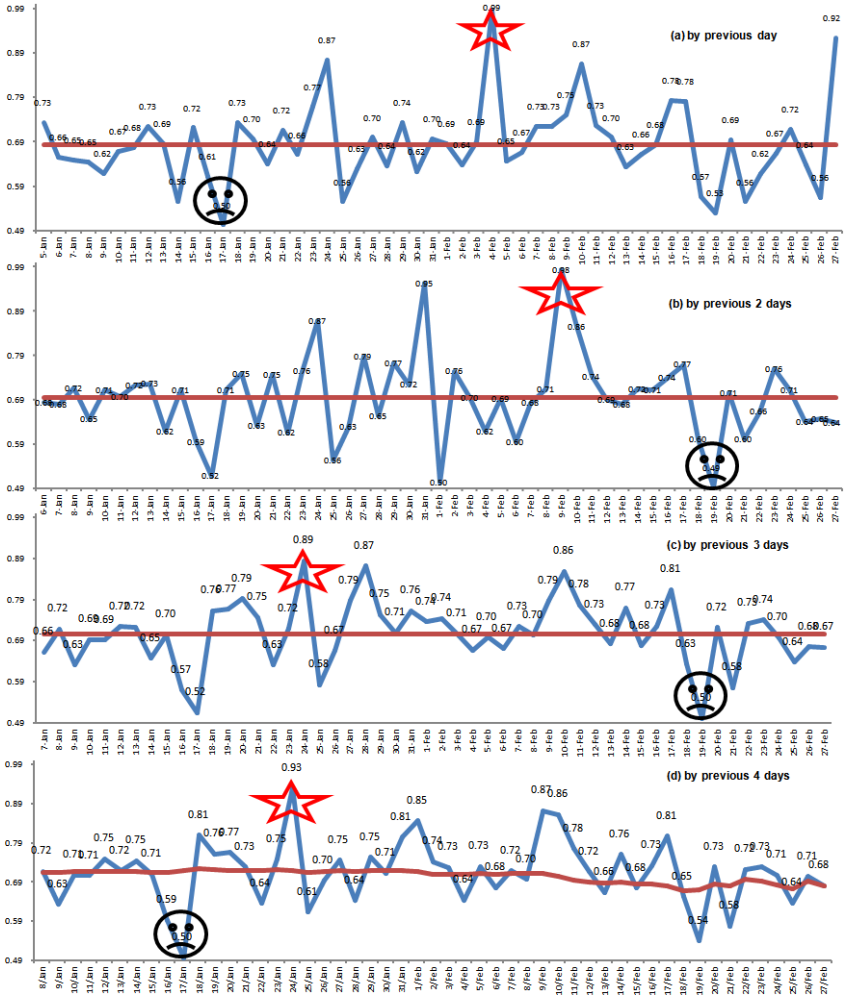

Fig. 7. Classification precision for each day of four experiments, respectively

For four experiments, the best occurs when Baidu hot words happen 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, price rising 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 to automatically identify risk category of Baidu hot words. Four experiments using different previous day's 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 4 days' data gets the highest precision. Based on the risk classification, we may have a vision of societal risk through Baidu hot words.

A lot of work needs to be done in the future. As a classification problem, empirical comparison with other methods is needed in our paper. And the effect of different removal percentage of feature words and different kernel types on classification precision is promising [11]. We also need to apply SVM to Baidu hot word using longer longitudinal data as more data are accumulated.

Acknowledgements. 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. Yang, Y.M., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development Information Retrieval, pp. 42–49 (1999)
2. Tong, S., Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research 2, 45–66 (2001)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34, 1–47 (2002)
4. Tsai, C.H.: MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm, `http://www.geocities.com/hao510/mmseg/`
5. Yang, Y.M., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceeding of the 14th International Learning Conference on Machine Learning, pp. 412–420. Morgan Kaufmann Publishers (1997)
6. 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)

7. Wu, D., Tang, X.J.: Preliminary analysis of Baidu hot words. In: Proceedings of the 11th Workshop of Systems Science and Management Science of Youth and 7th Conference of Logistic Systems Technology, pp. 478–483. Wuhan University of Science and Engineering Press (2011) (in Chinese)
8. Tang, X.: Qualitative meta-synthesis techniques for analysis of public opinions for in-depth study. In: Zhou, J. (ed.) Complex 2009. LNICST, vol. 5, pp. 2338–2353. Springer, Heidelberg (2009)
9. Zheng, R., Shi, K., Li, S.: The Influence Factors and Mechanism of Societal Risk Perception. In: Zhou, J. (ed.) Complex 2009. LNICST, vol. 5, pp. 2266–2275. Springer, Heidelberg (2009)
10. Chang, C.C., Lin, C.J.: libsvm2.8.3., http://www.csie.ntu.edu.tw/~cjlin/libsvm/
11. Zhang, W., Yoshida, T., Tang, X.: A Study on Multi-word Extraction from Chinese Documents. In: Ishikawa, Y., He, J., Xu, G., Shi, Y., Huang, G., Pang, C., Zhang, Q., Wang, G. (eds.) APWeb 2008 Workshops. LNCS, vol. 4977, pp. 42–53. Springer, Heidelberg (2008)