Internet public opinion hotspot detection and analysis based on Kmeans and SVM algorithm

Hong Liu
College of Computer and Information Engineering, Zhejiang Gongshang University
HangZhou, China
LLH@mail.hzic.edu.cn

Abstract—Rapid progress of network arouses much attention on Internet public opinion, it is important to grasp the internet public opinion in time and understand the trends of their opinion correctly. Text mining plays a fundamental role in categorization and monitoring of internet public opinion, but internet public opinion is much more difficult than pure-text process because of their semi-structured characteristic. To address this issue, we propose a model for internet public opinion hotspot detection and analysis. Due to the text format of internet public opinion, we introduce the traditional vector space model (VSM) to express them, and then use Kmeans algorithm to perform text clustering on a corpus collected from some news website, and use SVM classifier to perform text categorization for new text opinion analysis, the result of the experiment shows that the efficiency and effectiveness of such method.

Keywords- Internet public opinion; hotspot detection; text categorization; vector space model; SVM;Kmeans clustering

I. INTRODUCTION

With the advent of the Web and the enormous growth of digital content in Internet, network has become a primary carrier to reflect public opinion which is one of study of social psychology. News reports, BBS, forums, blogs, and etc are the main sources of public opinion information. So, such public opinions are regarded as Internet public opinions. Such Internet public opinion can have a great impact on real-world society security if we can dynamically detect hotspot internet public opinion, make useful information quickly exposed to those seekers, and real-timely monitor the tendency of public opinion. So this has motivated the research on detection of online public opinion.

The current research on the Internet public opinion is mainly to investigate on its impact on the real-world society or government, and analyze its activity pattern from psychology or sociology point of view [3]. Text mining plays a fundamental role in a number of information management and retrieval tasks. But the majority of these web data is in unstructured or loosely formatted texts often appears at a variety of tangible or intangible dynamic interacting networks [2]. A variety of heterogeneous online information website embodies the interacting networks nowadays. A web page is different from regular corpora of text documents. A text document can be treated as a bag of words whereas a web page has additional structural information.

As efficient business intelligence methods, data mining and machine learning provide alternative tools to dynamically process large amounts. In this paper, online public opinion hotspot detection is studied using text categorization. Our research is to provide a comprehensive and timely description of the interacting structural natural groupings of various public opinions, which will dynamically enable efficient detect of hotspot opinion, thus benefit Internet social network members in the decision making process.

The organization of this paper is as follows: section 2 reviews related work of IPO. Section 3 describes our methods of feature selection and dimensionality reduction. Section 4 gives a short description of Kmeans clustering and SVM classifiers. Section 5 describes our experimental method and reports the experimental results. Section 6 is a conclusion.

II. RELATED WORKS

Internet public opinion is a wide topic. Compared to text documents, web pages have extra features, such as HTML tags, URLs, hyperlinks and anchor texts, which have been shown to be useful in extracting sentiment. Recently much research [1,4] has been done on Web-page summarization to utilize these features in extracting sentiment. Dou[4] show that Web-page summarization techniques for preprocessing in Web-page classification is a viable and effective technique. In the paper, we apply text clustering algorithm for web-page through extracting the main relevant content from the web pages. Nowadays, there are also some research of web-pages mining based on the method of mathematic algorithm. P. D. Turney[11] introduces a simple algorithm for unsupervised learning of semantic orientation from extremely large corpora. The method involves issuing queries to a Web search engine and using pointwise mutual information to analyze the results. Similarly, Peter Jorgensen[12] explores the use of an interactive activation with competition (IAC) artificial neural network (ANN) to find relationships in email texts. PJianping Zeng[5] introduced Hidden Markov Model (HMM) to describe the activity of IPO. G.M. Sacco [15] describes a way to identify to topic relevant portions of a hierarchical space, while L. Terveen[16] gives a methodology to derive the sites that
pertain to a given topic. Other research has used the method of text classification or text clustering. Text classification is currently a hot subject of research in information search and data mining field [6]. It has a rapid development in recent years with wide-ranging applications in information filtering, natural language processing and organization and management of information.

III. FEATURE SELECTION AND DIMENSIONALITY REDUCTION

Feature selection process aims to obtain a substantial reduction of the features, whose methods based on document frequency, mutual information, or information gain could be used to reduce the number of words [8,9]. During the classification process of network text opinion, not all the words are used in the classification process. Instead, the important words that distinguish a certain text opinion are extracted and used later on detecting network opinion hotspot. These words are called features.

There are two types of features: the full word features and the stemmed features. The full word features are the words extracted from the articles as they are. The stemmed features are the stems of the extracted words. Considering the integrity and accuracy of hotspot detection and analysis, we use the full word features in this paper. Besides, we used the common term-weighting tf.idf measure that is used commonly to weight features. The tf-idf weight (term frequency–inverse document frequency) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus, which combines the term frequency (tf) which measure the term density in a document multiplied by the inverse document frequency (idf) which is a measure of informativeness of a term (its rarity across the whole corpus).

The tf.idf weighting is identified by the formula:

\[
W(t,d) = TF(t,d) \times IDF(t) = TF(t,d) \times \log \left( \frac{D}{DF(t)} \right)
\]

(1)

Where the term frequency \(TF(t,d)\) is the frequency (number of times) of word \(t\) in the document \(d\). The document frequency \(DF(t)\) is the number of documents that contain word \(t\).

In order to use this weighting measure, each text opinion must be converted into a vector of words that exist in the article. Where we apply vector space model (VSM) for expressing the text opinion. After that we compute the tf.idf weight of each word and then select the word features that have the highest tf.idf weights for each article to construct the VSM of document, whose form is shown as follows:

\[
d = \{ (t_1, w_{1d}), \ldots, (t_k, w_{kd}) \}
\]

(2)

where \(t_k\) is feature of document and \(w_{kd}\) is the weight of \(t_k\). VSM does not take the account of the position information or grammar implication of features, so in this sense a document vector is a Bag of Word (BOW). VSM adopts similarity between two documents to express the correlation between them, that is, higher similarity means more correlation. Since the documents are represented by a vector, operations on documents can be implemented on the corresponding vectors.

The problem of VSM is the independent assumption of words, and this is always unsatisfactory. Documents in some news websites are most related to a particular theme thus context-sensitive. According to the consideration of the properties of IPO, we propose a new model, analyses and detection of Internet Public Opinion hotspot, as shown in Fig.1.

IV. HOTSPOT DETECTION AND ANALYSIS APPROACH

In recent years, there have been extensive study and rapid progress in automatic text categorization, including the traditional machine learning approaches such as k-nearest neighbor classifier, Naïve Bayes classifier, and decision tree. Even these methods are not so significant compared to those new approaches such as Rocchio classification, Widrow-hoff classification and SVM (support vector machine), however, they have the advantages of simple algorithms and relatively high efficiency, the modified and improved methods based on these traditional approaches are continuing to cause concern. In this paper, we use Kmeans clustering and SVM classifier to categorize web texts. Therefore, only these two methods are presented in this section.

A. Kmeans clustering

Kmeans is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume \(k\) clusters) fixed a priori. The main idea is to define \(k\) centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this
point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function:

$$J = \sum_{j=1}^{n} \sum_{i=1}^{m} ||x_{ij}^{(i)} - c||^2$$  \hspace{1cm} (3)

Where $\gamma$ is a chosen distance measure between a data point and the cluster centre, $c$, is an indicator of the distance of the $n$ data points from their respective cluster centres.

Suppose that we set time segments and get internet public opinion in the time segments, through feature selection and dimensionality reduction, we express them as follow:

$$D = \begin{bmatrix} T_1, T_2, \ldots, T_m \\
D_1 \begin{bmatrix} w_{11}^T, w_{12}^T, \ldots, w_{1m}^T \\
D_2 \begin{bmatrix} w_{21}^T, w_{22}^T, \ldots, w_{2m}^T \\
\vdots & & \ddots & & \vdots \\
\vdots & & \vdots & & \ddots \\
D_n \begin{bmatrix} w_{n1}^T, w_{n2}^T, \ldots, w_{nm}^T \\
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}$$  \hspace{1cm} (4)

Where $Di$ means text public opinion, $Ti$ means feature, and $w_{ij}$ means feature weight. The data set used as the input of the Kmeans clustering, which will be clustered into k groups, with a center topic for each cluster. The hotspot is those closest to the theoretical centers of the clusters.

**B. SVM classification**

Support Vector Machine (Vapnik, 1998; Cortes and Vapnik, 1995) is a state-of-the-art classification algorithm that is known to be successful in a wide variety of applications. High generalization ability of the method makes it particularly suited for high dimensional data such as text. Indeed, it has been shown that SVM outperformed most of the other classification algorithms in text categorization tasks (Joachims, 1998; Dumais et al., 1998). In the paper, we apply the SVM algorithm in detecting network opinion hotspot.

The numbers of network opinion hotspot is uncertain, so it is a multi-classification problem. Here we translate input space into high-dimension space through nonlinear function. In the high-dimension space, we construct linear distinguish function to realize nonlinear distinguish of original space, and get the classification decision-making function $f(x)$ shown as follows:

$$f(x) = \text{sgn} \left( \frac{1}{n} \sum_{j=1}^{n} \gamma_{ij} \Phi(x_j) \cdot \Phi(x) + b \right) = \text{sgn} \left( \frac{1}{n} \sum_{j=1}^{n} \gamma_{ij} K(x_j, x) + b \right)$$  \hspace{1cm} (5)

Where, $K(x, x)$ selects radial basis function as inner product kernel function, whose form shown as follows:

$$K(x, x) = \exp \left( \gamma ||x_i - x_j||^2 \right), \gamma > 0$$  \hspace{1cm} (6)

Apart from Kmeans clustering, SVM is utilized in this section to realize hotspot forecasting. In order to forecast the hotspot distribution with the current time segments, we fed into the SVM model with the historical data we obtain from the last time segments. As for the output of the SVM, which serves as the supervised learning tool in our work, the clustering result by the Kmeans approach with the current time segment is used. A well-trained SVM is utilized to carry out prediction for the next time segments by inputting the data obtained from the current one. Suppose the current time segment is $s_i$, if a forecast for $s_{i+1}$ is expected, we first train a SVM by inputting representation vectors of $s_i$ and setting the output as the clustering result for $s_i$ by Kmeans. Then the trained SVM generates classification outputs for data of $s_i$. Finally, SVM result is compared to the Kmeans clustering result for data of $s_{i+1}$.

**V. EXPERIMENT RESULT AND ANALYSIS**

To evaluate the text categorization result, we use macro-averaging precision, macro-averaging recall and macro-averaging F1-measure. Whose forms are shown as follows:

- product kernel function, whose form shown as follows:

$$\text{Macro}_R = \frac{1}{n} \sum_{i=1}^{n} R_i$$  \hspace{1cm} (7)

$$\text{Macro}_P = \frac{1}{n} \sum_{i=1}^{n} P_i$$  \hspace{1cm} (8)

$$\text{Macro}_F1 = \frac{2 \times \text{Macro}_R \times \text{Macro}_P} {\text{Macro}_R + \text{Macro}_P}$$  \hspace{1cm} (9)

Where Precision (P) is the proportion of actual positive class members returned by the system among all predicted positive class members returned by the system. $P = TP/(TP+FP)$ . Recall(R) is the proportion of predicted positive members among all actual positive class members in the data. $R = TP/(TP+FN)$. F1 is the harmonic average of precision and recall as shown: $F1 = \frac{2 \times P \times R}{P + R}$.

**A. Computation using Kmeans Clustering**

In this section, we conduct Kmeans clustering among the data from some web sites as: news.sohu.com, bbs.qianlong.com, news.cctv.com, www.china.com, www.zaobao.com, unn.people.com.cn and category of data
include finance and economics, humanistic, life, entertainment, etc. One deficiency of Kmeans is that a predetermined value of k is required. To overcome this drawback, Kmeans cluster analysis is conducted for a set of k values, we calculate the same vector space model over different k values ranged from 5 to 10. The evaluation value of the VSM over different k values is shown as Table 1. Such these data show the average values of the same vector metrics over different k values. From them, we can see our method is generally sufficient to achieve a satisfying result for accuracy. And for current case, we can get good result when k is set to a value of 9. So, during the subsequent experiments, we set the value of k as 9 and calculate the different vector matrix over the different time span. The result is shown as Table 2.

<table>
<thead>
<tr>
<th>k-means</th>
<th>Macro_P</th>
<th>Macro_R</th>
<th>Macro_F1</th>
<th>Vector dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=5</td>
<td>0.642276</td>
<td>0.940476</td>
<td>0.763285</td>
<td>961</td>
</tr>
<tr>
<td>K=6</td>
<td>0.763889</td>
<td>0.8</td>
<td>0.781528</td>
<td>1231</td>
</tr>
<tr>
<td>K=7</td>
<td>0.693694</td>
<td>0.916667</td>
<td>0.789744</td>
<td>1702</td>
</tr>
<tr>
<td>K=8</td>
<td>0.71028</td>
<td>0.904762</td>
<td>0.795812</td>
<td>1918</td>
</tr>
<tr>
<td>K=9</td>
<td>0.761905</td>
<td>0.836735</td>
<td>0.797568</td>
<td>2383</td>
</tr>
<tr>
<td>K=10</td>
<td>0.842105</td>
<td>0.72619</td>
<td>0.779864</td>
<td>2459</td>
</tr>
</tbody>
</table>

### Table II. Experiment Result over Different Vector Matrix and the Same k Value

<table>
<thead>
<tr>
<th>Feature dimension</th>
<th>Computer</th>
<th>Education</th>
<th>Sport</th>
<th>Technology</th>
<th>automobile</th>
<th>entertainment</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>87.5</td>
<td>71.32</td>
<td>80.69</td>
<td>80.69</td>
<td>72.39</td>
<td>84.52</td>
</tr>
<tr>
<td>150</td>
<td>46.42</td>
<td>73.71</td>
<td>48.92</td>
<td>54.55</td>
<td>75.18</td>
<td>72.39</td>
</tr>
<tr>
<td>200</td>
<td>73.71</td>
<td>54.55</td>
<td>48.92</td>
<td>75.18</td>
<td>72.39</td>
<td>84.52</td>
</tr>
<tr>
<td>250</td>
<td>60.95</td>
<td>75.18</td>
<td>84.52</td>
<td>72.39</td>
<td>84.52</td>
<td>72.39</td>
</tr>
<tr>
<td>300</td>
<td>48.92</td>
<td>54.55</td>
<td>72.39</td>
<td>84.52</td>
<td>72.39</td>
<td>72.39</td>
</tr>
<tr>
<td>350</td>
<td>94.64</td>
<td>75.18</td>
<td>84.52</td>
<td>72.39</td>
<td>84.52</td>
<td>72.39</td>
</tr>
<tr>
<td>400</td>
<td>84.52</td>
<td>72.39</td>
<td>84.52</td>
<td>72.39</td>
<td>84.52</td>
<td>72.39</td>
</tr>
</tbody>
</table>

### Table IV. Feature Dimension of Five Categorization

<table>
<thead>
<tr>
<th>Feature Dimension</th>
<th>Computer</th>
<th>Education</th>
<th>Sport</th>
<th>Technology</th>
<th>Automobile</th>
<th>Entertainment</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>74</td>
<td>30</td>
<td>42</td>
<td>20</td>
<td>313</td>
<td></td>
</tr>
</tbody>
</table>

### Table V. F1 Measures for Two Methods

<table>
<thead>
<tr>
<th>Feature Dimension</th>
<th>K-means</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>82.55</td>
<td>87.5</td>
</tr>
<tr>
<td>Education</td>
<td>46.42</td>
<td>71.32</td>
</tr>
<tr>
<td>Sport</td>
<td>73.71</td>
<td>84.52</td>
</tr>
<tr>
<td>Technology</td>
<td>60.95</td>
<td>80.69</td>
</tr>
<tr>
<td>automobile</td>
<td>48.92</td>
<td>75.18</td>
</tr>
<tr>
<td>entertainment</td>
<td>54.55</td>
<td>72.39</td>
</tr>
</tbody>
</table>

### VI. Conclusion

In this work, we propose a model about internet public opinion hotspot detection and analysis, the main technique of categorization is text categorization. According to the text properties of internet public opinion, we introduce VSM to express text opinion. Text corpora are chosen from some new websites. We perform Kmeans clustering and SVM classifier on the documents, the experimental result shows that the efficiency and effectiveness of such method. What’s more, our future works are shown as follows:

1. Commerce application research is the other one of our future works. Using our hotspot detection approaches can help seeker get quickly information what they want to get. For market department, it can help them understand what their specific customers’ concerns regarding goods and services information. It is beneficial for us to provide best personalized services for people.

2. Deeply approach research of internet public opinion detection. Refinement for each step of the approach proposed above is needed. Dynamic monitoring technology is in demand which can monitor the web sites to detect change in time. Data cleaning is time-consuming and labor-intensive. Web content analysis can not stop at word frequency analysis because sometimes the result is polysemantic. How to confirm the optimal k value of clustering algorithm, how to improve process speed of massive data is also main research content of our future works.

### ACKNOWLEDGMENT

This paper is supported by the Zhejiang provincial department of science and technology as grand science and technology special social development project (No.2008C13082). And this paper is also supported by the Zhejiang provincial natural science foundation as general science and technology research project (No. Y1080565),
and supported by the National Natural Science Foundation of China under Grant No. 60903053, supported by the Special Funds for Key Program of the China No. 2009ZX01039-002-001-04, 2009ZX03001-016, 2009ZX03004-005.

REFERENCES


