I.J. Computer Network and Information Security, 2011, 3, 48-54
Published Online April 2011 in MECS (http://www.mecs-press.org/)

A public opinion classification algorithm based on micro-blog text sentiment intensity：Design and implementation

Xin Mingjun, Wu Hanxiang, Li Weimin, Niu Zhihua
School of Computer Engineering and Science, Shanghai University, Shanghai 20072, China
{ xinmj, newwhx, wmli}@shu.edu.cn; zhniu@staff.shu.edu.cn

***Abstract*—on the features of short content and nearly real-time broadcasting velocity of micro-blog information, our lab constructed a public opinion corpus named MPO Corpus. Then, based on the analysis of the status of the network public opinion, it proposes an approach to calculate the sentiment intensity from three levels on words, sentences and documents respectively in this paper. Furthermore, on the basis of the MPO Corpus and HowNet Knowledge-base and sentiment analysis set, the feature words' semantic information is brought into the traditional vector space model to represent micro-blog documents. At the same time, the documents are classified by the subjects and sentiment intensity. Therefore, the experiment result indicates that the proposed method improves the efficiency and accuracy of the micro-blog content classification，the public opinion characteristics analysis and supervision in this paper. Thus, it provides a better technical support for content auditing and public opinion monitoring for micro-blog platform.**

***Index Terms*—micro-blog; sentiment intensity; public opinion classification algorithm**

I. INTRODUCTION

Micro-blog is a kind of blogging variant that has risen in recent years. It becomes an important network platform for public opinion expression，which gains more attention and recognition for its short format and real-time characteristics. In Wikipedia, micro-blog is described as "a broadcast medium in the form of blogging allows users to exchange small elements of content such as short sentences, individual images, or video links [1]." The differences between micro-blog and traditional blog are that users of micro-blog could make use of web browsers, mobiles and other network terminals to read and publish text, images, audio，video links and other types of information anywhere at any time. Because the content of micro-blog is shorter (generally no more than 140 chars or Chinese words), the transmission speed among users is faster, and the expression means are freer.

The "Social Blue Book", published in December 2009 by the Chinese Academy of Sciences, considered micro-blog as "the most lethal carriers of public opinion"; The "2010 third-quarter Assessment Analysis Report of China's Response capacity to Social public opinion", published in October 2010 by Shanghai Jiaotong University, claimed that micro-blog was becoming an important channel for enterprises and individuals to respond to public opinion. People can publish their own opinions about all aspects of social life on the one hand, on the other hand, people can make use of the "follow" links to form micro-blog groups as the communication platforms between government information and social public opinion information. These make micro-blog as a facilitative role in the social development . Each coin has two sides, and so does micro-blog. The complete liberalization of speaking one's opinion and fast speed of information broadcasting has made the traditional oversight mechanism broken up. Because of the loss of gatekeepers and uncertified information , the micro-blog platform is filled with a lot of unreal information and harmful content which push people who do not know the truth and have litter discernment into the untruth information world.

Therefore, new challenges are brought to the government monitoring public opinion trends and discovering public opinion crisis. Unfortunately, research on micro-blog for public opinion in China has just started, and lacks of sophisticated systems and applications. To study and analyze micro-blog text semantic tendency, the lab has constructed a public opinion corpus on the content of micro-blog information, and proposed an approach to marking corpus from the point of semantic. Based on the analysis of the status of network public opinion and the features of micro-blog, we propose a micro-blog document sentiment intensity computing model on words, sentences and documents respectively and a classification algorithm based on subjects and sentiment intensity, which would improve the efficiency of the public opinion characteristics analysis and supervision.

II. RELATED WORK

Text classification as the management and organization technology of unstructured text information has been widely studied and used [2] [5] [7] [9] [11] [12], which is a complex process including documents representation, classification algorithm designation, and performance evaluation. The main task is documents formalization and classification algorithm designation.

The purpose of document formal representation is to transform text documents to a mathematical model which the computer can understand and process. The

representational model should reflect the text documents' inherent semantic information and other features as much as possible. In the last years, some traditional text models were designed, taking the Bayes model [13] [14] [15] and the vector space model (VSM) [16] [17] [18] [19] [20] [21] as the examples. Vector space model which is the most popular text representation model has been widely used in the text classification and retrieval researches. The center idea of VSM is to use the vector<w1, w2, w3, …, wn> to represent a text, where 'wi' is the i-th characteristic word weight in the text[4] [6] [8]. Generally single Chinese characters, words, or phrases can be used to be one item of the vector, while many experiments show that words being the unit are more effective than others [8]. The feature words in the VSM are selected on the basis of linear independence among words. But with the characteristics of the natural language, the semantic links exist in the context, which cannot meet the assumptions most of time. So the accuracy of classification is influenced.

HowNet built by Professor Dong Zhendong is a common sense knowledge base for Chinese words, which reveals and reflects the relationships among concepts abstracted from Chinese characters or attributes of concepts. The crux of the HowNet philosophy is all matters are in constant motion and ate ever changing in a given time and space in the corresponding change in their attributes [10]. HowNet extracts sememes from about 6000 characters with a bottom-up grouping approach, respectively, classified as event class, entity class, attribute or quantity class, attribute or quantity values class [4]. Event Role is a semantic relation between concepts. Event role is the possible participants and roles playing in the event. HowNet also describes the entity class as event role in some events that it plays in. Relations among concepts mainly include hypernym-hyponym, synonym, antonym, converse, part-whole, attribute-host, material-product, agent-event, patient-event, instrument-event, location-event, time-event, value-attribute, entity-value, event-role, and concepts co-relation.

After the analysis to the information spreading mode under internet-based micro-blog platform and the sender's mood tendency to some specific topics in the content of micro-blog message, we have studied a model under the micro-blog service platform and a algorithm to quickly classify and audit the related micro-blog content in the spreading process. Based on the researches , we propose a micro-blog text documents emotional intensity calculating model on the basis of the HowNet Knowledge Base. Then a improved public opinion classification algorithm of micro-blog documents is brought up from the subject and text emotional intensity on the basis of MPO Corpus. And the classification algorithm is a update version of the tendency classification algorithm in reference [4].The rest of this paper is organized as follows. Section **III** instructs the MOP Corpus. Section **IV** proposes the sentiment intensity computation model. Section **V** studies the classification algorithm. Section **VI** analyzes experimental result and evaluates the algorithm performance. Finally, conclusion remarks are given in Section **VII**.

III. INSTRUCTION OF THE MOP CORPUS

The MPO Corpus is a semantic corpus consisting of a set of micro-blog documents, which is the foundation of text classification, retrieval, comprehensive and comparison. The corpus services for the public opinion research oriented micro-blog platform. The original sources are collected from Sina, Tencent and Sohu that are three main micro-blog platforms in China. The construction of corpus is a long process which include many original sources collection, and text annotation.

TABLE I .
THE BASIC INFORMATION OF INITIAL CORPUS SOURCES

Sources	Num of characters	Num of words	Num of sentences	Num of micro-blogs
Tencent	4265	3072	107	50
Sina	3613	2601	129	50
souhu	4128	2968	117	50
Total	12006	8641	353	150

Currently, for experimental tests, the corpus only includes 150 documents, 12006 characters, 8641words and 353 sentences, and about 58 positive tendency blogs, 92 negative tendency notes and about 17 subjects. The following table (Table **I**) is the basic information of the initial corpus sources distribution.

A micro-blog document in the MPO Corpus is considered as a doc, consisting of head and body. The main annotation information contains the meta-information of micro-blogs and the segment information of texts. The head is marked about the meta-information which reveals the micro-blog self information, including the serial number 'index' in the classification of information, author information 'author', source information 'source', subject information 'topic', list of keywords 'keywords', emotional intensity 'intensity' and so on. And the body is about micro-blog text labeling information composed by a series of sentences to reveal the grammatical information and the semantic information in the sentences and context. Sentences labeling is the core of the corpus annotation including the index of sentences 's_no' in the doc, sentence length 's_len', the original text 'origin', word text after the segment 'segmentation', rhetoric 'rhetoric', opinion or fact information 'opinionFact', agent 'agent', patient 'patient', and keywords 'keywords', and semantic annotation based on the HowNet. The semantic annotation includes syntax part and semantic part. The syntax part of words includes the serial number of words 'w_no', the start position 'start', the length of the word 'w_len' and parts of speech in the sentences. The semantic part includes the concept annotation 'class' based on HowNet, and sentiment tendency 'polarity' for emotional words based on HowNet emotion word set. More information about the MPO Corpus can be found in

the reference [3]. The labeled documents are stored in the corpus with the XML format, and one finished example is shown in Fig.1.

Figure1 Example of a micro-blog corpus entity.

IV. Sentiment Intensity Computation Model

The sentiment intensity computing model oriented micro-blog is the foundation of the classification of documents based on the text emotional intensity, and is also the basis of public opinion research about micro-blog information platform. The model proposed in the paper includes three levels with the emotional intensity calculation from words, sentences and documents. The model frame is show in Fig.2.

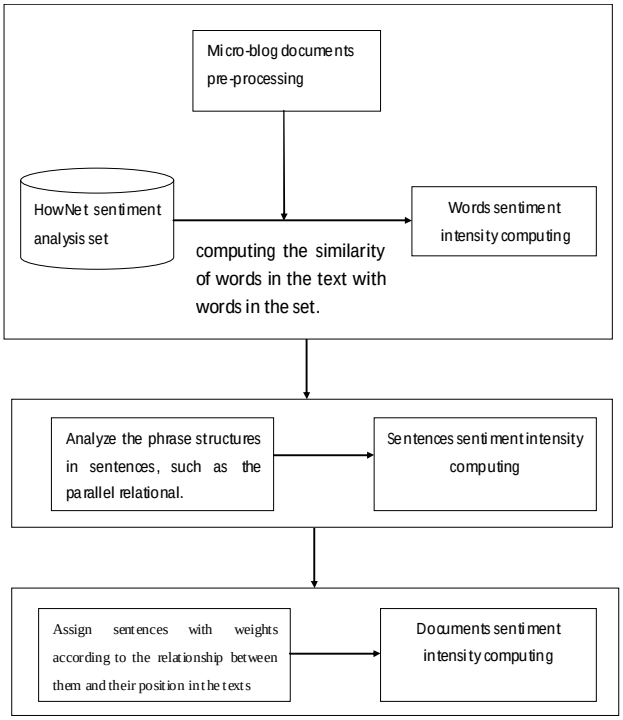

Figure2 Sentiment Intensity Computing Model Frame.

In Fig.2, for the starts, the original documents will be pre-processed as words segmentation, and then the program will compute the similarity based the algorithm in reference [11] between the words in the documents and HowNet sentiment analysis set to set the words' emotional intensity. These are the words' level process. For the sentences level, the program will analyze the relationships between words consisting of phrases like modified relationships, parallel relationship.etc. The sentimental intensity of sentences is computed based on the relationships of words and words' intensity. Finally, with different positions of sentences, the program analyzes the position of sentences in the context and sets each sentence different weight to calculate the document's intensity. The detailed instructions will be shown as follows respectively.

A. Word Emotional Strength

Word emotional intensity computing is based on HowNet sentiment analysis set, which consists of Chinese and English emotion analysis words sets, including positive and negative evaluation words, positive and negative emotion words, degree-level words and words and claim words. Because the emotion difference between the evaluation words and emotional words are not very obvious in the research, the sentiment analysis words are merged into the positive words set and negative words set in the paper, such as:

Positive words: love and dote, love and esteem, caress, love.

Negative words: sad, pity, grieved, deep sorrow, dump.

Pre-process the words in the set and give all the positive emotional words the weight 1, all the negative emotional words the weight -1 for the emotion sets. The degree-level words in the set do not contain any emotional information, but modify emotions degree intensity. So the paper gives these words a positive real number weight between 1 and 10. Then compute the similarity based the algorithm in reference [11] between words in the documents and the processed HowNet sentiment analysis set. The proposed computing algorithm of words emotional intensity list as follows:

if speech of word is Degree Adverb then
Calculate the similarity between word and word' in the HowNet degree-level words set;
Note the biggest similarity 'sim' and weight 'weight'of word';
else if speech of word is one of nouns, verbs, adjectives then
Calculate the similarity between word and word' in the HowNet emotional words set;
Note the biggest similarity 'sim' and weight 'weight'of word';
else intensity(word) = 0;
*intensity(word) = weight * sim;*

B. Sentence Emotional Strength

Words are the basic unit of the sentences, but sometimes a single word does not accurately reflect the semantics of a sentence such as:

Sentence 1: Fuel consumption of Excelle is really high

Sentence 2: Etta's cost performance is very high

Sentence 1 and sentence 2 are emotional sentences, but the emotional word 'high' shows different polarities when modified different objects: 'high' indicates derogatory in the sentence 1while compliment in the sentence 2. Therefore, we study the modified relationship between the adjacent words before calculating the sentences emotional intensity. Some researchers have found the phrases structures with certain emotional meaning are usually nouns, verbs, adjectives, adverbs phrases. The common Chinese phrase types such as prejudiced phrase are shown in Table Ⅱ.

TABLE Ⅱ.
THE COMMON PHRASE CONSTRUCTS

	Grammar Structures	Examples
One center word	adjective+noun	A clever girl(Chinese meanings：聪明的女孩)
	noun+verb, noun+adjective	Wang likes (Chinese meanings：小王喜欢)
	verb+noun, verb+adjective	Like clean(Chinese meanings：爱干净)
	noun+'of'+noun	The affinity of idols(Chinese meanings：偶像的亲和力)
	degree-adv.+ adj./adv., adj./adv.+ degree-adverb	Very good(Chinese meanings：很好)
	Negative word +adj./ verb/adv.	Do not like (Chinese meanings：不喜欢)
Multiple center words	Adjective+adjective, noun+noun, verb+verb	Bright and smart(Chinese meanings：聪明伶俐)

To compute the emotional intensity, it obeys the following rules in the paper:

a) The emotional strength of the parallel structure phrases such as: "noun + noun", "adjective + adjective" is equal to the sum of the each word' strength.

b) The emotional strength of the modified structure phrases such as: "adjective + noun", "adjective + adverb" is equal to the product of multiplying like intensity(adverb) * intensity(adjective)

To facilitate the calculation of the emotional intensity of the sentence, two presumption are made:

a) Each sentence is a single sentence, and complex sentences composed by the conjunction artificially are split into two sentences;
b) The similarity based on HowNet is increased by 10 times.

Under the analysis of semantic links between words in the phrases and the context relationships in the sentences, the sentence emotional intensity algorithm is designed and shown as follows:

```
intensity = 0;
While(word1is not the last word){
    If there is modified relationship between word1and word2
        Combine word1 and word2 into word;
        Intensity(word) += intensity(word1) * intensity(word2);
        word1 = word;
    else intensity += intensity(word1) + intensity(word2);
}
```

C. Document Emotional Strength

In a document, the relationships between sentences, such as the assumed, transition and progressive, affect the document emotion intensity. The topic sentence in the document occupies a central position having significant impact on document emotion intensity. Therefore, it gives each sentence a different weight to reveal diffident positions in a micro-blog text. The calculation is according to the following formula (α , β i is the correlation coefficient):

$$\text{intensity} = \alpha * \text{intensity}(\text{topic sentence}) + \beta 1 * \text{intensity}(\text{sentence1}) + \ldots + \beta n * \text{intensity}(\text{sentenceN}) \quad (1)$$

In the formula discussed above, the position is more important in the document, the coefficient is larger. Usually, the coefficients about the topic sentences are set a float number among0.5-1 and other sentences' coefficients are set among 0-0.5.

Ⅴ. PUBLIC OPINION CLASSIFICATION ALGORITHM

The micro-blog documents Classification is the groundwork of the analysis of micro-blog public opinion. In the paper, it studies the features of micro-blog documents and the traditional text classification algorithms and document models. A category method is proposed to micro-blog text documents based on the MPO Corpus and text emotional tendency. The algorithm classifies documents from two levels with subjects and sentiment intensity of the text.

There are two forms of the subject words exited in the micro-blog:

a) The shown topics micro-blog. These blogs' topics are shown up between the labels '#' in the blog documents. The labels '#' and '#' are a convention labels in the micro-blog community meaning a topic between them, which are designed to collect the same topic blogs conveniently.

b) The hidden topics micro-blog. These blogs' topics are hidden in the texts and need some analysis of the whole text from the semantic view point. Some blogs may have different content but belong to the same subject.

For the hidden topic blogs, the main task is to find out the topic words under the research on document model. The category method for shown topic micro-blogs is simple, and the classification detailed below mainly directs to the hidden class. Fig.3 is shown the flow chart of the classification algorithm proposed.

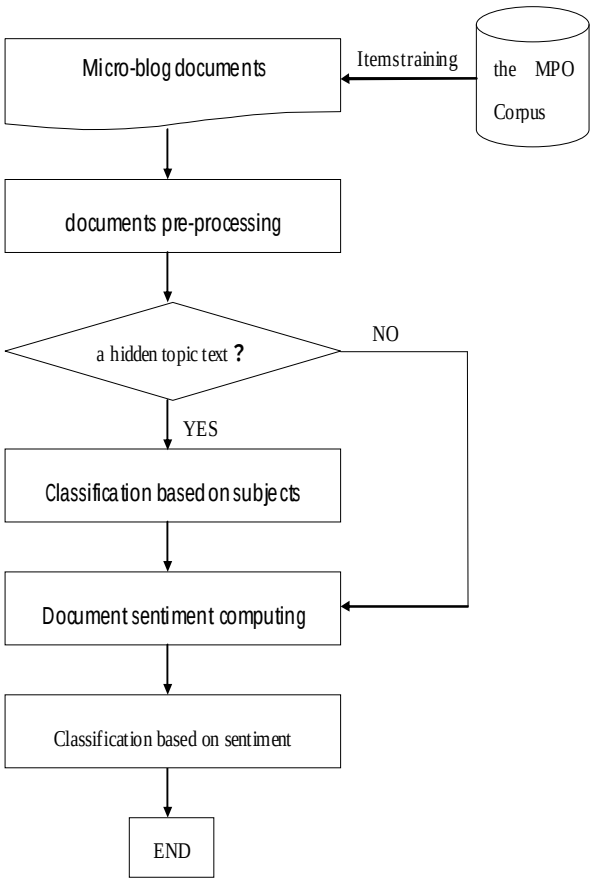

Figure3 the flow chart of the classification algorithm.

In Fig.3, the unclassified documents in the MPO Corpus will be preprocessed firstly. Then the classification program will analyze whether the one is a hidden topic document or not. If it is a hidden topic blog, represent it using the improved vector space model proposed in the paper. After the document is classified into a subject class, the program calls for the document sentimental intensity calculating algorithm instructed above to compute the sentimental intensity of it. At last the document will be classified into a public opinion class, such as the Red class, the Orange class, the Yellow class, or the Green class according to the document intensity. The paper below will detail the category process.

The topic-based classification is a process to classify micro-blogs into different subjects according to different semantic scopes of the documents. Its main task includes text representation model and subject words selection. And the process details as follows.

a) pre-process the documents, model the documents in a mathematics way and select the feature words for the subject words.

b) select and train a proper classifier, such as KNN algorithm.

c) make the ready classifier classify documents based on subjects.

In the paper, we choose the vector space model (VSM) as the micro-blog document mathematic model. The VSM selects the keywords by using the IF-IDF model to calculate the words frequency weight and sorting to choose the biggest ones in the vector of documents and documents set. The keywords consist of the vector. Then the cosine value between the vectors is used to determine whether two documents belong to the same subject or not. The traditional model ignores the semantic links between key words, and only takes words frequency information into consideration, which fluencies the classification accuracy. And in this paper, it considers the semantic factors and hypernym-hyponym relationships based on HowNet while selecting the keywords.

Emotional intensity of micro-blog documents in the classification model reflects the emotional tendencies and the public opinion warning level and other information with the same subject. In the paper, it calculates the documents sentiment intensity and classifies the documents into Red, Orange, Yellow and Green four classes under the subject category. In the topic class set, the Red class includes the highest level of negative emotional intensity documents , the Orange class includes the higher level of negative emotional intensity documents,, the Yellow class includes the little high level of emotional intensity documents, and the Green class the lowest level documents. So in the public opinion work, if documents percentage in the negative class is large under one subject means the maybe public opinion harm is higher, and the related apartments would take much attention to those subjects.

To classify a micro-blog text document, a pre-processing process for origin documents is needed before the proposed algorithm first. Because of many colloquial terms, and inaccurate punctuations or sub-sentences, the sources would be pre-processed obeyed the rules instituted in the paper shown as follows:

a) The first line is the author's information.

b) Each context sentence holds one line.

c) Following the context line is the segmentation line that is the result of segmented sentence and the speech tagging of each word.

d) Replace a space with a comma or a full stop after analyzing the context of short sentences separated by apace.

e) The pre-processed file named by format like "date + time".

After pre-processed, documents are in a uniform format. Then they are taken into the classification flow to be classified. The proposed algorithm descripted as follows:

// document collection and pre-process

Step 1: micro-blog documents collection

Step 2: documents preprocessing

Step 3: semantic annotation for the preprocessed documents based on HowNet

// the classification of the text documents on micro-blog topics based on HowNet Knowledge Base.

Step 4: get words' statics information 'w' in the text and texts set。Through lots of text analysis, parts of speech of the feature words mainly focus on noun, verb, and adjective. So the particle, pronoun, preposition words etc. can be removed.

Step 5: merge the words with hypernym-hyponym relationships each other to the hypernym words whose frequencies is the sum of the hypernym-hyponym words.

Step 6: sort the words by the value 'w', and select the top 20 words as the feature words by experiences ;

if the document is the first one in the subject class

then compare the feature words vectors with documents and subject vectors based on the semantic tree, and replace the hyponym words with hypernym words. And note the cosine value as similarity 'sim' of the document and the subject vector.

if sim > threshold 'TH'

then class the document to the subject class 'T' ;

update the subject vector ;

else build a new subject ;

// the classification of the text documents on text sentiment under each subjects.

Step 7: calculate documents emotional intensity in each subject class

Step 8: if intensity >= 0 classified into the Green class;

else if -25 <= intensity < 0 classified into the Yellow class;

else if -50 <= intensity < -25 classified into the Orange class;

else if intensity < -50 classified into the Red class;

In the public opinion algorithm flow described above, the intensity value being greater than 0 indicates that the information is positive, and the negative value negative. The more negative value of the subjects may cause the greater public opinion harm. This algorithm is convenient for the apartments to find out the harmful public opinion subjects and events from statistics of each public opinion class's percentage in one subject class and improve the effective of monitoring work. So the related government apartments will take more attention about the micro-blog subjects with large percent of negative documents.

VI. *Experiment and Results Analysis*

The accuracy is a common indicator to evaluate the performance of text classification. For a given category, 'a' is the number of the correct assigned instances of the class, and 'b' is the number of the mistakenly assigned to other class but belonged to the class, the accuracy rate (p) is defined as:

$$r = a / (a + b), \text{ if } a + b > 0; \text{ else } r = 1; \quad (2)$$

To verify the performance of the emotional intensity calculation algorithms, the paper makes uses of the KNN classification algorithm to class the documents from the MPO Corpus constructed by out lab, and takes the accuracy as the experimental results evaluation criteria.

TABLE III.
THE EXPERIMENTAL RESULTS TABLE

	Guo Degang's disciple beating incident	Film "The One 2"	TV "IPARTMENT"
RED	93.28%	94.13%	93.53%
ORANGE	93.79%	92.68%	92.47%
YELLOW	93.21%	93.59%	93.19%
GREEN	92.96%	93.35%	93.17%

The experimental results are shown as follows (in Table III and Fig.4).

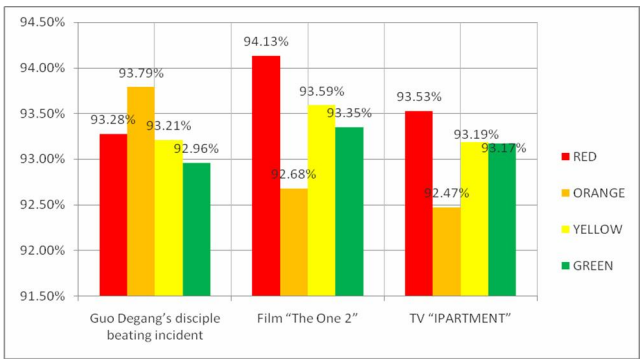

Figure4 the experimental results chart.

From the result in Table III and Fig. 4 shown above, a good classification accuracy is shown by the algorithm proposed in the paper. The public opinion classification based on the micro-blog subjects and document sentiment intensity has some advantages and feasibility. But because of the scale of the MPO Corpus relatively small, some parameters are not given accurately enough, which influences the classification efficiency.

VII. Conclusions and Future Work

In the paper, it analysis the status of network public opinion and the features of micro-blog in China, and proposes an approach to calculate the micro-blog documents emotional intensity firstly based on the analysis of documents from the semantic point. It calculates the intensity from three levels with words, sentences and documents based on the HowNet Knowledge Base. Due to the background research of the Classification information in the analysis of micro-blog public opinion, a classification algorithm has been developed from subjects and sentiment intensity taking the MPO Corpus as the data set. Finally, the experiment test for the algorithm indicates that the algorithm is practical and efficient.

In the future, we will keep our further research on how to improve the efficiency and accuracy of the algorithm discussed in this paper. Then, we will continue to study the key technical problems including context message intensity audit, public opinion transmission chain, text sentiment intensity experiment and emergency security strategy under the micro-blog platform environment.

ACKNOWLEDGMENT

First of all, we should show our great thanks to NSWCTC2011, we are pleased to have the opportunity to express our research in this journal. Our research work is supported by National Natural Science Foundation of China (Project Number. 61074135 and 60903187), Shanghai Creative Foundation project of Educational Development (Project Number. 09YZ14), and Shanghai Leading Academic Discipline Project (Project Number.J50103), Great thanks to all of our hard working fellows in the above projects.

REFERENCES

[1] Wikipedia [R]. http://en.wikipedia.org/wiki/Micro-blog.
[2] A.-H.Tan and P.Yu, A Comparative Study on Chinese Text Categorization Methods, PRICAI 2000 Workshop on Text and Web Ming, Melbourne,pp.24-35,August 2000
[3] Wu,Hanxiang, Xin Minjun, An approach to micro-blog sentiment intensity computing based on public opinion Corpus, unpublished.
[4] Qin Zhenhua, Xin Mingjun, Niu Zhihua. A Content Tendency Judgment Algorithm for Micro-blog Platform[C]. IEEE International Conference on Intelligent Computing and Intelligent Systems, in Xiamen, China, 2010.
[5] Chang Yi, Zhang Xin, Research and Implementation of Text Categorization System based on keyword Expressions, unpublished.
[6] Jian-Yun Nie, Jiangfeng Gao, Jian Zhang and Ming Zhou, On the Use of Words and N-grams for Chinese Information Retrieval. Proceeding of the fifth international workshop on Information retrieval with Asian languages, pages: 141-148, November, 2000, Hong Kong,China.
[7] LI Xuelei, ZHANG DONGMO, A Text Categorization Method Based on VSM, Computer Engineering, October 2003, Vol.29 No.17, pages: 90-92.
[8] K.L. Kwok, Comparing Representations in Chinese Information Retrieval. SIGIR'97, PAGES: 34-41, Philadelphia, Pennsylvania, United States.
[9] Zheng Wei, WANG Rui, Comparative Study of Feature Selection in Chinese Text Categorization, Journal of Hebei Norh University (Natural Science Edition) Dec.2007. Vol. 23 No. 6 pages: 51-64.
[10] Thorsten Joachims, A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Cateorization.Pages:143-151, Proceeding of ICML-97, 14th International Conference on Machine Learning.
[11] How net's Home Page. http:// www.keenage.com
[12] Qun LIU, Sujian LI, Word Similarity Computing Based on How-net [C], the 3rd Chinese Lexical semantics workshop, in Taipei, 2002
[13] Yiming Yang, A Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, Vol1 No. 1/2.69-90.
[14] Yiming Yan and Xin Liu. A re-examination of text categorization methods. Proceedings of SIGIR'99, pages: 42-49
[15] Liu Ying, The Application of Naïve Bayes in Text Classification Preprocessing, Computer and Information Technology, Dec.2010, Vol.18 No.6, pages:26-27.
[16] Jingnian Chen, HouKuan Huang, Shengfeng Tian, Youli Qu. Feature selection for text classification with Naïve Bayes. Expert Systems with Application 36(2009)5432-5435.
[17] Kim, S., Han, K., Rim, H., &Myaeng, s. (2006). Some effective techniques for Naïve Bayes text classification. IEEE Transactions on Knowledge and Data Engineering, 18(11), 1457-1466.
[18] SU Li-hua, ZHU Zhang-hua, BAI Wen-hua, Term Weighting Algorithm in Text Categorization Based on VSM, Computer Knowledge and Technology, Nov 2010, Vol.6 No.33, pp.9327-9329
[19] HU Xue-gang, DONG Xue-chun, XIE Fei, Method of Chinese text categorization based on the word vector space model, Journal of Hefei university of Technology, Oct. 2007, Vol.30 No.10, pages:1262-1264 .
[20] ZHANG Yun-liang, ZHANG Quan, Research of Automatic Text Categorization Based on Sentence Category VSM, Computer Engineering, Nov 2007, Vol.33 No.22, pages:45-47
[21] HUANG Xuan-Jing, XIA Ying-Ju, WU Li-De, A Text Filtering System Based on Vector Space method, Journal of Software, Vol.14, No.3, 2003, pages: 435-442.
[22] PANG Jian-feng, BU Dong-bo, BAI Shuo, Research and Implementation of Text Categorization System Based on VSM, Application Research of Computers, No 9, 2001, pages:

Xin Mingjun was born in 1970. received the PhD degree in computer science from Northwestern Polytechnical University, China. He is currently a vice professor in School of Computer Engineering and Science in Shanghai University, China.

His research interests include service computing, decision support system and information system.

Wu Hanxiang was born in 1985, earned B.S degree in the field computer science and technology in 2009 from Tianjin University of Technology and Education. He is currently a post student of School of Computer Engineering and Science in Shanghai University, China.

His research interests include web services security, content audit, public opinion monitor.

Li Weimin was born in 1972, he received the PhD degree in computer science from Donghua University, China. He is currently a researcher in School of Computer Engineering and Science in Shanghai University, China.

His research interests include service computing, data stream management, sensor networks, and massive data management.

Niu Zhihua was born in 1976. She received her Ph.D degree at Xidian University. Now shi is a lecturer at the School of Computer Engineering and Science, Shanghai University.

Her main research fields are cryptography and information security.

