I.J. Wireless and Microwave Technologies, **2012, 5, 53-61**
Published Online October 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012.05.08

Available online at http://www.mecs-press.net/ijwmt

An Approach to Micro-blog Sentiment Intensity Computing Based on Public Opinion Corpus

Wu Hanxiang[a,*1], Xin Mingjun[a,*2], Li Weimin[a,*3], Niu Zhihua[a,*4]

[a] *School of Computer Engineering and Science, Shanghai University, Shanghai 20072, China*

Abstract

Based on the analysis of the status of network public opinion, the features of short content and nearly real-time broadcasting velocity in this paper, it constructs a public opinion corpus on the content of micro-blog information, and proposes an approach to marking corpus on the basis of sentiment tendency from the semantic point of view; Furthermore, considering the characteristics of micro-blog, it calculates the sentiment intensity from three levels on words, sentences and documents respectively, which improves the efficiency of the public opinion characteristics analysis and supervision. So as to provide a better technical support for content auditing and public opinion monitoring for micro-blog platform.

Index Terms: Micro-blog; Public Opinion Corpus; Sentiment Intensity

1. Introduction

Micro-blog is a kind of blogging variants arising in recent years. It gains more and more attention and recognition for its short format and real-time characteristics, and becomes an important platform for public opinion expression. 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 and 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. Therefore, new challenges are brought to the government monitoring public opinion trends and discovering public opinion crisis. At present, research on micro-blog for public opinion in China has

* Corresponding author:
E-mail address: *1,*2,*3 {newwhx, xinmj, wmli}@shu.edu.cn; *4 zhniu@staff.shu.edu.cn

just started, and lacks of sophisticated systems and applications. To study and analysis micro-blog text semantic tendency, this paper constructs a public opinion corpus on the content of micro-blog information, and proposes 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, it proposes three algorithm on computing sentiment intensity on words, sentences and documents respectively. Thus, it can improve the efficiency of the public opinion characteristics analysis and supervision.

2. Related Work

Corpus is an order set of documents, which is the foundation of text classification, retrieval, synthesis and comparison, raw materials for further language analysis or other work [2]. Corpus studies in foreign countries have started very early, and completed many large-scale corpuses, such as the BROWN corpus constructed by Brown University, the United States, and the LOB corpus jointly established by Lancaster University, UK and University of Oslo and Bergen University, Norway, and so on. Chinese corpus studies start from 1920s when some scholars had established text corpuses, using statistical methods to study the frequency of Chinese characters. Those were not machine-readable corpus, and the scales were very small. Until 1979, the construction of machine-readable corpus began to be studied, such as Chinese Modern literature corpus of Wuhan University, and Modern Chinese corpus in Beijing University of Aeronautics and Astronautics and so on [3].

How to Build Knowledge Base is the key problem for semantic corpuses construction. 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. 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. Steps of the corpus construction will be discussed in Section 3, emotional strength calculation algorithm in section 4, and conclusions and further work in section 5.

3. A Public Opinion Corpus Constructing

Generally corpus construction process includes 4 main modules: the collection and pre-process, labeling specifications formulation, corpuses marking and quality control and so on [5]. In this paper, it collects micro-blog texts as the source corpus entities, and marks every entity on the basis of HowNet as the common sense knowledge base from the perspective of the text sentiment tend. The corpus construction flow chart is shown in Fig. 1.

3.1. Corpus Collection and Preprocessing

The first problem for Corpus selection and collection is the coverage of the sources. Recently, the market's leading suppliers of micro-blog include twitter abroad, and some internal providers such as Tencent, Sina, Souhu, etc. With the micro-blog's features of short content and relatively colloquial style, this paper selects the sources with clear emotions tend and fitting in with Chinese syntax rules as much as possible. The number of micro-blogs with all kinds of emotional tendencies is equal more or less. The following table (Table 1) is the basic information of the initial corpus sources distribution.

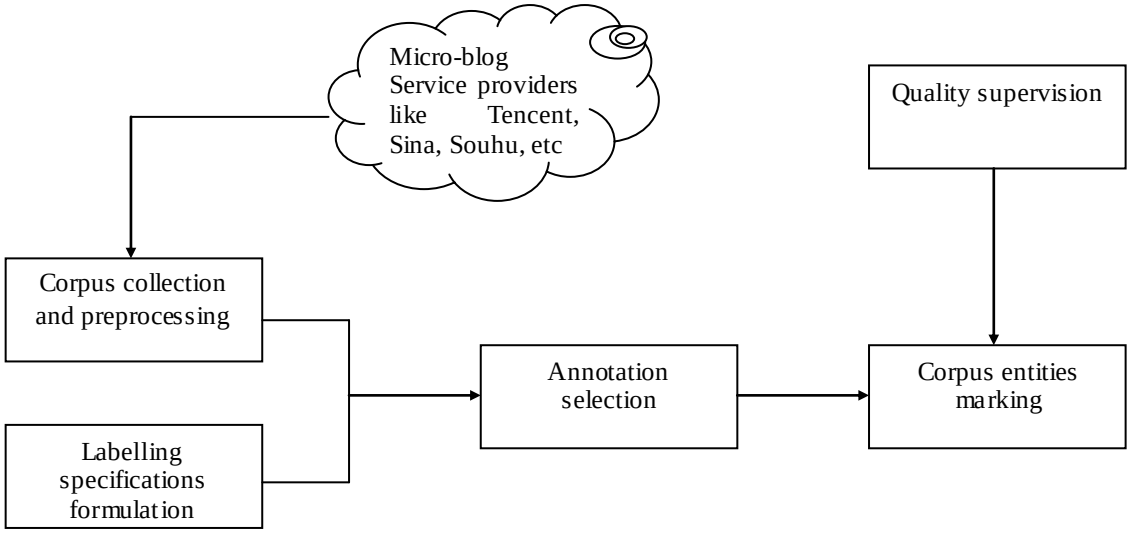

Fig. 1. Corpus onstruction flow chart

Table 1. 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

With many colloquial terms, and inaccurate punctuations or sub-sentences, the sources would be pre-processed before being marked into corpus. And some pre-process rules are instituted as follows:

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

2) Each context sentence holds one line.

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

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

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

3.2. Corpus Annotation

Corpus annotation is a process that fills necessary information in the source corpuses and reveals the hidden semantic knowledge, preparing for modeling documents. In principle the corpus annotation specifications are determined before the corpus annotation, but in fact, with the complexity of the origin corpus, the final specifications in the corpus are determined by a number of repeated pre-marks with amendments and modifications.

This paper uses the Institute of Computing Technology and Chinese Lexical Analysis System (ICTCLAS) designed by the Chinese Academy of Sciences to divide sentences of the initial corpus sources into words and tagging with parts of speech. The next corpus marking job includes two aspects: the one is about the basic information of micro blog, including author information, source, subject, keyword, parts of speech, etc; the other is about the semantic and sentiment intensity information of the text based on the common sense knowledge and HowNet Emotion word set.

1) Basic Information

It considers a micro-blog document as a doc, consisting of head and body in this paper. The basic information contains meta-information of the micro-blog and the segment information of the text. The head is marked about the meta-information, including the serial number in the classification of information, author information, source information, subject information, list of keywords, emotional intensity and so on. And the body is about micro-blog text labeling, composed by a series of sentences. Sentences labeling is the core of the corpus annotation including the index of sentences in the doc, sentence length, the original text, word text after the segment, rhetoric, agent, patient, and semantic annotation.

2) Semantic Information

Based on HowNet and sentences segment, the words annotation of each sentence consists of two parts: syntax and semantic. The former is the syntax part of words, including the serial number of words, the start position, the length of the and parts of speech in the sentences. The semantic part includes the concept annotation based on HowNet, and sentiment tendency for emotional words based on HowNet emotion word set.

HowNet makes use of the Knowledge Dictionary Mark-up Language (KDML) formally descript concepts of words. The feature at the first position in the definition of words' concept is the natural attribute, revealing the concept's class. Features appearing at other positions in the definition are the characteristics attributes. Such as:

Man: DEF = { human， male}

In the definition of concept of 'Man', 'male' is the nature attribute of 'Man'.

In HowNet some special identifiers are used to describe the special relationship between the concepts of common sense, such as: coded with pointer '%' before the whole part meaning part-whole, coded with pointer '&' before the host meaning attribute-host, coded with pointer '?' before the product meaning material-product.

There are many words' concepts having more than one definitions in HowNet, which makes the corpus annotation difficult. In this paper, it abides by the following measures: 1) Reference to Context of words. 2) Reference to the Sememes' Hypernym-Hyponym. 3) Reference to the Concepts' Event-role. 4) Reference to the English Description of the Concept

It distinguishes each word's emotion tendency between positive and negative words based on HowNet sentiment analysis set in this paper. It sets the following two rules when labeling semantic properties polarity: One is about prepositions, auxiliary, pronouns and other non-emotional tendencies, polarity = 0; The other is about nouns, verbs, adjectives and adverbs, based on the calculation results of similarity with words in HowNet sentiment analysis set: polarity(negative words) = -1, polarity(positive words) = 1, polarity(neutral words) = 0. At Doc and sentence level the paper makes emotional intensity. The following is the formal description of Corpus annotation specification:

---

doc = (head, body)
head = (index, author, date, source, topic, keywords, intensity)
body = (sentence1, sentence2, sentence3, ……, sentenceN)
sentence = (opinion, rhetoric, agent, patient, segment , word1, word2, word3, ……, wordN)
word=(syntax, semantic)

3.3. XML Format Annotation

One of the representatives of Corpus Linguistics Leech holds that corpus annotation should follow the basic principles that the additional code can be deleted to be restored to the original corpus; The end-users should understand the annotation specification and means of the corpus annotation codes; Labels should be selected which is widely accepted as a neutral mode; Any annotation pattern cannot be used as the first standards and so on. The semi-structured XML format is made to store the marked corpus entries. A corpus labeled sample is shown below (Fig. 2).

```
<doc>
    <head>
        <index>1</index>
        <author>刘晓</author>
        <date>201101081427</date>
        <source>http://t.qq.com</source>
        <topic>非诚勿扰2</topic>
        <keywords>喜欢</keywords>
        <intensity>0.8</intensity>
    </head>
    <body>
        <sentence s_no=1 s_len=25 orgin="《非诚勿扰2》里，我很喜欢一句话：活着是一种修行。">
            <opinionFact>主观</opinionFact>
            <intensity>0.8</intensity>
            <rhetoric>无</rhetoric>
            <agent>我</agent>
            <patient>非诚勿扰2》里的一句话</patient>
            <keywords>喜欢</keywords>
            <segmentation >《/w 非诚勿扰2/nz 》/w 里/f ，/w 我/r 很/d 喜欢/v 一/m 句/q 话/n ：/w 活着/v 是/v 一/m 种/q 修行/v 。/w   </segmentation>
            <word w_no=1  start=1 w_len=1><syntax class="w"><semantic class="标点，《" polarity=0>《</semantic></syntax></word>
            <word w_no=2  start=2 w_len=5><syntax class="nz"><semantic class="电影，专" polarity=0>非诚勿扰2</semantic></syntax></word>
            <word w_no=3  start=7 w_len=1><syntax class="w"><semantic class="标点" polarity=0>》</semantic></syntax></word>
            <word w_no=4  start=8 w_len=1><syntax class="f"><semantic class="位置，内，？" polarity=0>里</semantic></syntax></word>
            <word w_no=5  start=9 w_len=1><syntax class="w"><semantic class="标点" polarity=0>，</semantic></syntax></word>
            <word w_no=6  start=10 w_len=1><syntax class="r"><semantic class="人，第一人称，我" polarity=0>我</semantic></syntax></word>
            <word w_no=7  start=11 w_len=1><syntax class="d"><semantic class="属性值，程度，很" polarity=1>很</semantic></syntax></word>
            <word w_no=8  start=12 w_len=2><syntax class="v"><semantic class="喜欢" polarity=1>喜欢</semantic></syntax></word>
            <word w_no=9  start=14 w_len=1><syntax class="m"><semantic class="数量值，多少，单" polarity=0>一</semantic></syntax></word>
            <word w_no=10 start=15 w_len=1><syntax class="q"><semantic class="名量，词语" polarity=0>句</semantic></syntax></word>
            <word w_no=11 start=16 w_len=1><syntax class="n"><semantic class="语文，说" polarity=0>话</semantic></syntax></word>
            <word w_no=12 start=17 w_len=1><syntax class="w"><semantic class="标点" polarity=0>：</semantic></syntax></word>
            <word w_no=13 start=18 w_len=2><syntax class="v"><semantic class="活，长时间" polarity=0>活着</semantic></syntax></word>
            <word w_no=14 start=20 w_len=1><syntax class="v"><semantic class="是" polarity=0>是</semantic></syntax></word>
            <word w_no=15 start=21 w_len=1><syntax class="m"><semantic class="数量词，多少，单" polarity=0>一</semantic></syntax></word>
            <word w_no=16 start=22 w_len=1><syntax class="q"><semantic class="名量，类型" polarity=0>种</semantic></syntax></word>
            <word w_no=17 start=23 w_len=2><syntax class="v"><semantic class="实施，宗教" polarity=0>修行</semantic></syntax></word>
            <word w_no=18 start=25 w_len=1><syntax class="w"><semantic class="标点" polarity=0>。</semantic></syntax></word>
        </sentence>
    </body>
<doc>
```

Fig. 2. Example of a micro-blog corpus entity

Entire corpus is marked between <doc> and </ doc>. Information between <head> and </ head> is about the meta-information part of the micro-blog, such as the index number ‘index’, author information ‘author’, source ‘source’, keyword-list ‘keywords’, theme ‘topic’, emotional intensity ‘intensity’. Information between <body> and </ body> is about the micro-blog text part, in which <sentence> and </ sentence> is for sentence tags. The origin sentence in the markup model can restore from the original properties ‘origin’. It could get words frequency information, part of speech information and emotional information. And the performance of the calculation sentiment intensity algorithm will be tested and verified in the next section.

4. Sentiment Intensity Computing

Sentiment intensity computing is the foundation of text classification (the future work). This paper calculates the sentiment intensity of micro-blog from three levels: words, sentences and documents emotional strength.

4.1. 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. As the emotion difference between the evaluation words and emotional words are not very obvious, the sentiment analysis words are merged into the positive words set and negative words set, such as:

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

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

And set 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 include emotional information, but modify emotions degree intensity. So this paper gives these words a real number weight between 1 and 10. The calculation 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;*

4.2. 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, this paper studies 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. A common Chinese phrase type is prejudiced phrase. To compute the emotional intensity, the paper obeys the following rules: 1) the emotional strength of the parallel structure phrases such as: "noun + noun", "adjective + adjective" is equal to the sum of the each word' strength. 2) 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: One is that each sentence is a single sentence, and complex sentences composed by the conjunction artificially are split into two sentences; And the other is that the similarity based on HowNet is increased by 10 times. Sentence emotional intensity is calculated 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);
}
```

4.3. 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, βi is the correlation coefficient):

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

To reduce artificial labeling errors, a java-based Corpus Tagging System is designed to improve the efficiency and accuracy. The system reads a text file and outputs an xml file after artificial annotation. The system interface's screenshot is shown below (Fig. 3).

Fig. 3. The Tagging System Interface

4.4. Experimental 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 test the performance of the emotional intensity calculation algorithms, this paper uses KNN classification algorithm to class the documents from the corpus constructed above based on emotion intensity, and takes the accuracy as the experimental results evaluation criteria. The experimental results are shown as follows (Table 2).

Table 2. The Result of the Test Experiment

Source	Sina	Tencent	Souhu
positive	17	23	18
negative	33	27	32
accuracy	92.34%	94.25%	91.78%

It sets 0 as the default threshold. If the sentiment intensity of document is greater than 0, then the document is taken to be a positive one. If the intensity is less than 0, then it is judged to be a negative document. As shown in Table 2, the results indicate that the proposed algorithms have a higher accuracy and practicality.

5. Conclusions and Future Work

To improve the efficiency of the public opinion characteristics analysis and supervision based on micro-blog platform, firstly in this paper, it constructs a public opinion corpus which has 150 entries. Secondly it analyses the semantic of the text on the basis of HowNet and proposes the emotional intensity algorithms by computing emotion intensity of words, sentences and documents. Finally, the experimental results show that the proposed algorithms have certain advantages and feasibility. In the future research work, we will pay more attention to complete another 500 micro-blog entries marking job. Any corpus cannot be perfect. There must be some faults and deficiencies, we will improve and refine the corpus.

Acknowledgements

This 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] Xu Rui, The Research and Construction of a Chinese Semantic Corpus. SuZhou University, 2006

[3] Zhiwei Feng, Evolution and Present Situation of Corpus Research In China. Journal of Chinese Language and Computing (in Chinese), vol. 12, No.1. pp. 43–62, July 2002.
[4] How Net [R]. How Net's Home Page. http:// www.keenage.com
[5] Lin-hong XU, Hong-fei LIN. Construction and Analysis of Emotional Corpus, Journal of Chinese Information Processing (in Chinese), vol. 22, NO. 1, pp. 116-122, Jan 2008