Recognition Method for Emoticons Based on Network Text

Yiming Lei [a], Yong Liu, Hua Huo

College of Information Engineering, Henan University of Science and Technology, Luoyang, China

[a]lymhunt@163.com

Keywords: Emoticon; Recognition method; Conditional entropy; Mutual Information; Network text; UTF-8 code

Abstract. Emoticons are widely used in the network text, which has enhanced effect for the author's emotional experience and semantic representation. Effective identification of the emoticons in the network text has great significance to sentiment analysis, public opinion survey and commercial investigation. Through analyzing the information about a number of emoticons on the Internet and based on the UTF-8 coding, new method of recognition emoticons is presented. First of all, the emoticons are differentiated and analyzed by using the statistical method, and then, the method describes the process by the steps of conditional information entropy are defined, computing mutual information and filter the candidates with using written rules. Finally, experimental results prove that the recognition algorithm has correctness, precise and practicality.

Introduction

With the development of computer technology and network technology, when people interact and communicate by using the computer network, they cannot get the additional information like expression, action or voice tone from others, this situation bound to reduce the efficiency of the communication [1]. To make up for the defects, many Internet users start to use the emoticons.

In the past papers about Chinese natural language processing, researchers [2-5] usually see emoticons as ordinary symbols or noise strings [6]. But when process the corpus which contains lots of network language such as the emoticons for the research of sentiment analysis, public opinion survey and commercial investigation, the accuracy of the semantic understanding and emotional expressions for the whole text will be decreased if the emoticons are fully removed. Therefore, when new words discovery is studied, emoticons should be identify and define as vocabulary [7, 8], this has practical importance for the whole semantic understanding and sentiment analysis when the network corpus and Internet phrases are processed by using the Chinese information processing system.

Definition and Analysis of Emoticons

Emoticon is non-traditional word as a pictorial representation of a facial expression composed of punctuation marks, number, letter or words. It can simulate specific images to express the writer's feeling, mood and actions [9]. For example, "눈_눈" just like a person's face, "o(*≧▽≦)ツ" means happy. Emoticons could appear in the sentence to assist the web text writers express a particular emotion or independent appear to express the unique meaning [10].

Emoticons could be divided into three categories which are shown in Table 1 by analyzing more than fifty kinds of different emoticons.

Table 1 Three kinds of emoticon

	Name	Example	Meaning	UTF-8 code
First	Normal Continuous characters	^_^ T_T	smile cry	^ => &#x5E _ => &#x5F T => &#x54
Second	Include some language words	(/TДT)/ 눈_눈	cry stare	Д => &#x414 눈=> &#xB208
Third	Based on the brackets as the core	o(*≧▽≦)ツ (。・`ω´・)	happy cheer	ツ => &#x30C4 ω => &#x3C9

The Method of Emoticons Discovery

Establish The Candidates Set Because the corpus for experiment and research is Chinese language, all the elements in the emoticons can be divided into two categories: Chinese character and non-Chinese character with using UTF-8 code. The appear position of them is i, the non-Chinese character in the corpus is C_i, to $n \in i$, a candidate emoticon is $(C_1,C_n)=C_1C_2\ldots C_n$. The rules of emoticons appear is (words)+'emoticons'+(words) in the corpus, emoticons appear between words in the sentence or be independent. So the candidate set could be established by computing the continuous non-Chinese characters and collect them into with using Eq. 1.

$$(C_1, C_n) = \{C_1 \mid C_1 \in corpus, C_n \mid C_n \in corpus, 1 < n\} \tag{1}$$

Sieve Method Based On Conditional Entropy Most elements in the emoticons are symbols, punctuation marks or other languages words, to the Chinese corpus, these elements have no real meanings and have a lower frequency of them. But the emoticon has a higher coupling degree between each element because its significance. So a sieve method is established based on conditional entropy (*CE*). To an element *X*, the frequency of the occurrence *P(x)* is random distribution in the corpus. The entropy of *X* is:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \tag{2}$$

To another element *Y*, the conditional entropy of *X* is (2) with the condition of *Y*:

$$H(X \mid Y) = -\sum_{x \in X, y \in Y} P(x, y) \log P(x \mid y) \tag{3}$$

In the same way, to the third element *Z*, the conditional entropy of *X* is:

$$H(X \mid Y, Z) = -\sum_{x \in X, y \in Y, z \in Z} P(x, y, z) \log P(x \mid y, z) \tag{4}$$

According Eq. 2, Eq. 3, and Eq. 4 can filter the first and second kind of candidate emoticon in Table 1.But for the third case, conditional entropy has been unable to meet the requirements because brackets is a common symbol and has a higher frequency in the corpus. If a candidate emoticon is the third kind, the frequency of elements beside the brackets should be similar. So a new suitable sieve method is established according the mutual information (*MI*) between the elements which are beside the bracket. To one candidate emoticon of the third kind, its left-bracket is *LB*, the *LB*'s left neighbor is LB_{left}, *LB*'s right neighbor is LB_{right}, same to the right-bracket. For example, to the emoticon (ノ #-_-)ノ , "*NULL*" is LB_{left}, "ノ "is LB_{right}, RB_{left} is "-", "ノ " is RB_{right}.

$$MI(LB) = I(LB_{left}; LB_{right}) = \sum_{x \in LB_{left}, y \in LB_{right}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \tag{5}$$

$$MI(RB) = I(RB_{left}; RB_{right}) = \sum_{x \in RB_{left}, y \in RB_{right}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{6}$$

$$DV = \left| \frac{MI(LB)}{MI(RB)} - 1 \right| \tag{7}$$

The *DV* (difference value) more closed to zero, the more similar of the frequencies of the four elements beside the two brackets and the candidate more likely is a real emoticon.

Filtering Method Based On Written Rules There will have some meaningless character combination of low frequency in the candidate set because of the write error, keyboard wrong operation, etc. These characters combinations have no significance and can never be the emoticon. Filter processing these low-frequency characters combinations is conducive to improve the precision.

When people write text online, some Internet users will adopt the way of continuous writing the same punctuation marks to assist in semantic expression like "!!!" or "???". These strings of the continuous same symbols is not in conformity with the definition of emoticons, so filtering them from the candidate set is conducive to improve the precision.

Experiment and Analysis

To test and verify the correctness of the method, the content of Sina Weibo in July, August and September in 2015 with using web crawler tools to collect is used as the experimental corpus. Sina Weibo is the most frequently used social platform and most of writer are young people which are always use the emoticons. The evaluation of experimental results using symbol *P* (precision rate), *R* (recall rate) and F_1 *score* (also *F-score* or *F-measure*). In statistical analysis of binary classification, the F_1 *score* is a measure of a test's accuracy. It considers both the *P* and the *R* of the test to compute the score: *P* is the number of correct discovery results divided by the number of all discovery results as shown in Eq. 8, and *R* is the number of correct discovery results divided by the number of all the correct results that should have been discovered from the corpus as shown in Eq. 9. The F_1 *score* can be interpreted as a weighted average of the precision and recall, where an F_1 *score* reaches its best value at 1 and worst at 0 (Eq. 10).

$$P(precision) = \frac{|D(E) \cap D^*(E)|}{|D(E)|} \tag{8}$$

Where,

D(E) is all of the emoticons which are the results in this experiment.

$D^*(E)$ is all of the emoticons which are should be discovered in the corpus in this experiment.

$$R(recall) = \frac{|D(E) \cap D^*(E)|}{|D^*(E)|} \tag{9}$$

$$F = \frac{2PR}{P+R} \tag{10}$$

The candidate emoticons set can be gained according to Eq. 1, iterative screen the set by using conditional entropy (*CE*) and mutual information (*MI*), and in the end with filtering method based on written rule to get the emoticons from the corpus.

R and *P* change with the variation of threshold value (*TV*) of *DV* in the process of data processing, *P&R* does not increase or decrease at the same time when the *TV* changed as shown in the Fig. 1. With F_1 *score* is an evaluation criteria of the precision rate and recall rate, the best fit of threshold value could be found when *F* reaches maximum value.

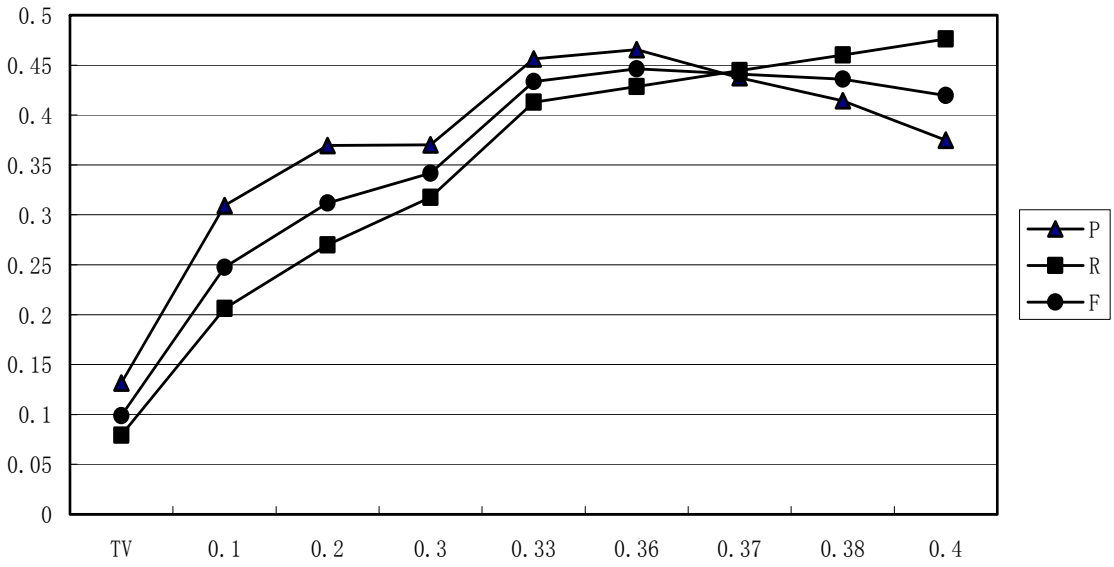

Figure 1 the value of *R&R&F* from different *TV*.

Can be seen in the Fig. 1, the precision rate and recall rate changed with the different threshold values. The value of *R* increase with the improving of the *TV* and *P* reduced after rise before. 0.45 is the maximum value of *F*, and *TV* is 0.37 corresponding to the *max-F*.

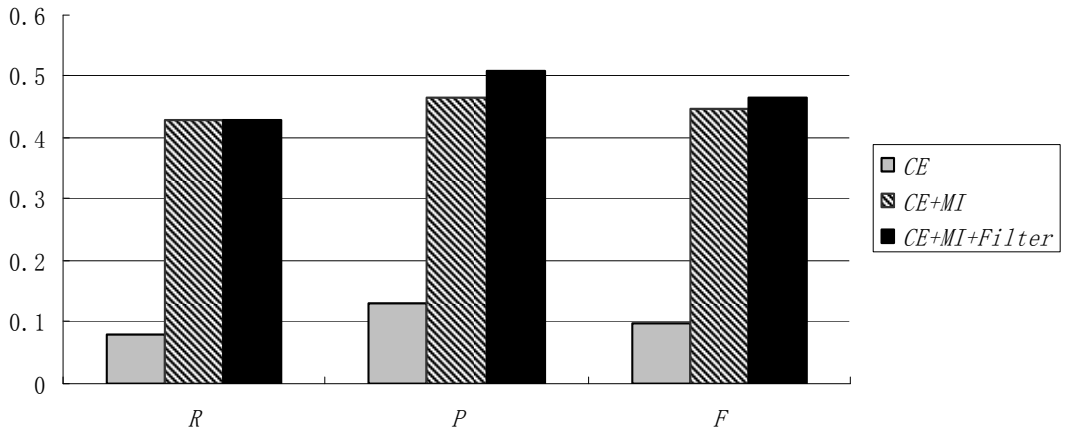

Figure 2 The results from different steps.

The result of this experiment with the best fit of threshold value (*TV*=0.37) is shown in the Fig. 2. It could be seen through the different steps, recall and precision have the corresponding increase, and all of *R&R&F* are the maximum at the situation of "*CE+MI+Filter*", *R&P* have a corresponding increase by using the method of mutual information (*MI*), and the enhancement of *R&P* prove the filter method is also effective, the results show the effectiveness of the present method.

Conclusion

As a kind of words which often used by people, the emoticon is an effective complement of the natural language, have a large number in the network text and be widely used, this paper try to put the emoticons as new words for research and the feasibility of the method is verified by using the network text experiment, this method provided a useful reference for Chinese natural language processing.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (60743008), the Key Scientific and Technological Projects of Henan Province in China (142102210045).

References

[1] A.J.Kan: Journal of Huaibei Normal University (Philosophy and Social Sciences), Vol.34(2013)No.4, p.84-86.(in Chinese)

[2] S.Huo, M.Zhang, Y.Q. Liu, et al: Pattern Recognition & Artificial Intelligence, Vol.27(2014)No.2, p.141-145.(in Chinese)

[3] Wu Yue, Yan Pengju and Zhai Lufeng: Journal of Tsinghua University, Vol.51(2011)No.9, p.1317-1320.(in Chinese)

[4] J.H.Zheng, W.H.Li: Journal of Shanxi University, Vol.25(2002)No.2, p.115-119.(in Chinese)

[5] G.Zou, Y.Liu, Q.Liu, et al: Journal of Chinese Information Processing, Vol.18(2004)No.6, p.1-9.(in Chinese)

[6] Z.Y.Jia, Z.Z.Shi: Computer Engineering, Vol.30(2004)No.20, p.19-21.(in Chinese)

[7] J.Zhao, L.Dong, J.Wu, et al: ACM SIGKDD international conference on Knowledge discovery and data mining.(2012), p.1528-1531.

[8] M.Jakob, G.Luderer, J.Steckel, et al: Affective Computing IEEE Transactions on, Vol.1(2010), p.46-59.

[9] Information of Emoticon on https://en.wikipedia.org/wiki/Emoticon.

[10] M.Kawakami: Human Science Research Bulletin of Osaka Shoin Womens University, Vol.7(2008), p.67-82.(in Janpanese)

