Sentiment Orientation Classification of Webpage Online Commentary Based on Intuitionistic Fuzzy Reasoning

Xiaofeng Li
Department of Computer Science and Technology,
ChengDong College of Northeast Agricultural University
Harbin, 150025, China
e-mail:Mberse@126.com

Dong Li
Department of Computer Science and Technology,
Harbin Institute of Technology
Harbin, 150001,China
e-mail:lee@hit.edu.cn

***Abstract*—An approach of sentiment classification for online comments based on intuitionistic fuzzy reasoning is presented on the basis of the analysis of characteristics of sentiment classification. The approach employs membership function, non-membership function and hesitant function to depict uncertainties of features, quantitatively, by sample training, as well as sentiment expressions influenced by adverbs of degree, conjunctions and negative words are considered. Then the semantic orientation of a text is synthesized on the level of phrases, sentences and texts in sequence by means of aggregations of intuitionistic fuzzy information of features. The presented approach obtains high precision and recall when test using public corpus**

Keywords Chinese information process; text categorization; sentiment orientation; intuitionistic fuzzy set

I. INTRODUCTION

Internet popularization and emergence of many new network media have made people access a huge amount of information and also provided various platforms for them to express opinions or feelings, such as BLOG, BBS and online comment platform. Therefore, how to manage effectively and scientifically online comments on websites which are available for the discharge of personal sentiments matters a lot to individuals, enterprises and social security as well [1]. However, the management of texts on those websites is significantly different from that of common ones by virtue of these features:

A they have no fixed grammatical structure and are generally short, even with new words coined, which is nicknamed new text [2];

B the top priority to that management is to understand netizens' sentimental attitudes towards the principal body of commentary, for instance, what film-making companies care about the most is whether audiences love the film and a hotel manager is concerned firstly with customers' degree of satisfaction at the hotel service.

Texts or new texts can be classified in many ways. Fuzzy decision analysis theory has applications in the text classification and clustering. In the paper [3], DS evidence theory, fuzzy set theory and fuzzy -rough set theory were applied to improve traditional k-NN method to have achieved higher accuracy and recalling rate. The paper [4] introduced an efficient text TF vector reduction algorithm based on rough set attribute reduction algorithm and used it for automatic categorization. Similar studies revealed that through the description of fuzziness and roughness of classification, the effect could be enhanced. However, no work has been done on sentiment orientation classification with the use of those theories. Differing from the conventional fuzzy set, intuitionistic fuzzy set considers both the membership and non-membership of elements belonging or not belonging to such set. It can determine more flexibly the uncertainty of classification decision problems. Here we discussed the use of intuitionistic fuzzy theory for text sentiment orientation classification, which could specify the degree of feature-supportive texts subject or not subject to one category. Besides, after processing quantitatively degree adverbs, transitional words or negative adverbs, it improved the accuracy of classification.

II. BASIC HYPOTHESES FOR SENTIMENT ORIENTATION CLASSIFICATION OF WEBPAGE ONLINE COMMENTARY

For any web text, it is common to classify by just basing on texts themselves such as text caption and the body, instead of considering other expression forms which could represent emotional tendencies too, e.g. web elements like grading or rating tab. As a matter of fact, if the alike element is contained in webpage commentary, there must be the two possibilities:

A It's possible to judge the emotional tendency of corresponding texts, in no need to analyze them;

B Words are the content to be really delivered, e.g. on Taobao.com, many make good comments but only dissatisfaction is presented in the content box. In this case, texts themselves are the only consideration. So from the viewpoint of the author, it is meaningless to think these comments positive: facilities in the room are older; the air-conditioner is noisy; guests living in superior rooms are charged for swimming in the pool, which is unacceptable as it's as high as in the outside; water of the pool is too deep; nobody cares about it even if it's mentioned; breakfast is just-so-so.

Above all to text sentiment orientation classification is the writer's emotional inclination he wants to signify, or that of majorities. To a specific text, apart from pos or neg classification, being neutral is an emotional attitude too. It is insignificant to sort a text whose sentiment inclination can be hardly itemized intuitively. Take for instance these comments: the room is good; hot spring is too bad and

Published by Atlantis Press, Paris, France.

costly; it's satisfying to see a few squirrels. Another is film reviews, of which one piece is an ads excerpted from other movie sites. It is obvious the comment should not be classified into any one of them: pos or neg.

Just as demonstrated in the work [5], no one sorting algorithm is stable, in other words, on the one hand to such methods, it may be the best to find the most suitable method as much as possible for a given context but not intentionally resort to the universal and steady one. For example, since characteristics of text classification and music classification are represented in different ways, their applicable algorithms should differ from each other too; on the other hand, a classifier which is trained under this environment may hardly work well when applied in others. Training corpus and testing corpus we are utilizing in the paper are different corpora in the same background.

III. THE METHOD BASED ON INTUITIONISTIC FUZZY REASONING

Basic concepts, operational rules and comparison methods of intuitionistic fuzzy set and aggregation operator of intuitionistic fuzzy information were introduced in the work [6] and [7]. We won't mention again.

A. Description of classification problems

Use a training method to get classifier before using it to categorize texts. Supposing the classifier has L features like, $ch_l, l = 1,2,\cdots, L$. The probability of the feature ch_l in support of text belonging to pos or neg class is expressed by an intuitionistic fuzzy set as: $\langle ch_l, u_l, v_l \rangle$, where, u_l and v_l stand respectively for the minimum probability of the text subject to pos or neg class during the existence of the feature ch_l, $0 \le u_l \le 1$, $0 \le v_l \le 1$ and $0 \le u_l + v_l \le 1$. For the to-be-categorized text *x*, mark participles and part of speech as to find out feature words. For the convenience of weight computation and composition in the following part, we firstly divide *x* into *N* sentences with punctuation marks (space ignored). Set M_n feature words in the n^{th} sentence as x_{ni} $(i = 1,2,,M_n)$. The feature x_{ni} has K_{ni} modifiers (limited to degree adverbs). The task of this problem is to determine sentiment orientation classification of text *x* with the help of information about feature words, degree adverbs, transitional words and negative adverbs.

B. Determination of intuitionistic fuzzy set

As to the feature ch_l $(l = 1,2,\cdots, L)$, the membership and non-membership grade of its corresponding intuitionistic fuzzy set can be defined through following expressions:

$$u_l = P(pos|ch_l) = \frac{P(pos, ch_l)}{P(ch_l)} \quad (1)$$

$$v_l = P(neg|ch_l) = \frac{P(neg, ch_l)}{P(ch_l)} \quad (2)$$

Apparently,since $P(pos, ch_l) + P(neg, ch_l) \le 1$, $0 \le u_l + v_l \le 1$, $\langle ch_l, u_l, v_l \rangle$ can constitute intuitionistic fuzzy set. Yet, it is improper to use directly the statistics for categorization. That is because, say, for this approving word "good", which has two membership degrees-0.7594 and 0.2304, we can't deem the possibility for the word to support text subject neg class is 0.2304. It is unreasonable. In the paper, it stipulates that the non-membership degree of approving words is 0 and the membership degree is 0. Thus, the intuitionistic fuzzy set $\langle ch_l, u_l, v_l \rangle$ means the degree of the feature ch_l supporting the text belonging to pos and neg category, in which u_l, v_l has and only has the one degree, 0. The degree of membership or non-membership of approving words and derogatory words can be decided intuitively or through the method in [8].

C. Treatment of degree adverbs, transitional words and negators

Adverbs of degree play a critical role in suggesting emotional inclination, e.g. a little expensive and rather expensive, representing greatly different emotions. For the computation here, degree adverbs are regarded as one of the main factors to the weight of those feature words that are modified by them and labeled with numbers from 1/9 to 9. If the weighting coefficient of a degree adverb is less than 1, it means it has alleviative effects on the sentimental color of features. On the contrary, it has intensive effects. Based on the knowledge in cognitive psychology and linguistics, we used these degree adverbs and their weighting coefficients for this experiment as shown in Table I..

Table I.. Degree adverbs and weighting coefficients

Adverbs of Degree	Weighting Coefficients
Extremely, utterly, fantastically	9
Extraordinarily	7
Exceptionally	5
Very much, very	3
Fairly, quite, rather	2
A little bit	1/3
Slightly	1/5

Some conjunctions especially transitional words often imply the controlling idea of text. They are very few and can be directly sorted out in advance. Likewise, we can choose numbers from 1/9 to 9 to signify their roles. For the convenience of the experiment, we disregarded the role of connectors which suggest coordinating and incremental relations as during the composition of text sentiment

orientation, such relations have been embodied. Transitional words we used for the test and their weighting coefficients are noted in Table II.

Table II. Transitional words and weighting coefficients

Transitional Words	Weighting Coefficients
Generally speaking	8
But	6
However	3
Only	1/3
Except for	1/5

When classifying feeling inclination, negators cannot be neglected as stop words. If the intuitionistic fuzzy set in correspondence to a feature is $\langle ch_l, u_l, v_l \rangle$, then, the intuitionistic fuzzy set relative to the phrase "negator+ch_l" should not be visually the complementary set $\langle ch_l, v_l, u_l \rangle$, but we set $\langle ch_l, 0.8v_l, 0.8u_l \rangle$.

What's more, we set the rule in the experiment: the action range of a degree adverb is the first feature word just following it; while the range of a transitional word is from where it appears to where the next one appears or to the end of a text.

D. Steps of classification

Step one: use the above-mentioned method in *C* to treat negative words after pre-processing texts.

Step two: sentiment orientation composition at phrase levels. Separate a sentence into phrase set with the advantage of the feature word x_{ni}, i.e. the n^{th} sentence contains totally M_n phrases, where, $i = 1,2,,M_n$, $n = 1,2,\cdots,N$. Set the feature word x_{ni} corresponding to intuitionistic fuzzy set $\langle ch_x, u_x, v_x \rangle$ and K_{ni} weighting coefficients corresponding to modifiers of that word w_j, where $j = 1,2,,K_{ni}$. Here, we can use the intuitionistic fuzzy number (u_{ni}, v_{ni}) to represent the sentiment orientation of the i^{th} phrase like: $(u_{ni}, v_{ni}) = \sum_{j=1}^{K_{ni}} w_j (u_x, v_x)$

Step three: sentiment orientation composition at sentence levels. Use (u_n, v_n) to stand for the emotional tendency of the $M_n{}^{th}$ phrase in the $n\ (n = 1,2,\cdots,N)^{th}$ sentence, then we can compose it with the use of arithmetic mean aggregation operator into:

$$(u_n, v_n) = \frac{1}{M_n} \bigoplus_{i=1}^{M_n} (u_{ni}, v_{ni}) \qquad (3)$$

Step four: sentiment orientation composition at textual levels 1. Sentiment tendency of sentences is expressed by (u, v). When composing, use the effect of transitional words as the weight of sentences, then use the weighted mean aggregation operator to synthetize it into:

$$(u_{pos}, v_{pos}) = \bigoplus_{i=1}^{M_n} \omega_n (u_n, v_n) \qquad (4)$$

in which, ω_n is the weight modifying transitional words in the n $(n = 1,2,\cdots,N)$ th sentence.

Step five: sentiment orientation composition at textual levels 2. The essence of the operation of intuitionistic fuzzy set is the evidence of aggregation supporting rather than opposing elements belonging to a set, so, we can't simply make classification according to the size of u and v. In order to categorize appropriately, we can firstly switch positions of features' membership and non-membership degree, then, repeat step two to four for the operation. The result will be (u_{neg}, v_{neg}).

Step six: classification. (i) If $(u_{pos}, v_{pos}) > (u_{neg}, v_{neg})$, then, text *x* belongs to pos type; (ii) if $(u_{pos}, v_{pos}) < (u_{neg}, v_{neg})$, text *x* belongs to neg type; (iii) otherwise, we can't determine its sentiment orientation classification.

IV. Experiment and Result Analysis

A. Text base and feature selection

Through the selection of test texts for the experiment, balanced corpora are provided without removing repetitive corpora [9], of them, pos and neg class each has 2000 reviews. The context is about hotel review. Hence, training corpora choose the same context. From CTRP, 2390 reviews are selected, 1241 reviews being pos class, 1081 reviews being neg class and the rest 68 reviews having no definite emotional inclination.

Feature words of sentiment orientation classification differ from those of general text classification, which are to be selected from words with emotional tendency. In the paper [10], the author chose 40 groups of words, some of which are commendatory and the others are derogatory. The work used emotional words and degree adverbs of moods as features. In [11], the distance between words and reference words selected manually was counted according to the distance between sememes and feature words were chosen with TFIDF formula. Here, appreciative terms and derogatory words were selected as feature words. In the experiment, we just chose words with obvious emotional colors and occurring frequently as feature ones, i.e. adjectives, verbs and modal particles. The first 150 words formed feature word list, corresponding to some of which, their intuitionistic fuzzy numbers are shown in Table III.

Table III Some feature words and their corresponding intuitionistic fuzzy numbers

Feature Words	Intuitionistic Fuzzy Numbers	Feature Words	Intuitionistic Fuzzy Numbers	Feature Words	Intuitionistic Fuzzy Numbers
Clean	(0.6889, 0)	Dirty	(0, 0.9265)	Disappointed	(0, 0.8293)
Warm	(0.9024, 0)	Good	(0.7594, 0)	Depressed	(0, 0.7586)
Cheap	(0.5341, 0)	Happy	(0.3824, 0)	Dad	(0, 0.9667)
Beautiful	(0.755, 0)	Ho-ho	(0.7778, 0)	Impartial	(0.4286, 0)

B. Results

We chose accuracy and recalling rate as evaluation indicators, which are listed in Table V as follows:

Table V. Experimental Results of Classification

Classification	Number of Text	Accuracy	Recalling Rate	Unclassifiable Texts
Pos	2000	89.96%	88.75%	53
Neg	2000	90.77%	90.1%	110

C. Discussions

From the above table II, we can discover the classification method based on intuitionistic fuzzy reasoning has higher accuracy and recalling rate, that's because it adopts fully the intuitionistic fuzzy theory to quantify the degree of membership, non-membership as well as hesitancy, supporting fairly well the classification decision through the fusion of uncertain information. Additionally, we made the following analysis:

1)Texts as indicated in section II exist in the test corpora. They should belong to neg class by intuitive judgment of text's semantics, but in the corpora, they are pos class, which, to some extent, leads to neg's better effect of classification than pos'. Take for example, contents of 1201 texts in pos class are: too general; internet speed as slow as a snail; what a high speed ! OMG! Add on April 2, 2008: no breakfast.

2)From table I, we can conclude that words having better discrimination are not necessarily adjectives with strong emotional colors, like, "ho-ho", whose membership degree to pos class is 0.7778; while "happy" is only 0.3824.

V. CONCLUSIONS

In the paper, it discussed the application of the degree of membership, non-membership and hesitancy of intuitionistic fuzzy set to quantitatively describe the information provided by features for the classification. Together with the use of information about the intuitionistic fuzzy information fusion theory to compose features, classification decision was eventually supported. Satisfactory accuracy and recalling rate were obtained through the experiment on sentiment orientation classification of webpage online reviews. The proposed method proved to be valid and practical. It is useful to acquire network information on the whole, understand the populace's emotions in real time, and provide decision support for hot issues or emergencies. Yet, as to how to improve the classification precision and how to use the proposed method to discover hidden information, such as how to provide decision support for the fuzzy integrated reviews about hotel management under such circumstances as in the experiment, it will be the next research field.

REFERENCES

[1] Wei Jiuchang, Zhao Dingtao. The crisis information communication model and its influence factor research . Information science, 2006, pp. 1782-1785.

[2] The workshop on new text Wikis and blogs and other dynamic text sources (EACL22006) http://www.sics.se/jussi/newtext/, 2006

[3] Xuefeng, Liu Qiuyun uncertainty reasoning in text classification. Journal of Jiangxi Normal University (NATURAL SCIENCE EDITION), 2007, pp.383-386

[4] Sheng Xiaowei, Jiang Minghu. Based on Rough reduction algorithm of Chinese text automatic classification system. Journal of Electronics & information technology, 2005, pp.1047-0.52.

[5] Jean-Pierre Barthélemy, Gentian Gusho. On the stability of hierarchical classification: Qualitative approaches . Mathematical and Computer Modelling, 2009, pp.329-332

[6] Zeshui Xu. Intuitionistic fuzzy aggregation operators. IEEE TRANSACTIONS ON FUZZY SYSTEMS, 2007, pp.1179-1187.

[7] Zhu Yanlan, min Jin, Zhou Yaqian. HowNet based word semantic orientation calculation . Journal of Chinese information processing, 2006, pp.14-20

[8] Li Chengwei, Peng Qin Ke, Xu Tao. Information inference based sentiment classification for online news comments. Journal of Chinese information processing, 2009, pp.75-79.

[9] Alistair Kennedy, Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifter. Computational Intelligence, 2006, pp.110-125.

[10] Shen Fengxian, Zhu Qiaoming. Based on the characteristics of dispositional webpage feature extraction method of.computer engineering and design, 2009, pp.3894-3896.

[11] Xu Linhong, Lin Hongfei, Yang Zhihao. Based on semantic understanding text tendency recognition mechanism.Journal of Chinese information processing, 2007, pp.96-100.