# A New Term Weighting Method by Introducing Class Information for Sentiment Classification of Textual Data

Long-Sheng Chen* and Chia-Wei Chang

***Abstract*—With the popularity of text based communication tools such as blogs, Plurk, Twitter, and so on, customers can easily express their opinions, reviews or comments about purchased products/services. These personal opinions, especially negative comments, might have a significant influence on other consumers' purchasing decisions. Therefore, how to detect users' sentiment from textual data to assist companies to carefully respond to customers' comments has become a crucial task. Recently, machine learning methods have been considered as one of solutions in sentiment classification. When applying machine learning approaches to classify sentiment, Term Frequency (TF), Term Presence (TP) and Term Frequency-Inverse Document Frequency (TF-IDF) usually have been employed to describe collected textual data. However, these traditional term weighting methods cannot have positive influence on improving classification performance. Therefore, this work proposes a new term weighting method called Categorical Difference Weights (CDW) by introducing class information. Besides, CDW will be integrated into Support Vector Machines (SVM). Finally, an actual case will be provided to illustrate the effectiveness of our proposed method. Compared with traditional term weighting methods, TF and TF-IDF, experimental results indicated that the proposed CDW method indeed can improve the classification performance.**

***Index Terms*—Term weighting, Sentiment classification, Text classification, Support Vector Machines.**

## I. Introduction

With the popularity of the Internet, the amount of comments in text based communication mechanisms such as blogs, Twitter, Plurk and so on is going to dramatically increase every day. Among this huge amount of comments, some opinions related to products or services might have a significant influence on consumers' purchasing decisions. For example, many people might learn how others' viewpoint of a product before buying or a company might improve the user satisfaction according to customer's opinions [1]. However, some comments or evaluations are usually negative and spread quickly, which could reduce consumers' purchase intentions and bring a great damage to enterprisers. Consequently, how to identify the sentiment of consumers effectively from a large number of online comments had become one of serious issues.

Recently, sentiment classification that classifies users' sentiment of text based communication tools into positive or negative has attracted lots of attention in web mining area [2]. Generally speaking, the objective of sentiment classification is to extract reviews from customers for certain products or services, and to identify type of sentiment of reviews [3]. Many works have been proposed to conduct textual sentiment classification. These studies could be divided into two categories [4]. The first group is to use machine learning techniques which build classifiers based on sentiment labeled textual comments and then identify the sentiment of new coming comments in blogs based on this constructed classifiers. The second group is to use semantic orientation approaches which classify terms into two classes (positive or negative), and then count the overall positive and negative scores in the documents to determine the sentiment of comments. In recent, machine learning techniques have been considered as one of effective solutions for sentiment classification. For example, Na et al. [5] used POS tags based negation phrases with SVM to improve the performance of classifying customers' comments. Ye et al. [3] used machine learning techniques of Naive Bayes, SVM and the character based N-gram model for sentiment classification regarding online travel destinations reviews. Tan and Zhang [4] compared four feature selection methods and five machine learning methods on sentiment classification of Chinese documents. Bai [6] proposed a heuristic search-enhanced Markov blanket model and used SVM as machine learning technique to predict consumer sentiments from online text.

In the process of using machine learning techniques, the textual data would be represented by feature vector. That is to calculate the weights of the terms in the documents and construct a term-document matrix (TDM). In the TDM, documents are represented by vectors which are expected to indicate as much information of the documents as possible [7]. In order to make the representation accurate and efficient, the term weighting method plays an important role in the process [7]. In related works of text classification, there are many term weight methods, such as term frequency (TF), inverse document frequency (IDF), term frequency-inverse document frequency (TF-IDF), and term presence (TP), etc. However, traditional term weighting methods cannot have positive influence on improving the performance of sentiment classification. They are calculated by the number

Manuscript received October 6, 2010; revised November 2, 2010. This work was supported in part by the National Science Council of Taiwan, R.O.C. (Grant No. NSC 98-2410-H-324-007-MY2).

*L.-S. Chen is with the Chaoyang University of Technology, Taichung 41349, Taiwan (phone: 886-4-23323000ext7752; fax: 886-4-23304902; e-mail: lschen@cyut.edu.tw).

C.-W. Chang is with the Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail: cwc0124@gmail.com).

of times a term occurs in a document or weather a term appears in a document. Therefore, this study will propose a new term weighting method by introducing class information while counting weight of a term in a document, which is called Categorical Difference Weights (CDW). Moreover, the most common and easiest feature selection method based feature frequency (FF) of unigrams have been employed to extract features and support vector machines (SVM) also employed to construct classifiers for identifying sentiments of text data. Finally, an actual case of online product reviews will be provided to illustrate the effectiveness of our proposed CDW method.

## II. Related works

### A. Feature selection

With increasing of the textual data in cyberspace, how to extract significant information from a huge amount of data have been become a serious problem. The objective of feature selection is to extract the important terms in the documents, and achieve the goal of dimension reduction. Feature frequency (FF) is the most common and easiest technique for selecting relevant terms in the documents. According to lots of published literatures [5], [8], feature frequency based unigrams have been obtained decent solutions. For example, in the experiments of Na et al. [5], using feature frequency based unigrams out-performed terms labeled with POS tags. Pang et al. [8] verified using only unigrams as features are better than bigrams, combinations of unigrams and bigrams, and POS tags. In this study, we calculated the times of each term occur in documents and selected the term whose frequency of appearance is higher.

### B. Term weighting method

Term weighting method aims to indicate the significant of a term in a document [9]. In sentiment classification, TF and TF-IDF are widely applied to count the weight of a term [10], [5], [11], [8]. TF represents the number of times a term occurs in a document, and TF-IDF is the combining of TF and IDF weights. IDF indicates the general importance of a term in overall documents. IDF and TF-IDF can be calculated as equations (1) and (2).

$$idf = \frac{The\ number\ of\ total\ documents}{The\ number\ of\ documents\ include\ a\ term} \quad (1)$$

$$tf = tf * idf \quad (2)$$

If a term's score of TF-IDF is high, it means this term occurs frequently and only appears in the part of overall documents. In this study, we compared our proposed CDW method with TF and TF-IDF weights.

### C. Support vector machine

SVM is a machine learning technique based risk minimization principle of statistical learning theory introduced by Vapnik [12], and it can deal with the problem of classification for multi-class or binary class. In the domain of sentiment classification, SVM aims to tackle the two-class problem by finding a hyperplane of maximal margin. Several studies [13], [10], [5], [11], [8], [4], [3] reported that SVM had a superior performance on sentiment classification.

## III. Proposed Methodology

Our CDW is inspired from Simeon and Hilderman's Categorical Proportional Difference (CPD) [15] which is originally proposed as a feature selection method on multi-class text classification. Then, O'Keefe and Koprinska [11] apply CPD to binary class sentiment classification. Therefore, we firstly introduce CPD that can be defined as equation (3).

CPD aims to count the positive document frequency (Positive DF) and negative document frequency (Negative DF) of a term separately, and then compute the proportional difference of a term in two classes.

$$CPD = \frac{\left|PositiveDF - NegativeDF\right|}{PositiveDF + NegativeDF} \quad (3)$$

From equation (3), we can know that the value of CPD will be located to the interval [0, 1]. If a term only occurs in positive document or negative document, we can find the value of CPD equal to one. Next this term will be viewed as significant for classification. In contrast, if a term occurs equally in positive and negative documents, we can get the value of CPD will be equal to zero. And this term will be considered as irrelevant. In practice, CPD can effectively extract the useful features by introducing class information. But, after implementing CPD method, we find that there is a drawback when counting the value of CPD to each term. Take Table 1 for example.

Table 1 Compare of CPD for three terms

| | Positive DF | Negative DF | CPD |
|---|---|---|---|
| Term A | 100 | 0 | 1 |
| Term B | 0 | 30 | 1 |
| Term C | 0 | 1 | 1 |

In this table, we can find that term A is more important than term B and term C. However, all of them have the same CPD score. Under such situation, we cannot know which one should be the most important. So, when we introduce CPD method to be a term weighting method. It cannot indicate the significance of a term in a document efficiently. In order to enhance CPD and solve the situation mentioned above, we modified CPD method which called Modified CPD (MCPD). Then based on MCPD, we propose a new term weighting method called Categorical Difference Weights (CDW).

To introduce CDW, we should know MCPD first. For implementing MCPD, a term's MCPD score can be defined as equations (4) ~ (6) by considering different situations. The considerations and equations of MCPD have described as bellow.

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Situation 1:

If a term's 'Positive DF' or 'Negative DF' is equal to zero, or 'Positive DF' is equal to 'Negative DF', the score of MCPD can be defined as equation (4).

$$MCPD = \left| PositiveDF - NegativeDF \right| \quad (4)$$

Situation 2:

If a term's 'Positive DF' is greater than 'Negative DF', the score of MCPD can be defined as equation (5).

$$MCPD = \frac{PositiveDF}{NegativeDF} \quad (5)$$

Situation 3:

If a term's 'Positive DF' is less than 'Negative DF', the score of MCPD can be defined as equation (6).

$$MCPD \frac{NegativeDF}{PositiveDF} \quad (6)$$

According to the definitions of MCPD, we can find that if a term occurs in a certain class frequently, the score of MCPD would be greater. In contrast, if a term occurs equally in each class, the score of MCPD would equal to zero. In addition, we testify our proposed MCPD method which can enhance CPD method. We take Table 2 as an example.

Table 2 Compare of CPD and MCPD

| | Positive DF | Negative DF | CPD | MCPD |
|---|---|---|---|---|
| Term A | 100 | 0 | 1 | 100 |
| Term B | 0 | 30 | 1 | 30 |
| Term C | 0 | 2 | 1 | 2 |
| Term D | 5 | 50 | 0.818 | 10 |

In table 2, we can easily find that our proposed method can indicate a term's importance in the documents efficiently. After counting MCPD score, we view a term's score of MCPD as weights in overall documents. Therefore, a term's CDW can be defined as follows:

$$CDW = \frac{MCPD}{PositiveDF + NegativeDF} \quad (7)$$

, where DF means the document frequency, that is the number of documents including this terms. After counting each term's CDW and constructing a TDM, we can find that if the difference of times a term occurs in two-class is greater; the weight of the term would be higher.

In order to illustrate the effectiveness of our method, we will compare our proposed method with traditional TF and TF-IDF methods, and then use a MP3 issues review data as our experiment source.

The implemental procedure of the experiment can be demonstrated as Figure1. It can be divided into four steps. They are described as follows.

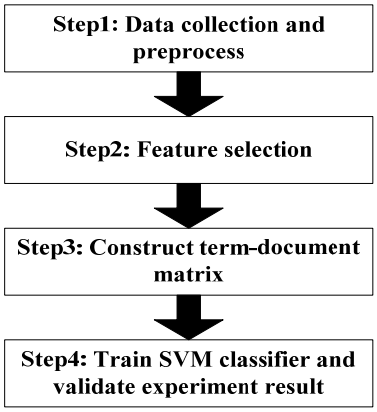


Fig. 1 The implemental procedure of the experiment

*Step 1*:

In this step, we use a MP3 issues review data as our experiment sources. In addition, we will remove some stop words and irrelevant words.

*Stept2*:

In step 2, we use feature frequency based on unigrams to select key words. Besides, we rank feature frequency and use its order to select key attributes. Six dimension sizes (1000, 700, 400, 200, 100, 50) will be considered.

*Step3*:

After selecting relevant terms, we calculated the weights based on selected feature space. Six term-document matrixes with six different dimension sizes would be constructed. Each TDM will be described in CDW, TF, and TF-IDF. In addition, a five-fold cross validation experiment has been employed in this study to generate training and test data.

*Step4*:

Finally, we use the training data to construct SVM classifier, and then input the test data to validate the built classifier. Besides, we compare our proposed CDW method with traditional TF and TF-IDF method.

## IV. Experiment Result

### A. The employed data

To conduct the research, an actual MP3 issues review data by retrieving corpus from Review Centre (http://www.reviewcentre.com/) have been utilized as experimental corpus. This corpus includes 200 positive reviews and 200 negative reviews. After data preprocessing, 1382 terms are left for further analysis. Besides, we use the LIBSVM which was developed by Chang and Lin [14] to build SVM classifier.

### B. The Results

First, we use the data which only implement data preprocess without employing feature selection method. After five-fold cross validation experiment, the results including average and standard deviation can be summarized in table 3.

Table 3 Results without implementing feature selection

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

method

| Weights / Corpus | TF | | TF-IDF | | CDW | |
|---|---|---|---|---|---|---|
| | Mean (%) | SD (%) | Mean (%) | SD (%) | Mean (%) | SD (%) |
| MP3 (1382) | 82.00 | 6.16 | 81.50 | 8.59 | 89.25 | 6.82 |

In table 3, we can find that our proposed CDW has the best performances (Mean: 89.25%, SD: 6.82%) compared to TF (Mean: 82.00%, SD: 6.16%) and TF-IDF (Mean: 81.50%, SD: 8.59%). We can say that CDW is better than TF and TF-IDF without implementing feature selection techniques.

Next, we wonder these weighting methods' performances if we introduce feature selection approaches. Therefore, we rank attributes by their feature frequency and select key attributes according to this ranks. Consequently, we select six feature sets including 1000, 700, 400, 200, 100, and 50. Tables 4 listed their results.

Table 4 Results of MP3 issues review for three term weighting methods with SVM

| Weights / Dimensions | TF | | TF-IDF | | CDW | |
|---|---|---|---|---|---|---|
| | Mean (%) | SD (%) | Mean (%) | SD (%) | Mean (%) | SD (%) |
| 1000 | 82.25 | 6.15 | 79.75 | 8.72 | 87.50 | 7.71 |
| 700 | 81.50 | 5.96 | 79.00 | 7.15 | 88.00 | 7.48 |
| 400 | 81.50 | 6.27 | 81.00 | 6.58 | 87.75 | 7.04 |
| 200 | 79.50 | 5.63 | 81.75 | 5.05 | 85.50 | 3.38 |
| 100 | 75.50 | 7.74 | 78.75 | 7.07 | 80.75 | 8.73 |
| 50 | 74.75 | 8.26 | 75.50 | 5.84 | 77.75 | 6.75 |

In tables 4, we can find that with the reduction of dimension, the performances of the three term weighting methods descend gradually. But, our proposed CDW method still has the best classification accuracy generally. Figure 2 also showed the same results.

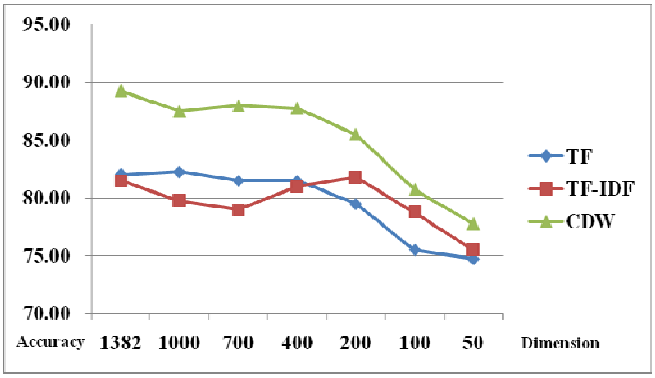


Fig. 2 Results of MP3 issues review when implementing feature selection with SVM

## V. Conclusions

In this study, we proposed a new term weighting method by introducing class information for sentiment classification of textual data. From experimental results, we can draw some conclusions. First, without considering feature reduction techniques, our CDW outperforms TF and TF-IDF. Second, after feature selection, although the performance of classification decreases gradually, our CDW method is still better than TF and TF-IDF methods generally. But, the performance gaps between our CDW and traditional weighting methods will be shortened. Therefore, our proposed method can measure the importance of a term in a document more effectively than TF and TF-IDF.

In the domain of text classification, there are many feature selection methods, such as information gain (IG), Chi-square, (CHI), etc. Integrating these methods into our CDW method might be potential direction of future works.

## References

[1] C. Zhang, W. Zuo, T. Peng, and F. He, "Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel," *The Third International Conference on Convergence and Hybrid Information Technology*, vol. 2, pp. 909-914, Nov. 2008.

[2] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *The 14th International Conference on World Wide Web*, pp. 342-351, May 2005.

[3] Q. Ye, Z. Zhang, and R. Law, "Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6527-6535, Apr. 2009.

[4] S. Tan and J. Zhang, "An Empirical Study of Sentiment Analysis for Chinese Documents," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622-2629, May 2008.

[5] J.C. Na, C. Khoo, and P.H.J. Wu, "Use of Negation Phrases in Automatic Sentiment Classification of Product Reviews," *Library Collections, Acquisitions, and Technical Services*, vol. 29, no. 2, pp. 180-191, Jun. 2005.

[6] X. Bai, "Predicting consumer sentiments from online text," Decision Support Systems, *doi:10.1016/j.dss.2010.08.024*, 2010.

[7] X. Tian and W. Tong (2010), "An Improvement to TF: Term Distribution Based Term Weight Algorithm," *The second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC)*, pp. 252-255, Apr. 2010.

[8] B. Pang, L. Lee, and S. Vaithyanathan (2002), "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *EMNLP*, pp.79-86, 2002.

[9] A. Aizawa, "An Information-theoretic Perspective of TF-IDF Measures," *Information Processing and Management*, vol. 39, no. 1, pp. 45-65, Jan. 2003.

[10] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," *The third AAAI International Conference on Weblogs and Social Media*, May 2009.

[11] T. O'Keefe and I. Koprinska, "Feature Selection and Weighting Methods in Sentiment Analysis," *The 14th Australasian Document Computing Symposium*, 2009.

[12] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[13] S. Li, C. Zong, and X. Wang, "Sentiment Classification through Combining Classifiers with Multiple Feature Sets," *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, pp. 135-140, Aug. 2007.

[14] C.C. Chang and C.J. Lin, "LIBSVM: a Library for Support Vector Machines," Software, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[15] M. Simeon and R. Hilderman, "Categorical Proportional Difference: A Feature Selection Method for Text Categorization," *The Australasian Data Mining Conference (Aus DM)*, pp. 201-208, Nov. 2008.

ISBN: 978-988-18210-3-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)