A New Supervised Term Ranking Method for Text Categorization

Musa Mammadov, John Yearwood, and Lei Zhao

Graduate School of information Technology and Mathematical Science,
University of Ballarat, Ballarat, VIC, 3350, Australia
{m.mammadov,j.yearwood,l.zhao}@ballarat.edu.au

Abstract. In text categorization, different supervised term weighting methods have been applied to improve classification performance by weighting terms with respect to different categories, for example, Information Gain, χ^2 statistic, and Odds Ratio. From the literature there are three term ranking methods to summarize term weights of different categories for multi-class text categorization. They are Summation, Average, and Maximum methods. In this paper we present a new term ranking method to summarize term weights, i.e. Maximum Gap. Using two different methods of information gain and χ^2 statistic, we setup controlled experiments for different term ranking methods. Reuter-21578 text corpus is used as the dataset. Two popular classification algorithms SVM and Boostexter are adopted to evaluate the performance of different term ranking methods. Experimental results show that the new term ranking method performs better.

1 Introduction

The task of text categorization is to assign unlabelled documents to predefined categories (topics or themes) according to their contents. Due to the growth in the volume of electronic documents, text categorization has been widely researched and applied in organizing as well as in finding information on the huge electronic resources.

Term weighting is an important issue for text categorization. In recent years, we have witnessed an increasing number of term weighting methods published. [1] classify the term weighting methods into two categories according to whether the method makes use of known information on the membership of training documents or not, namely, *supervised term weighting methods* and *unsupervised term weighting methods*. For example, *tf* and *tf-idf* [2] weighting methods belong to the unsupervised term weighting methods; information gain, χ^2 statistic, and odds ratio [3,4] are classified as supervised term weighting methods.

Although different approaches have been explored [4], not much attention has been paid towards specific class-oriented and local, context-dependent filters [5]. In particular, for multi-class text categorization, after being weighted by certain weighting methods, for example, information gain, term weights need to be summarized in to a single weight according to different categories.

J. Li (Ed.): AI 2010, LNAI 6464, pp. 102–111, 2010.

The literature indicates that there are only three methods to summarize term weights. The most common method is called the Summation method [3,6], which sums up all terms' weights with respect to all categories. We denote this method by $f_{sum}(t_i) = \sum_{k=1}^{c} f(t_i, c_k)$, where $f(t_i, c_k)$ denotes term t_i's weight with respect to category c_k. [4] employ two other methods. One is the average weight of term t_i with respect to all categories, denoted by $f_{avg}(t_i) = \sum_{k=1}^{c} P(c_k) f(t_i, c_k)$, where $P(c_k)$ is the proportion of documents belong to category c_k in the whole corpus. The other is the Maximum method $f_{max}(t_i) = \max_{k=1}^{c}\{f(t_i, c_k)\}$, which ranks terms according to their maximum weights across all categories. Among these three methods, [6] prefer the salient terms which are unique from one category to another, that is the Maximum approach. [3] also declare that the Maximum method outperformed both the Summation method and the Average method, but the experimental results are not given. Therefore, a question arises here, " Can we perform extensive experimental comparison of these methods, moreover, can we propose a better method than these existing ones?"

In our research, based on existing supervised term weighting methods, we propose a new term ranking method, the Maximum Gap. We illustrate by concrete example that this method can better distinguish those terms which can better differentiate one or more categories from the others than some existing ones, namely, the Summation, Maximum, and Average methods (see [3,6,4]). We conduct a series of comparative experiments on the Reuter-21578 text corpus. SVM and Boostexter are adopted as the learning algorithms. Average precision is used as the evaluation method. In our numerical experiments, Maximum Gap outperforms the other three term ranking methods.

This paper is organized as follows: In Section 2 we survey the existing term-category weighting methods and term ranking methods, then we propose a new term ranking method. In Section 3 we describe the detailed experimental settings. In Section 4 we report experimental results and discussion. We conclude in Section 5.

2 A New Feature Ranking Method

In this section, we review existing term weighting methods (information gain and χ^2 statistic) as well as term ranking methods (Summation, Maximum, and Average methods), introduce a new term ranking method for text categorization, and provide an example to demonstrate the effectiveness of the new method on Reuter-21578 text corpus.

2.1 Term-Category Weighting

Over the years, a number of methods have been developed to measure the discriminating power of various terms with respect to different categories, such as

information gain, odds ratio, and χ^2 statistic. In this research, we discuss information gain and χ^2 statistic, which have been shown to be effective for text categorization [4].

For term t_i and class c_k, the general functions of information gain and χ^2 statistic can be shown as:

$$IG(t_i, c_k) = P(t_i, c_k) \log \frac{P(t_i, c_k)}{P(t_i)P(c_k)} + P(\overline{t}_i, c_k) \log \frac{P(\overline{t}_i, c_k)}{P(\overline{t}_i)P(c_k)} \tag{1}$$

$$\chi^2(t_i, c_k) = \frac{n[P(t_i, c_k)P(\overline{t}_i, \overline{c}_k) - P(t_i, \overline{c}_k)P(\overline{t}_i, c_k)]^2}{P(t_i)P(\overline{t}_i)P(c_k)P(\overline{c}_k)} \tag{2}$$

where $P(t_i, c_k)$ denotes the probability a document is from category c_k when term t_i occurs at least once in it, $P(t_i, \overline{c}_k)$ denotes the probability a document is not from category c_k when term t_i occurs at least once in it, $P(\overline{t}_i, c_k)$ denotes the probability a document is from category c_k when term t_i does not occur in it, $P(\overline{t}_i, \overline{c}_k)$ denotes the probability a document is not from category c_k when term t_i does not occur in it, n denotes the number of documents.

Text categorization problems on multi-class datasets can be simplified into multiple independent binary classification problems. In each experiment, a chosen category c_k can be tagged as 1, and the other categories in the same corpus are combined together as 0. A contingency table (see Table 1) can be used to record the number of documents which contain term t_i and do not contain term t_i under category c_k and $\overline{c}_k$, and the sum of these four elements, n, is the number of documents of the dataset.

Table 1. The contingency table for category c_k and term t_i

	t_i	$\overline{t}_i$
Positive Category: c_k	a	b
Negative Category: $\overline{c}_k$	c	d

Notation:
a: Number of documents in class c_k that contain term t_i
b: Number of documents in class c_k that do not contain term t_i
c: Number of documents in class $\overline{c}_k$ that contain term t_i
d: Number of documents in class $\overline{c}_k$ that does not contain term t_i

[6] use these four elements in Table 1 to estimate the probabilities in formula (1) and (2). The functions of information gain and χ^2 are rewritten as:

$$IG(t_i, c_k) = \frac{a}{n} \log \frac{an}{(a+b)(a+c)} + \frac{c}{n} \log \frac{cn}{(c+b)(a+c)} \tag{3}$$

$$\chi^2(t_i, c_k) = \frac{n(ad-bc)}{(a+c)(b+d)(a+b)(c+d)} \tag{4}$$

2.2 Maximum Gap

The formulas (3) and (4) define weights for each term t_i according to different categories $c_k, k = 1, \cdots, c$. We denote these weights by $f(t_i, c_k)$. In this paper, two cases will be considered: $f_{ig}(t_i, c_k) = IG(t_i, c_k)$ and $f_{\chi^2}(t_i, c_k) = \chi^2(t_i, c_k)$. To rank all the terms, we need to define a weight for each term t_i with respect to all categories. As mentioned before we will investigate three different methods – Maximum, Summation, and Average methods defined by: (see [3,6,4])

$$f_{max}(t_i) = \max_{k=1}^{c}\{f(t_i, c_k)\} \tag{5}$$

$$f_{sum}(t_i) = \sum_{k=1}^{c} f(t_i, c_k) \tag{6}$$

$$f_{avg}(t_i) = \sum_{k=1}^{c} P(c_k) f(t_i, c_k) \tag{7}$$

In this section, we propose a new term ranking method that will be called Maximum Gap (MG). Unlike the above approaches, this method aims to distinguish, in terms of weights, those terms which can better differentiate one or more categories from the others.

First, we organize term i's weights $\{f(t_i, c_k)\}_{k=1}^{c}$ as follows:

$$f(t_i, c_{k_1}) \geq f(t_i, c_{k_2}) \geq \cdots f(t_i, c_{k_c})$$

then the Maximum Gap of term t_i is defined as

$$f_{mg}(t_i) = \max_{j=1}^{c-1}\{f(t_i, c_{kj}) - f(t_i, c_{k_j+1})\} \tag{8}$$

In the following example, we demonstrate why MG might be more efficient than the other three methods.

Example: From Reuter-21578 corpus, we select the top 30 terms selected by different term ranking methods. For both clarity and briefness, we only compare Maximum Gap and Maximum methods, where terms are weighted by information gain. The Maximum method is chosen because it is accepted that (see for example, [3]) this method is better compared to Summation and Average methods. Note that our experimental results in Section 4 also support this opinion.

Among the top 30 selected terms ranked by Maximum Gap and Maximum methods respectively, Table 2 lists 14 terms that are not selected by the opposite term ranking method (those 23 terms selected by both methods are omitted).

It can be seen that the terms exclusively selected by the Maximum Gap method are more closely related to the top 10 categories (see Table 2) than those terms that selected by Maximum method. For example, **us** (this refers to us or USA), **the**, **central**, and **note** selected by Maximum method are less related to the top 10 categories, while all terms selected by Maximum Gap

Table 2. Terms exclusively selected by Maximum Gap and Maximum term weighting methods out of top 30 terms. The top10 categories of Reuter-21578 are acq, corn, crude, earn, grain, interest, money-fx, ship, trade, and wheat.

	Maximum	Maximum Gap
1	us	surplu
2	the	petroleum
3	market	acquisit
4	loss	bui
5	export	tariff
6	central	yen
7	note	energi

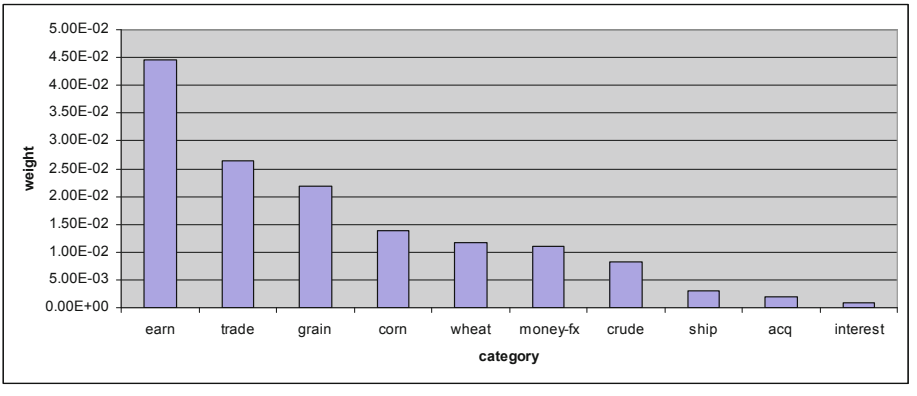

Fig. 1. Term-Category weights (calculated by information gain, Equation 3) of term "us" in Reuter-21578 top 10 categories (ordered by weights)

method are closely related to the top 10 categories. Here we should note that the selected terms have been stemmed by Porter Stemmer [7], for example, the original spelling of **surplu**, **bui**, and **energi** are **surplus**, **buy**, and **energy**. This concrete example can give us a direct explanation why Maximum Gap method might be better than Maximum method.

To provide insight into the performance of Maximum Gap, as an example we investigate the weights $f_{ig}(t_i, c_k)$ of term **us** and term **yen**[1], with respect to the top 10 categories. These two terms are selected respectively by Maximum method and Maximum Gap method. Fig. 1 and 2 lists the sorted weights of **us** and **yen** with respect to top 10 categories. Compared with term **us**, the

[1] In the data set, term **us** appears 6216 times representing the *United State (U.S.)* (5767 times), the *US Dollar* (171 times), and the word *us* (278 times) respectively. Term **yen** appears 629 times, which stands for Japanese currency only. It is clear that **yen** is a more predictable term, which is highly related to the category Money Foreign Exchange (money-fx), while term **us** appears frequently almost in all categories (See Figure 1, and Figure 2).

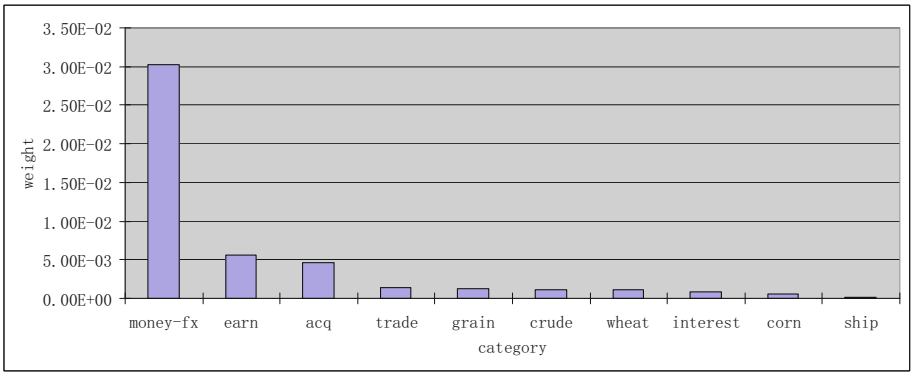

Fig. 2. Term-Category weights (calculated by information gain, Equation 4) of term "yen" in Reuter-21578 top 10 categories (ordered by weights)

Maximum Gap of term **yen** is bigger, but Maximum, Summation, and Average values are smaller. This is why term **us** is selected by Maximum, while term **yen** is selected by Maximum Gap.

3 Experiments

In this section, we describe the relevant details related to our experiments.

Dataset: In this controlled experiment, Reuter-21578 [8] is adopted as the benchmark dataset. In particular, the documents of the top 10 topics are extracted, from which 9393 related documents out of 21578 documents are extracted. Taking into account a large number of documents (9393) in the data set, we use 4-fold cross validation for the purpose of evaluation. Because documents are multilabel, we arrange these folds as follows. The first step, we consider all the combinations of multi-labeled classes and partition them based on the classes they belong to. The second, we fold each of the partitions, rather than the entire dataset, so that we could always keep the pattern for a particular classs combination from the testing set in the training set.

Learning Algorithms: In terms of learning algorithms, SVM and Boostexter are selected. For SVM, we use Chih-Jen Lin's LIBSVM (see [9]). Boostexter (see [10]) is based on the Boosting concept in Machine Learning. It has been proved as one of the most efficient classification algorithms and widely applied in many areas. Both SVM and Boostexter have shown competitive performance on text categorization [11,10].

Supervised Term Weighting Methods: In our experiment, we select two methods to weight the terms across different categories, namely information gain given by Equation (3) and the χ^2 statistic given by Equation (4). These methods have been shown effective and suitable for text categorization [4].

Performance Evaluation: In general, recall, accuracy or confusion matrix are used to evaluate the performance of the classification. These approaches are commonly used for binary or multiclass classification, where correct or not correct results can be evaluated without difficulty. According to multilabel classification problem, the ranking information of the predicted labels are also need to be considered. Average precision [12,10] is an evaluation method that are designed for multilabel classification problems, where the degree of accuracy can be measured by a single number that is more convinient for comparison purposes. Average precision is a performance measure previously used for evaluation of information retrieval (IR) systems [13]. In our experiments, we use a modified Average Precision (see [14]).

Given classifier (h, H) , predicted labels are denoted by $\mathcal{H}(x)$, actual labels are denoted by $\mathcal{Y}(x)$. Let $Y(x) = \{l \in \{1, \ldots, c\} : \mathcal{Y}_l(x) = 1\}$ be the set of actual labels of document x and $\mathcal{H}(x) = \{\mathcal{H}_1(x), \cdots, \mathcal{H}_c(x)\}$ be predicted labels. We denote by $\mathcal{T}(x)$ the set of all ordered labels $\tau = \{i_1, \ldots, i_c\}$ satisfying the condition

$$\mathcal{H}_{i_1}(x) \geq \ldots \geq \mathcal{H}_{i_c}(x);$$

where $i_k \in \{1, \ldots, c\}$ and $i_k \neq i_m$ if $k \neq m$.

In the case, when the numbers $\mathcal{H}_i(x), \quad i = 1, \cdots, c,$ are different, there is just one order τ satisfying this condition. But if there are labels having the same value then we can order the labels in different ways; that is, in this case the set $\mathcal{T}(x)$ contains more than one order.

Given order $\tau = \{\tau_1, \ldots, \tau_c\} \in \mathcal{T}(x),$ we define the rank for each label $l \in Y(x)$ as $rank_\tau(x; l) = k$, where the number k satisfies $\tau_k = l$. Then *Precision* is defined as:

$$P_\tau(x) = \frac{1}{|Y(x)|} \sum_{l \in Y(x)} \frac{|\{k \in Y(x) : \ rank_\tau(x; k) \leq rank_\tau(x; l)\}|}{rank_\tau(x; l)}.$$

Here, we use the notation $|S|$ for the cardinality of the set S. This measure has the following meaning. For instance, if all observed labels $Y(x)$ have occurred on the top of ordering τ then $P_\tau(x) = 1$. Clearly the number $P_\tau(x)$ depends on order τ. We define

$$P_{best}(x) = \max_{\tau \in \mathcal{T}(x)} P_\tau(x) \quad \text{and} \quad P_{worst}(x) = \min_{\tau \in \mathcal{T}(x)} P_\tau(x),$$

which are related to the "best" and "worst" ordering. Therefore, it is sensible to define the *Precision* as the midpoint of these two versions:

$$P(x) = \frac{P_{best}(x) + P_{worst}(x)}{2}.$$

Average Precision over all records $\mathcal{X}$ will be defined as:

$$P_{av} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} P(x).$$

4 Experimental Results

In this Section, we present the experimental results and compare the performance of the four term ranking methods discussed above.

To compare the performance of different term ranking methods, we use filter methods to select terms and test by SVM and Boostexter. We use S_n to denote the set of the top n terms ranked by certain term ranking methods. In our experiments, if a very small term subset is adopted, many documents of the corpus can not be included in training and test sets. Different term weighting methods can have different training and test subsets included, and they can not be compared appropriately. Actually, in our experiments, the selected terms can cover almost all the documents of our corpus if we have more than 20 terms selected. We only select 9 groups of sequential term subsets from Reuter-21578 corpus $S_{20} \subset S_{30} \subset S_{40} \subset S_{50} \subset S_{60} \subset S_{70} \subset S_{80} \subset S_{90} \subset S_{100}$.

The performance of four term ranking methods is shown in Table 3. Information gain and χ^2 statistic are used to weight terms across all categories respectively. Therefore, we have four different combinations of term weighting methods and text categorization algorithms shown in four columns in Table 3. The value in the table is the average performance among 9 groups of term subsets. The highest value according to different ranking methods is indicated by bold font.

In Table 4, 5, and 6, we make pairwise comparisons of different term ranking methods. The numbers in the second and the third columns of Table 4, 5, and 6 describe how many groups of term subsets show better evaluation performance.

Table 3. Terms are weighted by information gain and χ^2 statistic respectively. SVM and Boostexter are adopted as text categorization algorithms. Text categorization performances are evaluated by average precision. The average performance of 9 groups of term subsets are calculated according to four term ranking methods.

	IG-SVM	χ^2-SVM	IG-Boostexter	χ^2-Boostexter
mg	**93.751**	**93.264**	**95.345**	**95.121**
max	93.601	93.058	95.226	94.782
sum	93.128	92.395	94.923	94.493
avg	92.397	93.131	94.039	95.120

Table 4. Pairwise comparison of Maximum Gap and Maximum on 9 different term sets. 2 different term weighting methods (IG and χ^2 statistics) and 2 categorization algorithms (SVM and Boostexter) applied. In total, Maximum Gap perform better in 25 cases out of 36.

	Maximum Gap	Maximum
IG-SVM	**5**	4
IG-Boostexter	**7**	2
χ^2-SVM	**6**	3
χ^2-Boostexter	**7**	2
total	**25**	11

Table 5. Pairwise comparison of Maximum Gap and Summation on 9 different term sets. 2 different term weighting methods (IG and χ^2 statistics) and 2 categorization algorithms (SVM and Boostexter) applied. In total, Maximum Gap perform better in 34 cases out of 36.

	Maximum Gap	Summation
IG-SVM	**8**	1
IG-Boostexter	**9**	0
χ^2-SVM	**8**	1
χ^2-Boostexter	**9**	0
total	**34**	2

Table 6. Pairwise comparison of Maximum Gap and Average on 9 different term sets. 2 different term weighting methods (IG and χ^2 statistics) and 2 categorization algorithms (SVM and Boostexter) applied. In total, Maximum Gap perform better in 27 cases out of 36.

	Maximum Gap	Average
IG-SVM	**9**	0
IG-Boostexter	**5**	4
χ^2-SVM	**9**	0
χ^2-Boostexter	4	**5**
total	**27**	9

In all of our controlled experiments, Maximum Gap outperforms other term ranking methods in terms of the average performance of the 9 selected feature subsets. In the pairwise comparison with the existing methods, Maximum Gap method also performs very well. The only exception is the comparison with Average method by χ^2-Boostexter (see Table 6), but the difference between them is very close (4 to 5).

5 Conclusion

We present a new term ranking method for text categorization that is called Maximum Gap. This method is compared with three other similar methods: Maximum, Summation, and Average methods. Numerical experiments are carried out on the Reuter-21578 dataset. Experimental results show that the Maximum Gap outperforms other term ranking methods in selecting better terms for the text categorization task.

References

1. Lan, M., Tan, C.L., Low, H.-B.: Proposing a new term weighting scheme for text categorization. In: AAAI. AAAI Press, Menlo Park (2006)
2. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)

3. Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: SAC, pp. 784–788. ACM, New York (2003)
4. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Fisher, D.H. (ed.) ICML, pp. 412–420. Morgan Kaufmann, San Francisco (1997)
5. Duch, W., Duch, G.: Filter methods. In: Feature Extraction, Foundations and Applications, pp. 89–118. Physica Verlag, Springer (2004)
6. Liu, Y., Loh, H.T., Youcef-Toumi, K., Tor, S.B.: Handling of Imbalanced Data in Text Classification: Category-Based Term Weights. In: Kao, A., Poteet, S.R. (eds.) Natural Language Processing and Text Mining, p. 171 (2006)
7. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
8. Lewis, D.D.: Reuters-21578 text categorization test collection. Distribution 1.3 (2004)
9. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
10. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine Learning 39(2/3), 135–168 (2000)
11. Joachims, T., Nedellec, C., Rouveirol, C.: Text categorization with support vector machines: learning with many relevant. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
12. Li, T., Zhang, C., Zhu, S.: Empirical studies on multi-label classification. In: ICTAI, pp. 86–92. IEEE Computer Society, Los Alamitos (2006)
13. Salton, G.: Developments in automatic text retrieval. Science 253(5023), 974–980 (1991)
14. Mammadov, M.A., Rubinov, A.M., Yearwood, J.: The study of drug-reaction relationships using global optimization techniques. Optimization Methods and Software 22(1), 99–126 (2007)