A Data-drive Feature Selection Method in Text Categorization

Yan Xu [1,2]
1.Beijing Language And Culture University ,Beijing, China
2.Institute of Computing Technology,Chinese Academy of Sciences
xuy@blcu.edu.cn

***Abstract*—Text Categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. It has become a key technique for handling and organizing text data. One of the most important issues in TC is Feature Selection (FS). Many FS methods have been put forward and widely used in TC field, such as Information Gain (IG), Document Frequency thresholding (DF) and Mutual Information. Empirical studies show that some of these (e.g. IG, DF) produce better categorization performance than others (e.g. MI). A basic research question is why these FS methods cause different performance. Many existing works seek to answer this question based on empirical studies. In this paper, we present a formal study of FS in TC. We first define three desirable constraints that any reasonable FS function should satisfy, then check these constraints on some popular FS methods, including IG, DF, MI and two other methods. We find that IG satisfies the first two constraints, and that there are strong statistical correlations between DF and the first constraint, whilst MI does not satisfy any of the constraints. Experimental results indicate that the empirical performance of a FS function is tightly related to how well it satisfies these constraints and none of the investigated FS functions can satisfy all the three constraints at the same time. Finally we present a novel framework for developing FS functions which satisfy all the three constraints, and design several new FS functions using this framework. Experimental results on Reuters21578 and Newsgroup corpora show that our new FS function DFICF outperforms IG and DF when using either Micro- or Macro-averaged-measures.**

***Index Terms*—Feature selection, text categorization, Constraints**

I. INTRODUCTION

Text Categorization (TC) is the process of grouping texts into one or more predefined categories based on their content. Due to the increased availability of documents in digital form and the rapid growth of online information, text categorization has become a key technique for handling and organizing text data. One of the most important issues in TC is feature selection (FS), which is to select the features for TC from the available feature space.

Many FS methods have been proposed for TC, including Document Frequency thresholding (DF), Information Gain (IG) and Mutual Information (MI). Some comparative experiments have shown IG to be one of the most effective methods, while DF also performs very well, however MI has relatively poor performance [22][23]. Two other FS methods- CTD (Categorical Descriptor Term) [5] and SCIW (Strong Class Information Words) [11], have also been proposed and achieve promising performance. However, all these conclusions are based principally on empirical studies. The theoretical reasons why these methods behave in this way remains unanswered question.

Hui Fang et. al's work shed some light on how to make a formal study on Information Retrieval [2][3]. This work shows that intuitive retrieval heuristics can be formally defined as constraints on retrieval functions and that the empirical performance of a retrieval function is tightly related to how well it satisfies these constraints. A very interesting question is thus whether we can define some desirable constraints that any reasonable FS function should satisfy and develop new High-Performance FS functions that can satisfy all the desirable constraints.

In this paper, we give a formal study of FS in TC. We first formally define three desirable constraints that any reasonable FS function should satisfy. Then check these constraints on some existing feature selection methods, including IG, MI, DF, CTD and SCIW. We find that IG satisfies the first two constraints, and that there are strong statistical correlations between DF and the first constraint, while MI does not satisfy any constraint. Formal analysis and experiments indicate that the empirical performance of a FS function is tightly related to how well it satisfies the three constraints, and that none of the investigated FS functions can satisfy all the three constraints at the same time. Finally we put forward a framework for developing FS functions which satisfy all the three constrains and design two FS functions using this framework. Experimental results show that our new FS function DFICF outperforms IG and DF when using either Micro- or Macro-averaged-measure.

The remainder of this paper is organized as follows: Section 2 describes related work, Section 3 defines three basic desirable constraints, Section 4 checks the constraints on some existing FS methods, Section 5 gives a framework for developing good FS functions and develops several FS function based on the constraints, Section 6 presents the experiments and results, and finally our conclusions are summarized in Section 7.

© 2011 ACADEMY PUBLISHER
doi:10.4304/jsw.6.4.620-627

II. RELATED WORK

Many FS methods have been proposed for TC. This section reviews several popular and interesting FS methods, including DF, IG, MI, CTD and SCIW.

For simplicity of description, let us first introduce some notations. In this paper we use t to denote a term, $\bar{t}$ denotes absence of a term t, T={t, $\bar{t}$ } denotes all states for a term t, c or ci denotes a category, C=$\{c_i\}_{i=1}^{m}$ denotes the set of m categories, category distribution P(C) = (P(c1), P(c2),....., P(cm)) (P or p denotes probability). f(C, t) denotes a FS function, and gives the score of t with respect to C. CF(t) is the number of categories that a term t occurs and ICF(t) is inverse CF(t). N is the number of all the documents.

A. Document Frequency Thresholding

Document frequency is the number of documents in which a term occurs. Another form of DF-Inverse DF (IDF) is widely used in ranked retrieval such as the vector space model. DF thresholding (DF) only retains those terms with a higher DF value than a predefined threshold.

DF thresholding is the simplest technique for vocabulary reduction and it can easily scale to very large corpora due to its approximately linear computational complexity with respect to the number of training documents. Although DF is very simple, it can achieve good results as shown in [22] [23].

B. Information Gain

Information Gain (IG) is commonly used as a term goodness criterion in machine learning [14]. The IG of term t is defined as:

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

Given a training corpus, for each unique term its IG is computed and those terms whose IG is less than some predetermined threshold are removed. According to [22][23], IG can achieve very good results and is regarded as one of the most effective FS methods. Therefore, IG is widely used in TC tasks.

C. Mutual Information

Mutual Information (MI) is a criterion commonly used in statistical language modeling of word associations and related applications [18][22]. The MI between term t and category c is defined as:

$$MI(t,c) = \log_2 \frac{p\ (t \wedge c)}{p\ (t) \times p\ (c)}$$

In practice, two alternate MI functions are usually used:

$$MI_{avg}(t) = \sum_{i=1}^{m} p(c_i) MI(t,c_i)$$

,

$$MI_{\max}(t) = \max_{i=1}^{m}\{MI(t,c_i)\}$$

Previous work show that MI has relatively poor performance in TC [22][23].

D. Categorical Descriptor Term Method

The CTD (Categorical Descriptor Term) method utilizes the document frequency information in IDF as well as the category information in ICF (Inverse Category Frequency) [5]. CTD is defined as:

$$CTD(t,c_i) = TF(t,c_i) \times IDF(t,c_i) \times ICF(t)$$

$$\therefore CTD(t,c_i) \propto ICF(t)$$

here TF(t,ci) means the term frequency of t in all the documents of category ci.

Experiments show that CTD can perform comparatively well compared to other FS measures, especially on collections with highly overlapped topics.

E. Strong Class Information Words Method

SCIW (Strong Class Information Words) [11] is an approach to select the words with strong class information. For example, the word “football” usually appears in the class “sports”. SCIW is defined as:

$$SCIW(t) = \max\{p(c_i \mid t)\}$$

In the SCIW method, two thresholds are used, one is TS(0 ≤TS ≤ 1) which stands for the threshold of the SCIW value, the other is DS which stands for the threshold of the DF. Only the features which satisfy these two thresholds are selected.

Experiments show good precision performance for a linear classifier using SCIW.

$$\because \sum_{i=1}^{m} p(c_i \mid t) = 1$$

If only one category in which term t occurs, then $SCIW(t) = \max\{p(c_i \mid t)\} = 1$, it achieves its maximum value.

Thus generally speaking, the SCIW value tend to be dominated by the appearance probability of terms in rare category frequency, i.e. generally speaking, the more categories in which term t occurs, the less value SCIW(t) achieves, so there is strong correlations between SCIW(t) and ICF(t).

Some comparative experiments have shown that IG is one of the most effective methods, and DF also performs very well, while MI has relatively poor performance [22][23]. CTD and SCIW can also achieve good performance. However, these conclusions are mainly derived from empirical studies. Why these methods cause different results needs proper analysis and explanation. Let’s look back to the above conclusions, we observe that the excellent performance of DF and IG may indicate that common terms (the terms with high DF) are indeed informative, while the success of CTD and SCIW may indicate that ICF information is also informative. Starting from the observation, we next report a formal study of FS methods.

III. THREE CONSTRAINTS

A. Term-Category Dependence Constraints

The Term-Category Dependence Constraints (TCDCs) include two constraints. The first constraint is based on the following intuition: when the presence or absence of a

© 2011 ACADEMY PUBLISHER

term t has no relationship with category distribution P(C), f(C,t) should achieve its minimum value. Namely, if

$$(p(c_1 \mid t), p(c_2 \mid t),..., p(c_m \mid t)) = (p(c_1 \mid \bar{t}), p(c_2 \mid \bar{t}),..., p(c_m \mid \bar{t})) = (p(c_1), p(c_2),..., p(c_m))$$

then f(C, t) should achieve its minimum value

According to [19], equations (1) and (2) hold iff two discrete random variable C and T are independent:

$$(p(C = c_i; T = t) = P(C = c_i) \times p(T = t) \quad (0 \le i \le m) \quad (1)$$

$$(p(C = c_i; T = \bar{t}) = P(C = c_i) \times p(T = \bar{t}) \quad (0 \le i \le m) \quad (2)$$

$(p(c_1 \mid t), p(c_2 \mid t),..., p(c_m \mid t)) = (p(c_1 \mid \bar{t}), p(c_2 \mid \bar{t}),..., p(c_m \mid \bar{t}))$
$= (p(c_1), p(c_2),..., p(c_m))$ holds iff equations (1) and (2) hold

So, when C and T are regarded as two discrete random variables, the constraints of T with respect to C (these term-category dependence Constraints are denoted as TCDCs) are:

TCDC1: The value of f(C, t) should be the smallest, if and only if T and C are independent. (usually let f(C, t)=0)

Conclusion 1: when TCDC1 holds, i.e. $p(c_i) = p(c_i \mid t) = p(c_i \mid \bar{t})(0 \le i \le m)$ if and only if the value of f(C, t) is the smallest.

By contrast, when the value of C is completely determined by the value of T, f(C, t) should achieve the maximum value. This can be defined as:

TCDC2: The value of f(C, t) should be the largest, if and only if the value of C is completely determined by the value of T (though this case is infrequent in TC).

Conclusion 2: When TCDC2 holds, i.e. for m=2, C= $\{c_i\}_{i=1}^{2}$, $p(c_i \mid t) = 1, p(c_i \mid \bar{t}) = 0$ or $p(c_j \mid t) = 0, p(c_j \mid \bar{t}) = 1$ $(1 \le i,\ j \le 2, i \ne j)$ if and only if the value of C is completely determined by the value of T, f(C, t) is the largest.

B. Category Discrimination Constraint

TFIDF is the most common term weighting scheme used in the Vector Space Model [17]. First, it incorporates the word frequency in the document. Thus, the more a word appears in a document, the more significant it should be in the document. In addition, the measure of term specificity first proposed [7] and later became known as inverse document frequency (IDF). The intuition is that a term with high DF may not be a good discriminator, and should be given less weight.

Therefore, IDF means that a term which occurs in many documents is not a good document discriminator; similarly, ICF means that a term which occurs in many categories is not a good category discriminator. Furthermore, as described in Section 2, the success of CTD and SCIW may indicate that ICF information is also informative for TC.

Following the above consideration, the Category Discrimination Constraint (CDC) of term t is defined as:

CDC: Let m=|C| is the number of categories in the collection, CF(t) is the number of categories in which term t occurs, $ICF(t) = \log\frac{m}{CF(t)} + 1$, then f(C, t) ∝ ICF(t).

This constraint regulates the interaction between DF (or other factors) and ICF, and accurately describes the effect of using ICF in scoring. It ensures that, given a fixed number of DF (or other factors) of terms, we should favor a term that has a high ICF value.

IV. ANALYSIS OF SOME FEATURE SELECTION METHODS

In this section, we check these three constraints on some existing feature selection methods, such as IG, MI, DF, CDT and SCIW. Focusing in particular on analysis of IG, MI, DF. Our goal is to see how well each function satisfies the proposed constraints.

A. Analysis of Information Gain

According to [15], IG can be rewritten as

$$IG(C \mid T) = H(C) - H(C \mid T)$$

where $H(C)$ is the entropy of C, $H(C \mid T)$ is the average conditional entropy of T. Additional properties of $H(C \mid T)$ are[4]:

$0 \le H(C \mid T) \le H(C)$

$H(C \mid T) = 0$ if and only if the value of C is completely determined by the value of T

$H(C \mid T) = H(C)$ if and only if C and T are independent

Therefore

- $0 \le IG(C \mid T) \le H(C)$
- $IG(C \mid T) = 0$ if and only if C and *T* are independent
- $IG(C \mid T) = H(C)$ if and only if the value of C is completely determined by the value of T

So $IG(C \mid T)$ satisfies **TCDCs**. Obviously, it doesn't satisfy CDC.

Similar performance of IG and DF in term selection is observed in [22]. In this paper Yang and Pedersen found strong correlations between the DF and IG values of a term. Figure 1 plots the values of DF and IG given a term in the Reuters collection. Clearly there are indeed very strong correlations between the DF and IG values of a term, so IG favors common terms.

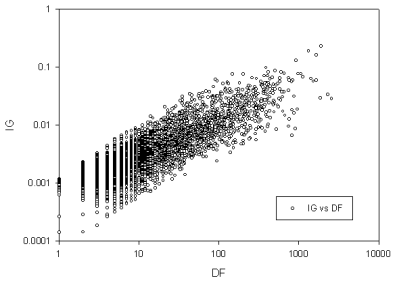


Fig. 1. Correlation between DF and IG values of words in Reuters

B. Analysis of Document Frequency Thresholding

DF thresholding does not use any category information, so it does not strictly satisfy TCDCs. However, empirical studies show that there are strong correlations between the DF and IG values (see Fig. 1).

© 2011 ACADEMY PUBLISHER

We give a statistical Interpretation of DF factor for FS in text categorization.

If term t never occurs in any document, or t occurs in each document, i.e. DF(t)=0 or DF(t)=N, here DF(t) is the Document Frequency of t.

Then equation (1) and (2) hold,

Hence, C and T are independent, therefore,

According to TCDC1, The value of f(C, t) should be the smallest. Therefore

$$\begin{cases} DF(t)=0 \Rightarrow f(C,t)=0 \\ DF(t)=N \Rightarrow f(C,t)=0 \end{cases}$$

If t appears infrequently, or t occurs too frequently, i.e. DF(t) near 0 or DF(t) near N

Then equation (1) and (2) nearly hold,

Hence, C and T are close to independent, therefore

$$\begin{cases} DF(t) \quad near \quad 0 \Rightarrow f(C,t) \approx 0 \\ DF(t) \quad near \quad N \Rightarrow f(C,t) \approx 0 \end{cases}$$

So, we suppose the curve for Correlation between f(C, t) which satisfies TCDCs and DF(t) should be given in figure 2 (the figure is similar to Luhn's[12]). We propose that:

- Let the line *L* in the figure represent the threshold of the rare terms, then the terms to its left are rare terms, they should be removed because of the presence or absence of these terms and *C* are nearly independent.
- Let the line *R* in the figure represent the threshold of the frequent terms, then the terms to its right are frequent terms, they should be removed because of the presence or absence of these terms and *C* are nearly independent. When we remove the stop words, these frequent terms have been removed.

So, DF factor is important for FS in text categorization.

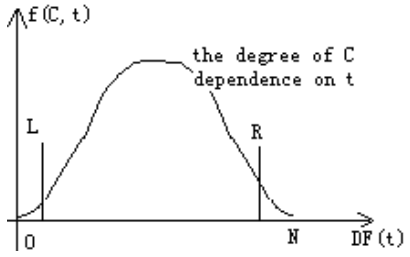


Fig. 2. Correlation between *f*(C, *t*) and DF(*t*), *f*(C, *t*) represents the degree of C dependence on term *t*.

Example 1: Let's define a function IND(t) as:

$$IND(t) = \sum_{i=1}^{m} | p(C=c_i; T=t) - p(C=c_i) \times p(T=t) | + \sum_{i=1}^{m} | p(C=c_i; T=\bar{t}) - p(C=c_i) \times p(T=\bar{t}) |$$

Then IND(t) is the smallest, if and only if T and C are independent, so IND(t) satisfies TCDC1.

If a term t only occurs once in the corpus, i.e. DF(t)=1, let term t only occurs in ci ,then

$$IND(t) = \sum_{i=1}^{m} | p(C=c_i; T=t) - p(C=c_i) \times p(T=t) | + | p(C=c_i; T=\bar{t}) - p(C=c_i) \times p(T=\bar{t}) | = \frac{4(N-|c_i|)}{N^2} < \frac{4}{N}$$

here |ci| is the number of documents which in category ci

∴ generally speaking, in a general way, there are a large number of documents in the corpus, the value of IND(t) above is very small, this implies that FS functions which satisfy TCDC1 do not favor rare terms, but rather favor common terms, that is, TCDC1 suggests that DF is a significant factor for feature selection in text categorization. This is an interpretation of the DF factor for FS in text categorization.

So there is strong correlation between the DF and **TCDC1**. But DF does not satisfy **CDC**.

C. Analysis of Mutual Information

According to [22][18], the MI of category *c* and term *t* is defined as:

$$MI(t,c) = \log \frac{p(t,c)}{p(t) \times p(c)} = \log \frac{p(t \wedge c)}{p(t) \times p(c)}$$

When *T* and *C* are independent, $p(t,c)=p(t)*p(c)$, thus $MI(t,c)=0$. But we can easily list a case where $MI(t,c)$ can be smaller than 0. Consider if *t* and *c* are complementary distribution, then $P(t,c)$ will be much less than $P(t)\,P(c)$, forcing $MI(t,c) < 0$, i.e. forcing $MI(t,c) < 0$, i.e. If $P(t,c)$ is near 0, $MI(t,c) <0$, $MI_{avg}(t) < 0$.

MI method does not satisfy the **TCDCs.**

Example 2. According to section 4.2, Let $DF(t)=N$ then *C* and *T* are independent, hence the value of $f(C, t)$ should be the smallest, but here $MI_{avg}(t) =0$, it does not achieve its minimum value, so MI method does not satisfy the **TCDC1.**

Figure 3 plots the values of DF and MI correspondingly(we select the average MI in this experiments). Clearly there are a lot of negative mutual information values. A lot of low document frequency terms have high *MI* values, and a lot of high document frequency terms have negative *MI* values, So, *MI* does not favor common terms, but rather rare terms, and does not satisfy the **TCDCs** which favors common terms. In addition, it does not satisfy **CDC**.

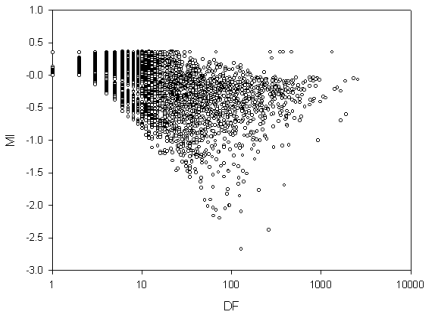


Fig. 3. Correlation between DF and MI values of words in Reuters

*D. Analysis of **SCIW** and **CTD** methods*

According to the analysis and discussion in section 2, the **CTD** satisfies **CDC** and **SCIW** tends to satisfy **CDC**. However, **CTD** does not satisfy constrain **TCDCs**, for **SCIW** method, two thresholds are used, one is **TS**($0 \leq TS \leq 1$) which stands for the threshold of the **SCIW** value, the other is DS which stands for the threshold of the DF, so there is strong correlation between the **SCIW** method and **TCDC1**.

© 2011 ACADEMY PUBLISHER

The above analysis indicates that none of these FS functions satisfy all the three constraints at the same time.

V. A FRAMEWORK FOR FEATURE SELECTION FUNCTIONS

In this section, we develop several FS function based on the constraints and give a framework for FS functions which satisfies all the three constraints.

A. Developing TCDC1-Based FS Methods

According section 4.2
IND(*t*) satisfies **TCDC1**, here

$$IND(t) = \sum_{i=1}^{m} | p(C = c_i; T = t) - p(C = c_i) \times p(T = t) | + | p(C = c_i; T = \bar{t}) - p(C = c_i) \times p(T = \bar{t}) |$$

similarly

$$IND_{k1,k2}(t) = k_1 \sum_{i=1}^{m} | p(C = c_i; T = t) - p(C = c_i) \times p(T = t) | + k_2 \sum_{i=1}^{m} | p(C = c_i; T = \bar{t}) - p(C = c_i) \times p(T = \bar{t}) |$$

Here, k_1 and k_2 are coefficient, $k_1 \geq 0$, $k_1 \geq 0$.
If $k_1 = k_2 = 1$ then $IND_{1,1}(t) = IND(t)$
$IND_{k1,k2}(t)$ also satisfies **TCDC1.**

B. Developing TCDCs-Based FS Methods

TCDCs (Term-Category Dependence Constraints**)** include two Constraints **TCDC1** and **TCDC2**. We call these functions which satisfy **TCDCs** as *Def*(C, *t*).
Developing steps:

1. for category distribution P(C) = ($P(c_1)$, $P(c_2)$,....., $P(c_m)$), design $g(C)$ as a function of category distribution.
2. calculate conditional category distribution g($C|T$): $g(C|T) = p(t)g(C|t) + p(\bar{t})g(C|\bar{t})$
3. calculate the contribution for category prediction by knowing the presence or absence of a term *t* in a document $g(C) - (p(t)g(C|t) + p(\bar{t})g(C|\bar{t}))$

Definition 1 *Term-Category Dependence measurement*: Let the category distribution function *g* satisfy the following:

- $0 \leq g(C|T) \leq g(C)$ (3)
- $g(C|T) = 0$ if and only if the value of C is completely determined by the value of T (4)
- $g(C|T) = g(C)$ if and only if C and *T* are independent (5)

Then the Term-Category Dependence measurement of *C* with respect to *t* is defined as:

$$Def(C,t) = g(C) - g(C|T) = g(C) - (p(t)g(C)|t) + p(\bar{t})g(C)|\bar{t}))$$

Conclusion 3: $Def(C,t)$ satisfies the Constraint **TCDCs**.
Proof (in appendix 1).

Obviously, *Def* is not unique and it depends on the function *g*. A different category distribution may correspond to different $g(C)$. Let's consider some examples of $g(C)$:

Example 3: $g(C) = H(C), H(C) = -\sum_{i=1}^{m} p(c_i) \log(p(c_i))$, here $H(C)$ is the entropy of *C*, obviously at this time, $Def1(C,t)$ = $IG(t)$, i.e. *IG* is a special case of *Def* functions.

Example 4: $g(C) = \sum_{1 \leq i < j \leq m} p(c_i) \times p(c_j)$, $Def2(C,t) = \sum_{1 \leq i < j \leq m} (p(c_i) \times p(c_j)) - (p(t) \sum_{1 \leq i < j \leq m} p(c_i|t)p(c_j|t) + p(\bar{t}) \sum_{1 \leq i < j \leq m} p(c_i|\bar{t})p(c_j|\bar{t}))$

we can prove that $Def2(C,t)$ satisfies **TCDCs** (the proof is omitted), here, $Def2(C,t)$ is a new FS function which satisfies **TCDCs.**

C. Framework for Feature Selection Functions

In the previous sections, we developed $IND_{k1,k2}(t)$ which satisfies the **TCDC1**, and *Defi*(C, *t*) which satisfies the **TCDCs**. In this section, we describe a framework for developing FS functions which satisfies all the three constraints. We believe that the first constraint suggests that *DF* is a significant factor, and the third constraint suggests that *ICF* is also an important factor, so we call the functions developed from this framework as *DFICF*:

$$DFICF(t) = Def(C,t) \times (\log \frac{m}{CF(t)} + 1)$$

The details of this framework for feature selection functions which satisfy all the three Constraints are given in Table 1.

Table 1. Framework for feature selection function

Step 1, for the classification category distribution P(C) = ($P(c_1)$, $P(c_2)$,....., $P(c_m)$), design a function *g*, call $g(C)$ as a function of category distribution. **Step 2**, calculate Conditional category distribution g($C|T$): $g(C|T) = p(t)g(C|t) + p(\bar{t})g(C|\bar{t})$, if ● $0 \leq g(C|T) \leq g(C)$ ● $g(C|T) = 0$ if and only if the value of C is completely determined by the value of T ● $g(C|T) = g(C)$ if and only if *C* and *T* are independent Then continue, else go to Step 1 **Step 3**, calculate the contribution for category prediction by knowing the presence or absence of a term t in a document $Def(C,t) = g(C) - (p(t)g(C|t) + p(\bar{t})g(C|\bar{t}))$ **Step 4**, $DFICF(t) = Def(C,t) \times (\log \frac{m}{CF(t)} + 1)$

Conclusion 4: Obviously, *DFICF*(*t*) is not unique, and it depends on the category distribution function *g*.
Conclusion 5: The TS function is a function of category distribution.

Conclusion 5 means different category distribution may correspond to different TS function. An interesting question for FS is how to design category distribution function $g(C)$ for a given category distribution, i.e. how to design optimal FS function for a given category distribution. This is one of our future works.

Continuing with the Examples 3 and 4, we respectively get Examples 5 and 6.

© 2011 ACADEMY PUBLISHER

Example 5: If $g(C)=H(C)$, $DFICF_1(t) = IG(t) \times (\log\frac{m}{CF(t)} + 1)$.

Example 6: If $g(C) = \sum_{1 \le i < j \le m} p(c_i) \times p(c_j)$,

$$DFICF_2(t) = Def_2(C,t) \times (\log\frac{m}{CF(t)} + 1) =$$

$$\sum_{1 \le i < j \le m} (p(c_i) \times p(c_j)) - (p(t) \sum_{1 \le i < j \le m} p(c_i \mid t) p(c_j \mid t) + p(\bar{t}) \sum_{1 \le i < j \le m} p(c_i \mid \bar{t}) p(c_j \mid \bar{t})) \times (\log\frac{m}{CF(t)} + 1)$$

D. Analysis and Summarization

The analyses of all the above FS functions are summarized in Table 2 ("Correl" means "correlation", "Number" means the number of Constraints which the FS method satisfies). The strong correlations between the DF and IG values of a term suggest that **TCDC1** favors common terms.

Table 2. Constraint analysis results

Constraint	**TCDC1**	**TCDC2**	**CDC**	Number
MI	No	No	No	0
DF	Correl	No	No	≈ 1
IND	Yes	No	No	1
$IND_{k1,k2}$	Yes	No	No	1
CTD	No	No	Yes	1
SCIW	Correl	No	Correl	≈ 2
IG(=Def1	Yes	Yes	No	2
Def2	Yes	Yes	No	2
$DFICF_1$	Yes	Yes	Yes	3
$DFICF_2$	Yes	Yes	Yes	3

In table 3, we summarize the connection between statistical information and constraints. For instance, **TCDC1** is connected to DF.

Table 3. The connection between statistical information and constraints

Constrain	**statistical information**
TCDC1	DF statistical information, category distribution statistical information
TCDC2	category distribution statistical information
CDC	ICF statistical information

VI. EXPERIMENT

A number of statistical classification and machine learning techniques have been applied to text categorization. Among them, two classifiers - the k-nearest-neighbor classifier (kNN) and the Naïve Bayes classifier are used in our experiments. We chose kNN because evaluations have shown that it outperforms nearly all the other systems[21], and we select Naïve Bayes because it is commonly studied in machine learning and an increasing number of evaluations of Naive Bayes methods on Reuters have been published[1,13]. We use the kNN and Naïve Bayes classifiers in Weka[20].

A. Performance Measurement

*F*1 was initially introduced by van Rijsbergen, and is a common measure in text categorization that combines both recall and precision [16]. The micro-averaged *F*1 (averaged over documents) has been widely used in cross-method comparisons while macro-averaged *F*1 (averaged over categories) has been used in some cases [9,10]. Generally speaking, the micro-averaged scores tend to be dominated by the performance on common categories, and the macro-averaged scores are more influenced by the performance on rare categories [21]. In order to examine the robustness of FS methods, both the macro-averaged *F*1 and the micro-averaged *F*1 are used in our experiments.

B. Data Collections

Two corpora are used for our experiments: Reuters-21578 collection[8] and the Newsgroups collection[6].

We use the ApteMod version of Reuters-21578, obtained by eliminating unlabelled and multi-labeled documents. We select the categories which have at least one document in the training set and the test set. This process resulted in 49 categories in both the training and test sets, we call this collection Reuters-21578(49), and then we select the categories which have at least five documents in the training set and the test set, this process resulted in 29 categories in both the training and test sets, we call this collection Reuters-21578(29).

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic. In our experiments, we select 10 categories out of all 20 by eliminating similar categories (e.g. we select one out of 5 categories of computer science). We use the "bydate" version of data whose training and testing data are split previously by the data provider. In the experiments, after stemming and stop word removal, there are 5769 documents as a training set and the 3837 documents as the test set. The total number of unique terms in the training set is 35996.

C. Results and Discussion

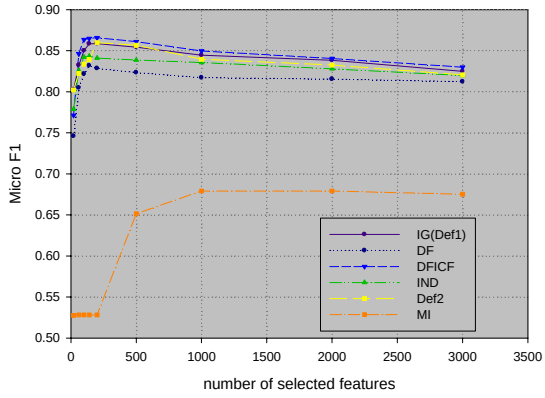
Figure 4 show the performance curves of kNN and Naïve Bayes on the Reuters-21578 collection after feature selection using DF, IND, IG(Def_1), Def_2, DFICF and MI. Here, DFICF refers to $DFICF_1$.

We can see from Table 2 that IG and Def2 satisfy **TCDC1** and **TCDC2**, IND only satisfies **TCDC1**, there are strong correlations between DF and **TCDC1**, MI does not satisfy any constraint, DFICF satisfy all three constraints. At the same time, we can see from Figure 4 that DFICF is the most effective method. IG and Def_2 perform similarly, IND and DF come next, MI has relatively poor performance. It can be concluded that the empirical performance of a FS function is tightly related to how well it satisfies the constraints.

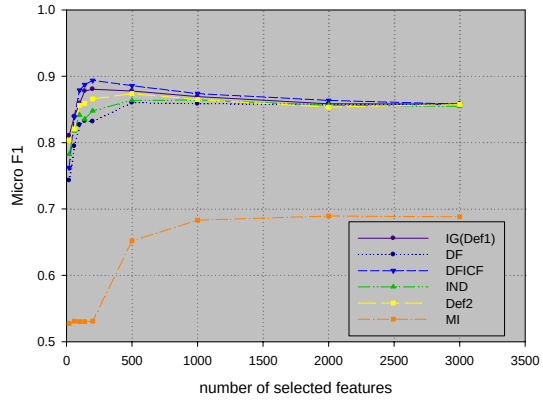
In order to see the difference of the performance clearly, Figures 5, 6 and 7 show the performance curves of kNN on the Reuters-21578 collection and Newsgroups using DF, IG and DFICF methods.

From the experiments we can see that the proposed DFICF is obviously better than IG and DF, especially using the Macro F1 performance measurement. From the experiments, we can reach these conclusions:

© 2011 ACADEMY PUBLISHER

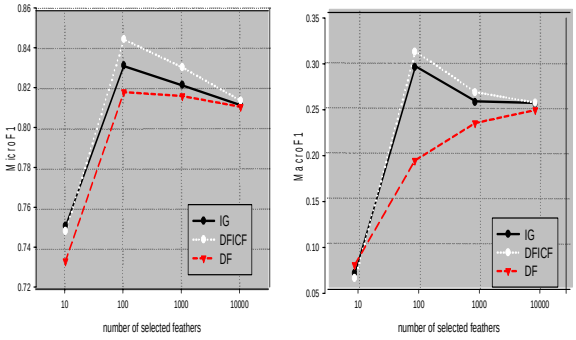


(a:KNN)



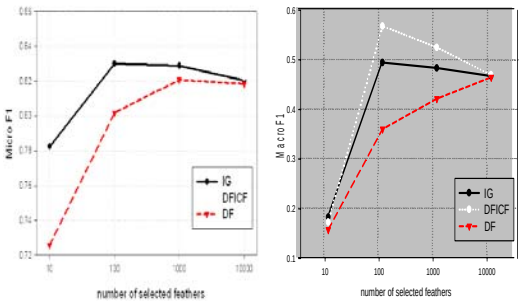
(b: NaiveBayes)

Fig. 4. Micro F1 of KNN or NaiveBayes vs. Number of selected features on Reauters-21578(29).



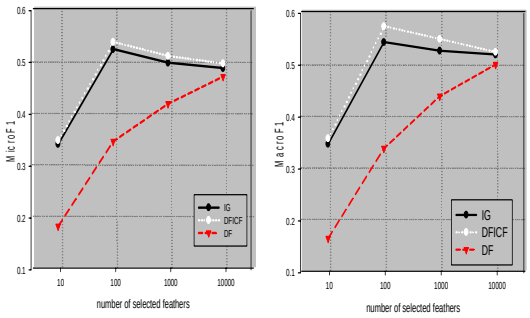
(a: Micro F1) (b: Macro F1)

Fig. 5. Micro or Macro F1 of KNN vs. Number of selected features on Reuters-21578(49)



(a: Micro F1) (b: Macro F1)

Fig. 6. Micro or Macro F1 of KNN vs. Number of selected features on Reuters-21578(29)



(a: Micro F1) (b: Macro F1)

Fig. 7. Micro or Macro F1 of KNN vs. Number of selected features on Newsgroups.

1. Figure 4 indicates that the empirical performance of a FS function is tightly related to how well it satisfies the three constraints.
2. Figures 5-7 indicate that DFICF is obviously better than IG and DF when using either Micro- or Macro-averaged-measures, especially using the Macro F1. As the macro-averaged scores are more influenced by the performance on rare categories, so ICF statistical information is useful for the corpus which has many rare categories.
3. Two corpora were used for our experiments: Reuters-21578 collection and the Newsgroups collection. The category distribution of Reauters-21578 is unbalanced (skewed): the most common category has a training-set frequency of 2877, but 33% of the categories have less than 10 instances [21]. It indicates that DFICF is useful for unbalanced corpus.
4. For the Newsgroups collection, we selected 10 categories out of all 20 by eliminating similar categories (e.g. we select one out of 5 categories of computer science), experimental result has given in Fig. 7. In addition, we select 5 computer science categories from the Newsgroups collection for another experiment, and found that DFICF and IG perform quite similarly (this experiment is omitted for due to lack of space in this paper), and indicate that DFICF is useful for the corpus which has great diversity in different categories.

VII. CONCLUSION

One of the most important issues in TC is FS. Many FS methods have been proposed for TC and many empirical study of FS method in TC. In this paper, we present a formal study of FS in TC and present a novel framework for developing FS functions which should satisfy three basic constraints and derive some new FS functions using this framework. Experimental results show that our new FS function DFICF outperforms IG and DF when using either Micro- or Macro-averaged-measures, especially using the Macro F1. Our work suggests that:

- DF information, ICF information, category distribution information are important statistical information that can lead good categorization performance. Especially, ICF statistical information is

© 2011 ACADEMY PUBLISHER

useful for the corpus which has many rare categories or has great diversity in different categories.

- The TS function should be a function of category distribution, it means different category distribution may correspond to different TS functions.

Therefore, our future research directions are:

- We need to extend our experiments to very large corpus tasks such as RCV1. Though we believe our method can scale to large-scale situation, we need to demonstrate this on large corpora.
- Since our constraints do not cover all the desirable properties, it would be interesting to explore additional necessary constraints in order to find other "necessary" statistical information which FS functions should use.
- Give a category distribution, we need to find a method to design category distribution functions $g(C)$ in order to get optimal FS function on the given category distribution.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Fundamental Research Project of China (60873166), the Ministry of Education of the People's Republic of China (109028) and the Education science Foundation of Beijing (AHA0911).

REFERENCES

[1] L. Douglas Baker and Andrew K. Mccallum. Distributional clustering of words for text categorization. In Proceedings of the.21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR' 98), pages 96-103

[2] Hui Fang, ChengXiang Zhai: An exploration of axiomatic approaches to information retrieval. SIGIR 2005: 480-487

[3] Hui Fang, Tao Tao, ChengXiang Zhai: A formal study of information retrieval heuristics. SIGIR 2004: 49-56

[4] R. M. Gray, Entropy and Information Theory, Springer-Verlag, 1990

[5] Bong,Chih How,Narayanan K. 2004. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence(WI'04).

[6] Lang K. Newsweeder: Learning to Filter Netnews. International Conference on Machine Learning (ICML). San Francisco:Morgan Kanfman Publishers,1995. 51-60.

[7] Sparck Jones, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, Vol. 28, pp. 11–21

[8] David D. Lewis,Reuters-21578 test collection. Reuters21578 http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[9] David D. Lewis, Li F, Rose T, Yang Y.RCV1:A new benchmark collection for text categorization research.Journal of Machine Learning Research,2004,5(3):361-397.

[10] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. (SIGIR'96), pp. 298-306, 1996

[11] Shoushan Li, Chenqing Zong. A New Approach to Feature Selection for Text, Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05.

[12] H.P. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, pages 159–165, 1958.

[13] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[14] T. Mitchell. Machine Learning. McCraw Hill, 1996

[15] Andrew Moore. Statistical Data Mining Tutorials. http://www.autonlab.org/tutorials/

[16] C.J. van Rijsbergen. Information Retrieval. Butterworths, London, 1979.

[17] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Commun. ACM, 18(11):613–620, 1975.

[18] F. Sebastiani, Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1-47. 2002.

[19] S.R.S.Varadhan. Probability Theory.Courant Institute of Mathematical Sciences. New York University. August 31, 2000

[20] Weka. http://www.cs.waikato.ac.nz/ml/weka/

[21] Y. Yang and X. Liu. A Re-examination of Text Categorization Methods. (SIGIR'99), pp. 42-49, 1999

[22] Y. Yang, Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, pp. 412-420

[23] Stewart M.Yang, Xiao-Bin Wu, Zhi-Hong Deng, Ming Zhang, Dong-Qing Yang. 2002 Modification of Feature Selection Methods Using Relative Term Frequency. Proceedings of ICMLC-2002, pp. 1432-1436

APPENDIX1

Conclusion 1: Term-Category Dependence measurement $def(C,t)$ satisfy the Constraint **TCDCs**.

Proof: If T and C are independent

Then $g(C|T) = g(C)$, equation (5) holds

$def(C,t) = g(C) - g(C|T) = 0$

$\because 0 \le g(C|T) \le g(C)$

$\therefore Def(t)$ is the smallest

vice versa

$\therefore$ $Def(t)$ satisfies the Constraint **TCDC1**

if the value of C is completely determined by the value of T

equation (4) is holds

$\therefore g(C|T) = 0$

$def(C,t) = g(C) - g(C|T) = g(C)$

$\therefore Def(t)$ is the largest

vice versa

$\therefore$ $Def(t)$ satisfies the Constraint **TCDC2**

$\therefore$ $Def(t)$ satisfies the Constraint **TCDCs** □

Yan Xu received the M.S. degree in computer science (1996) and the Ph.D. degree in computer science(2004) from Beijing University of Aeronautics & Astronautics, Beijing, China. She is an associate professor in Institute of Computing Technology, Chinese Academy of Sciences now. Her research interests include date mining, information retrieval.

© 2011 ACADEMY PUBLISHER