Akiko Aizawa

The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures

Proceedings of the 23rd Annual International ACM SIGIR Conference
on Research and Development in Information Retrieval (SIGIR 2000)
pages 104-111
2000

The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures

Akiko AIZAWA
National Institute of Informatics
2-1-2 Hitotsubashi Chiyoda-ku, Tokyo 101-8430, Japan
E-Mail: akiko@nii.ac.jp

Abstract

The *feature quantity*, a quantitative representation of specificity introduced in this paper, is based on an information theoretic perspective of co-occurrence events between terms and documents. Mathematically, the feature quantity is defined as a product of probability and information, and maintains a good correspondence with the $tfidf$-like measures popularly used in today's IR systems. In this paper, we present a formal description of the feature quantity, as well as some illustrative examples of applying such a quantity to different types of information retrieval tasks: representative term selection and text categorization.

1 Introduction

This paper presents the mathematical definition and applications of the feature quantity, a measure of specificity of terms or documents in a given document set. To introduce the basic idea, we first revisit the classical, but nevertheless important question of *'what is the mathematical implication of tfidf?'* in an information theoretic framework.

First of all, it is assumed that a document is given as an unordered set of terms. Let $D = \{d_1 \cdots d_N\}$ be a set of documents and $W = \{w_1 \cdots w_M\}$ be a set of distinct terms contained in D. The parameters N and M are the total numbers of documents and terms, respectively. In our adaptation of a probabilistic view, we also use the notion of d_j for an event of selecting a document from D. Similarly, w_i is used for an event of selecting a term from W. Now, let $\mathcal{D}$ and $\mathcal{W}$ be random variables defined over the events $\{d_1 \cdots d_N\}$ and $\{w_1 \cdots w_M\}$, respectively. Our objective here is to calculate the expected mutual information between $\mathcal{D}$ and $\mathcal{W}$ (Figure 1).

Assuming all the documents are equally likely at the initial stage, $P(d_j) = 1/N$ for all $d_j \in D$. Then, the amount of information calculated for each document is identically given by $-log(1/N)$. It follows that the self entropy of random variable $\mathcal{D}$, which is defined as the expected amount of information, is:

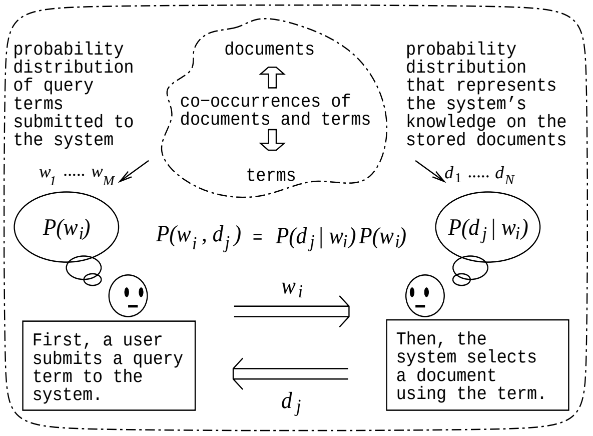

probability distribution of query terms submitted to the system
documents
co-occurrences of documents and terms
terms
probability distribution that represents the system's knowledge on the stored documents
w_1 ..... w_M
d_1 ..... d_N
$P(w_i)$
$P(w_i, d_j) = P(d_j | w_i)P(w_i)$
$P(d_j | w_i)$
w_i
d_j
First, a user submits a query term to the system.
Then, the system selects a document using the term.

Figure 1: An illustrative situation assumed in the calculation of the expected mutual information.

$$\mathcal{H}(\mathcal{D}) = -\sum_{d_j \in D} P(d_j) log P(d_j)$$

$$= -N \cdot \frac{1}{N} log \frac{1}{N} = -log \frac{1}{N}. \quad (1)$$

Next, consider a situation where a subset of specified documents that contain w_i ($\in W$) are known. Let N_i be the number of documents in the subset. Assuming that the N_i documents are equally likely, the amount of information calculated for each document in the subset is $-log(1/N_i)$. In this case, the self entropy of $\mathcal{D}$ given w_i becomes:

$$\mathcal{H}(\mathcal{D}|w_i) = -\sum_{d_j \in D} P(d_j|w_i) log P(d_j|w_i)$$

$$= -N_i \cdot \frac{1}{N_i} log \frac{1}{N_i} = -log \frac{1}{N_i}. \quad (2)$$

Since documents without w_i occur with probability zero, there is no contribution from these documents, i.e., the factor $(N - N_i)$ does not appear in the above equation.

Now, let us assume that a term w_i is randomly picked from the whole document set. Denoting the frequency of w_i within d_j as f_{ij}, the frequency of w_i in the whole document set as f_{w_i}, and the total frequency of all terms appearing in the whole document set as F, the probability that w_i is selected is $\sum_j \frac{f_{ij}}{F} = \frac{f_{w_i}}{F}$. Then, the expected information gain of the event, also referred to as the posterior entropy or the expected mutual information, is calculated as:

$$\mathcal{I}(\mathcal{D}; \mathcal{W}) = H(\mathcal{D}) - H(\mathcal{D}|\mathcal{W})$$

$$= \sum_{w_i \in W} P(w_i) \left(H(\mathcal{D}) - H(\mathcal{D}|w_i)\right)$$

† This manuscript appeared in Proc. of ACM SIGIR 2000, p.104-111.

$$
\begin{aligned}
&= \sum_{w_i \in W} \frac{f_{w_i}}{F} \left(-log\frac{1}{N} + log\frac{1}{N_i} \right) \\
&= \sum_{w_i \in W} \frac{f_{w_i}}{F}\, log\frac{N}{N_i} \\
&= \sum_{w_i \in W} \sum_{d_j \in D} \frac{f_{ij}}{F}\, log\frac{N}{N_i}. \qquad (3)
\end{aligned}
$$

Equation (3) equals the sum of the product of the *term frequency* (*tf*), either in the form of f_{ij} or f_{w_i}, and the *inverse document frequency* (*idf*) divided by a constant factor F. Hence, we conclude that from an information theoretic point of view, *tfidf* can be interpreted as the quantity needed for the calculation of the expected mutual information given by Eq. (3). When *tf* refers to f_{ij}, the *tfidf* values represent weights of terms within each document and are summed up for all the combination of terms and documents. When *tf* refers to f_{w_i}, the *tfidf* values represent the significance of corresponding terms in a whole document set and are summed up for all the words.

We should note here that in the derivation of Eq. (3), the condition $P(d_j) = \sum_{W(d_j)} \frac{f_{w_i}}{F} \cdot \frac{1}{N_i} = \frac{1}{N}$ is implicitly assumed for consistency, where $W(d_j)$ is the set of distinct terms contained in d_j. In our view, the specific assumption itself represents the heuristic that *tfidf* employs. Then, is there a possibility of extending the definition of *tfidf* into a more general form by applying the same information theoretic view?

Bearing this question in mind, the principal idea of this paper is that given a component of textual data such as a document or a term, the significance of the component is expressed as a product of the probability that it occurs and the amount of information it represents, i.e.,

(feature) = (probability) × (information).

Although conventional information theory does not explicitly deal with such a quantity (but uses the one in the calculation of entropy since entropy is generally defined as the expected amount of information), we have postulated that what the current popularity of the *tfidf* measure tells us is the usefulness of such a quantity as a measure of significance.

Another important implication of the above formulation is that the two probability distributions, $P(w_i)$ and $P(d_j|w_i)$ as shown in Figure 1, can be determined independently. In the figure, $P(w_i)$ represents the probability distribution of the query terms submitted to the system, while $P(d_j|w_i)$ is the conditional probability distribution of documents, given the query term. In other words, $P(w_i)$ serves as a model of the user, and $P(d_j|w_i)$ as a model of the retrieved documents. Such a formulation not only is closely connected to the previous theoretical development such as [1] [4] but also allows us to extend the classical definition of *tfidf* in more flexible ways, including the nonlinear scaling of term frequency, which is commonly practised in today's IR systems. Also, by adopting different term distribution models for the same document set, we can successfully connect the vector-space oriented view of the original *tfidf* to the probabilistic ones, as is shown in our text categorization application.

This paper reports some of the preliminary results of our attempt to expand such ideas. The subsequent sections are organized as follows. Section 2 presents the mathematical definition of the feature quantity. Section 3 deals with the problem of selecting representative terms where the feature quantity is used as a measure for the specificity of a term. Section 4 examines the text categorization problem where the feature quantity is used to identify the category best characterized by a given set of terms. Section 5 is the conclusion.

2 Mathematical Formulation

2.1 Notations and Basic Formulae of Information Theory

As before, let D and W be a set of documents and of terms, respectively, and $\mathcal{D}$ and $\mathcal{W}$ be random variables corresponding to D and W. Assume a joint probability distribution $P(w_i, d_j)$ is given for $d_j \in D$ and $w_i \in W$. Here, $P(w_i, d_j)$ provides a fundamental view of the problem. Naturally, a number of strategies are available to determine $P(w_i, d_j)$. For example, $P(w_i, d_j)$ can be determined directly from occurrences of terms in documents. In this case, standard techniques in probabilistic language modelling can be applied, including the simplest way of assigning probabilities proportional to the observed occurrences, or more computationally intensive ways, such as frequency discounting of n-gram statistics or the maximum entropy method. It is also possible to choose other distributions for $P(w_i)$ while still estimating $P(d_j|w_i)$ using the standard techniques. In this case, $P(w_i, d_j)$ is uniquely determined by the general probability formula $P(w_i, d_j) = P(d_j|w_i)P(w_i)$.

Given $P(w_i, d_j)$, it immediately follows that

$$P(w_i) = \sum_{d_j \in D} P(w_i, d_j), \qquad (4)$$

and

$$P(d_j) = \sum_{w_i \in W} P(w_i, d_j). \qquad (5)$$

By general definition of information theory, mutual information between w_i and d_j is given by

$$\mathcal{M}(w_i, d_j) = log\frac{P(w_i, d_j)}{P(w_i)P(d_j)}. \qquad (6)$$

The expected mutual information between $\mathcal{D}$ and $\mathcal{W}$ is:

$$
\begin{aligned}
\mathcal{I}(\mathcal{D};\mathcal{W}) &= \sum_{w_i \in W} \sum_{d_j \in D} P(w_i, d_j)\mathcal{M}(w_i, d_j) \qquad (7) \\
&= \sum_{w_i \in W} \sum_{d_j \in D} P(w_i, d_j)\, log\frac{P(w_i, d_j)}{P(w_i)P(d_j)}.
\end{aligned}
$$

$\mathcal{I}(\mathcal{D};\mathcal{W})$ can be viewed as the entropy of the co-occurrences of documents and terms. Note that by definition, $\mathcal{I}(\mathcal{D};\mathcal{W}) = \mathcal{I}(\mathcal{W};\mathcal{D})$ and Eq. (7) maintain duality regarding documents and terms.

The information increase of $\mathcal{D}$ after the event of observing w_i can be expressed using Kullback-Leibler information, which is a measure of the difference between two probability distributions. Kullback-Leibler information between $P(\mathcal{D}|w_i)$ and $P(\mathcal{D})$ is calculated as:

$$\mathcal{K}(P(\mathcal{D}|w_i), P(\mathcal{D})) = \sum_{d_j \in D} P(d_j|w_i) log\frac{P(d_j|w_i)}{P(d_j)}. \qquad (8)$$

Similarly, the information increase of $\mathcal{W}$ after the event of observing d_j is given by Kullback-Leibler information between $P(\mathcal{W}|d_j)$ and $P(\mathcal{W})$:

$$\mathcal{K}(P(\mathcal{W}|d_j), P(\mathcal{W})) = \sum_{w_i \in W} P(w_i|d_j) log \frac{P(w_i|d_j)}{P(w_i)}. \quad (9)$$

Applying $P(w_i, d_j) = P(d_j|w_i)P(w_i) = P(w_i|d_j)P(d_j)$ to Eqs. (7), (8) and (9), it is straightforward that the following relationships hold between the expected mutual information and Kullback-Leibler information:

$$\begin{aligned}\mathcal{I}(\mathcal{D}; \mathcal{W}) &= \sum_{d_i \in W} P(w_i)\mathcal{K}(P(\mathcal{D}|w_i), P(\mathcal{D})) \\ &= \sum_{d_j \in D} P(d_j)\mathcal{K}(P(\mathcal{W}|d_j), P(\mathcal{W})). \quad (10)\end{aligned}$$

2.2 Quantitative Representation of Features

Quantitative representation of the *feature* as formulated in this section is defined as the contribution of a specific co-occurrence event to the overall entropy calculation given by Eq.(7). The feature quantity of the occurrence of w_i and d_j is defined as:

$$\mathcal{F}(w_i, d_j) = P(w_i, d_j)\mathcal{M}(w_i, d_j). \quad (11)$$

Similarly, the feature quantity of the occurrence of w_i is defined as:

$$\mathcal{F}(w_i; \mathcal{D}) = P(w_i)\mathcal{K}(P(\mathcal{D}|w_i), P(\mathcal{D})), \quad (12)$$

and the feature quantity of the occurrence of d_j as:

$$\mathcal{F}(d_j; \mathcal{W}) = P(d_j)\mathcal{K}(P(\mathcal{W}|d_j), P(\mathcal{W})). \quad (13)$$

In all cases, the feature quantity is expressed as a product of probability and information, the latter being either mutual information, in Eq. (11), or Kullback-Leibler information, in Eqs. (12) and (13).

Equation (12) can further be rewritten as:

$$\begin{aligned}\mathcal{F}(w_i; \mathcal{D}) &= \sum_{d_j \in D} P(w_i)P(d_j|w_i)\ log \frac{P(d_j|w_i)}{P(d_j)} \\ &= \sum_{d_j \in D} \mathcal{F}(w_i, d_j), \quad (14)\end{aligned}$$

and Eq. (13) as:

$$\begin{aligned}\mathcal{F}(d_j; \mathcal{W}) &= \sum_{w_i \in W} P(w_i|d_j)P(d_j)\ log \frac{P(w_i|d_j)}{P(w_i)} \\ &= \sum_{w_i \in W} \mathcal{F}(w_i, d_j). \quad (15)\end{aligned}$$

Then, it follows that the entropy of all co-occurrences is simply expressed as the summation of feature quantity values of each case:

$$\begin{aligned}\mathcal{I}(\mathcal{D}; \mathcal{W}) &= \sum_{w_i \in W} \sum_{d_j \in D} \mathcal{F}(w_i, d_j) \\ &= \sum_{w_i \in W} \mathcal{F}(w_i; \mathcal{D}) \\ &= \sum_{d_j \in D} \mathcal{F}(d_j; \mathcal{W}). \quad (16)\end{aligned}$$

It is also important to note that the above definition is applicable, not only to document-to-term co-occurrences, but also to term-to-term, category-to-term, or document-to-descriptor co-occurrences. We will partly see in later sections how the definition is applied and extended for these different types of co-occurrence data.

3 Feature Quantity in Representative Terms Selection

3.1 Definition of $tfkli$ Measure

Relevant to the representative terms selection problem are: (i) automatic term extraction in computational terminology, and (ii) feature subset selection in machine learning. Approaches from computational terminology mainly concern the problem of determining the specificity of a term within a given document set, the result of which is used, for example, for information visualization in IR systems. On the other hand, approaches from the machine learning side mainly concern the problem of reducing the dimension of the features of the documents so that succeeding learning algorithms can effectively be applied, sometimes avoiding the over-fitting problem. In both cases, terms are characterized either by documents in which they occur, or by terms with which they co-occur.

Commonly used statistical measures in term extraction include term frequency, $tfidf$, document frequency, mutual information, log-likelihood ratio, signal-noise ratio, and Kullback-Leibler information [3] [14]. In machine learning, such statistical measures as mutual information, information gain, odds ratio, and expected cross entropy are used [16] [8]. Among these, the expected cross entropy [7] has exactly the same definition as our feature quantity. We would like to point out here that although there exist numbers of comparative studies in both fields, the expected cross entropy in the machine learning field has never been examined in the term extraction field as far as we know. Nor has it ever been pointed out that $tfidf$ and the expected cross entropy follow the same mathematical structure; that is, to express the significance of a term by the product of probability and information. This motivated us to compare these two measures in more detail using an actual data set.

As before, let f_{ij} be the number of occurrences of w_i within d_j, f_{w_i} be the total occurrences of w_i in all the documents, f_{d_j} be the total occurrences of all terms in d_j, and F be the total occurrences of all terms in all the documents, i.e., $F = \sum_{w_i \in W} \sum_{d_j \in D} f_{ij} = \sum_{d_j \in D} f_{d_j} = \sum_{w_i \in W} f_{w_i}$. The strategy to choose representative terms is to select ones with greater feature values. For simplicity, we assume here that the joint distribution $P(w_i, d_j)$ is simply determined by the observed occurrences, such that

$$P(w_i, d_j) = \frac{f_{ij}}{F}. \quad (17)$$

From Eqs. (12) and (17), the feature quantity of a term for the whole document set is calculated as:

$$\mathcal{F}(w_i; \mathcal{D}) = \frac{f_{w_i}}{F} \sum_{d_j \in D} \frac{f_{ij}}{f_{w_i}} log \frac{\frac{f_{ij}}{f_{w_i}}}{\frac{f_{d_j}}{F}}. \quad (18)$$

Since the selection criterion in the above equations is expressed as the product of term frequency and Kullback-Leibler information, we refer to such a measure as $tfkli$ in the following.

It is also possible to use joint distributions other than Eq. (17). For example, taking the influence of unobserved terms into account, the first term of Eqs. (19) and (18) can be substituted with $(f_{w_i} - \delta)/F$, where δ is the coefficient of absolute discounting [11]. When calculating a term's significance with respect to a specific document, the feature quantity is expressed as follows from Eqs. (11) and (17):

$$\mathcal{F}(w_i; d_j) = \frac{f_{ij}}{F} log \frac{\frac{f_{ij}}{f_{w_i}}}{\frac{f_{d_j}}{F}} . \qquad (19)$$

3.2 Experiments to Compare *tfidf* and *tfkli* Measures

Comparing Eqs. (18) and (19) with the traditional definition of $tfidf(w_i)$ $(= (f_{w_i}/F) \times log(N/N_i))$, it becomes clear that with our information theoretic formulation, *idf* and Kullback-Leibler information play similar roles. Specifically, these two quantities match when the following conditions are satisfied:

(C1) $f_{d_j}/F \approx 1/N$
(C2) $f_{ij}/f_{w_i} \approx 1/N_i$

Here, $C1$ means all the documents have almost equal sizes, while $C2$ indicates that the occurrence of a term does not differ much across the documents. For example, these conditions naturally hold when the document set under consideration is a collection of relatively short articles. Also, $C2$ is automatically satisfied when f_{ij} is given as a Boolean value, i.e., either 1 (occurs) or 0 (does not occur).

In the following experiments, we have actually calculated and compared the values of $tfidf$ and $tfkli$ for each term. The objective is to investigate the appropriateness of interpreting $tfidf$ as a variation of our feature quantity. Two different types of data sets are used in the experiments:

(D1) 2,106 abstracts of academic conference papers registered by the Japanese Society of Artificial Intelligence, and
(D2) 24 groups of abstracts of academic conference papers, in total 327,880, each group of which corresponds to a different academic society.

Each abstract is downloaded from the NACSIS Academic Conference Paper Database, also used in the NTCIR Workshop [10], and then is processed by a Japanese morphological analyser to extract nouns and also compound nouns. Although the language used in the corpus is Japanese, the result is language-independent since we only use the co-occurrence statistics for the numerical comparison of the two measures.

For data set $D1$, we assume that each abstract corresponds to a single document. The average size of a document is 93.4 words with the standard deviation being 33.0 words, figures which indicate conditions $C1$ and $C2$ are satisfied in this case. For data set $D2$, a group of abstracts presented at the same academic society is considered to be a single document. In this case, the size variation between documents is extremely large: while the largest document contains about 25% of the total 31,450,032 terms, the smallest one contains only 0.6% of the total. This implies the similarity conditions $C1$ and $C2$ no longer hold.

Figure 2 shows the results where $tfidf$ and $tfkli$ values are calculated for all the different terms using Eq. (18), and then summed up for the terms with the same frequency (f_{w_i}). The horizontal axis represents the frequency of the term and the vertical axis represents the averaged (for (a), (b), (d) and (e)) or the totalized (for (c) and (f)) feature values. For example, with $D2$, the $tfidf$ values for all the terms with frequency 1 is $log24$ and thus the point $(1, log24)$ is plotted on the graph (d).

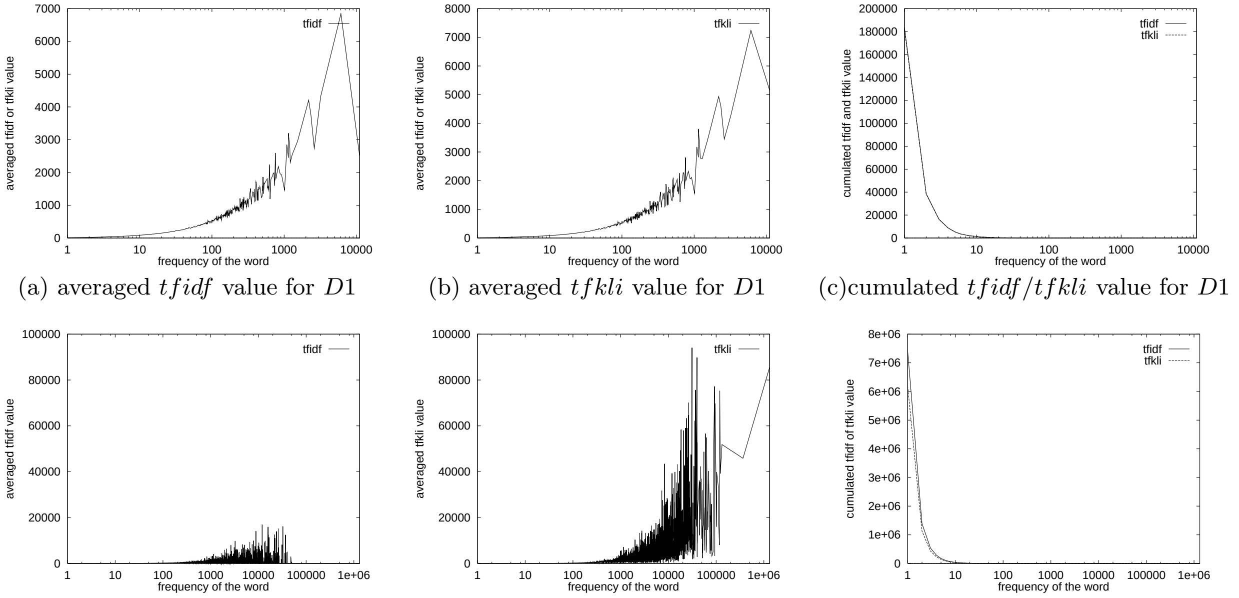
From these results, we can confirm our expectation that the values of $tfidf$ and $tfkli$ are almost identical for $D1$, but differ greatly for $D2$. Although only averaged values are shown in the figure, we have also examined the correlation between $tfidf$ and $tfkli$ on an individual term basis. The correlation coefficient for middle frequency terms is about 0.9 $\sim$ 1.0 for $D1$, and 0.6 $\sim$ 0.7 for $D2$. It is also important to note that on comparing the amount of information totalized for each frequency, low frequency terms contain a considerable amount of information, while the amount of information contributed by each term is small. Because of the exponential nature of term frequency distribution, known as Zipf's Law, this tendency remains unchanged even when considering the effect of unobserved terms.

3.3 Expanding the notion of *tfidf*

The theoretical development and experimental results show that $tfidf$ and $tfkli$ values are highly correlated for a relatively homogeneous document set, in which case, *idf* provides a simple but robust estimate of the information. It has also been observed that $tfidf$ and $tfkli$ values differ much for the data set composed of heterogeneous documents. Then, the next question is: *which works better under which conditions?*

We do not examine this issue in the present paper, however, for the following reason. Remember that $tfidf$ assumes equal probability for all the documents with w_i. In the information theoretic framework, such a strategy maximizes the entropy of $\mathcal{D}$ under the restriction that only the N_i documents with w_i have nonzero probabilities. On the other hand, $tfkli$ simply uses the observed frequency as the estimate of the real probability of w_i without considering the effect of finite sampling. The optimal allocation, if any, should be somewhere in the middle. However, the issue is related to the selection of a probabilistic model rather than the definition of a significance measure, and is beyond the scope of the paper. Instead of further comparing the behaviour of $tfidf$ and $tfkli$, we show in the following how the notion of $tfidf$ can be expanded using the information theoretic framework.

In the conventional studies, it is repeatedly pointed out that simple term frequency places too much emphasis on high frequency terms, while mutual and other information criteria allocate too much weight to low frequency terms. It is also widely recognized that the classical definition of $tfidf$, with its linear scaling in term frequency, again lays too much stress on high frequency terms. The difficulty is in establishing a good balance between frequency and information. This problem can easily be reformulated using the probabilistic framework shown in Figure 1. Namely, the "term frequency" of $tfidf$ actually refers to the probability of terms submitted to the system, and can be determined independently from the frequency of terms in the target documents. Therefore, the tf factor may be proportional to the square root of the term frequency, logged, or equally weighted, depending on the system's expectation about the query terms.



(a) averaged $tfidf$ value for $D1$ (b) averaged $tfkli$ value for $D1$ (c)cumulated $tfidf/tfkli$ value for $D1$

(d) averaged $tfidf$ value for $D2$ (e) averaged $tfkli$ value for $D2$ (f) cumulated $tfidf/tfkli$ value for $D2$

Figure 2: Comparison of $tfidf$ and $tfkli$ values.

In particular, the rough but simplified assumption of $tfidf$ allows us the following simple extension. Let us consider k independent trials, in each of which a single term is picked randomly from W with probability $P(w_i) = f_{w_i}/F$. Let $\mathcal{W}_i^k$ be an event that w_i appears at least once in the selected k terms. It immediately follows that the probability of $\mathcal{W}_i^k$ equals the probability that w_i is not selected at all subtracted from 1; that is, $P(\mathcal{W}_i^k) = 1 - \left(1 - \frac{f_{w_i}}{F}\right)^k$. Since $tfidf$ employs the rule-of-thumb assumption that all the documents are equally likely at the initial stage and also that all the documents that contain w_i are equally likely given w_i, the information obtained by $\mathcal{W}_i^k$ does not depend on the number of times w_i is observed. Then, the feature quantity using *idf* can be simply expressed as

$$tfidf^k(w_i) = \left(1 - \left(1 - \frac{f_{w_i}}{F}\right)^k\right) \; log\frac{N}{N_i}. \qquad (20)$$

If we let $k = 1$, Eq. (20) equals the definition of $tfidf$. On the other hand, if we let $k = \infty$, Eq. (20) equals the definition of *idf*. However, in general situations where conditions $C1$ and $C2$ do not hold, the calculation becomes more complex.

Although the above extension seems trivial, the important implication is that the specificity of terms can be defined so that it changes depending on the hypothesized situation, i.e., the distribution of terms submitted to the system. Eq. (20) provides a flexible way of trading off frequency and information ranging from $tfidf$ to *idf*. Accordingly, the abstract level of the selected terms is changed from general to concrete. Table 1 illustrates how the top 10 ranked terms of Reuters-21578 *acq* topic category change for values $k = 1, 10^3, 10^5$, where f_{ij} is used instead of f_{w_i} in calculating Eq. (20). The figures on the left side indicate the frequencies of the corresponding terms in the *acq* topic category. It can be seen that the terms become more specific as the k values increase.

Table 1: Example of effect of changing k with $tfidf^k$ measure with Reuters-21578 acq category.

$tfidf$		$tfidf^{1000}$		$tfidf^{100000}$	
5140	share	696	merger	221	cyclop
696	merger	313	usair	215	twa
977	acquir	977	acquir	154	purol
934	sharehold	934	sharehold	102	chemlawn
313	usair	221	cyclop	77	cyacq
847	stake	847	stake	60	emeri
221	cyclop	215	twa	54	comdata
215	twa	268	gencorp	50	pesch
268	gencorp	5140	share	47	norstar
4060	inc	901	acquisit	44	conrac

4 Feature Quantity in Text Categorization

4.1 Equations for fq-σ and fq-π Measures

Let $C = \{c_1, \cdots, c_N\}$ be a specified set of categories. Also, let $w^* = w_{i_1}, \cdots, w_{i_k}$ be a sequence of k terms representing a document to be categorized. The strategy for text categorization is to identify $c_j \in C$ that maximizes the feature quantity value, given w^* (Figure 3). Here, a category is viewed as a single collection of terms rather than a collection of independent documents and thus, Eq. (11), used for the calculation of the feature quantity across terms and documents, is also applicable for the calculation of the feature quantity across terms and categories. For notational simplicity, we denote the set of different terms in w^* as w^+, and the number of times w_i ($\in w^+$) occurs in w^* as h_i in the following.

Concerning the selection of w^*, two different formulations are considered.

(F1) In the first case, it is assumed that the k terms are selected from some unknown distribution.

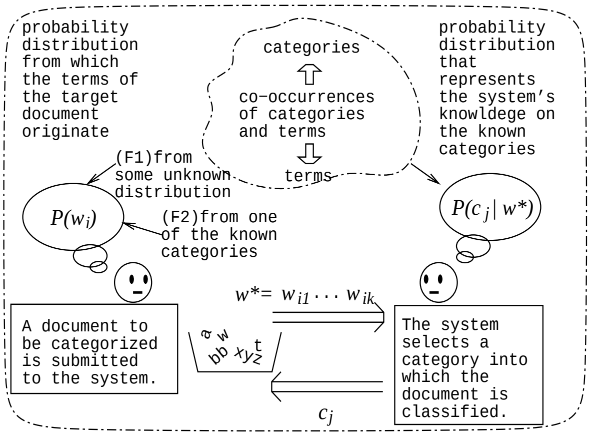

probability distribution from which the terms of the target document originate
categories
co-occurrences of categories and terms
terms
probability distribution that represents the system's knowldege on the known categories
(F1)from some unknown distribution
P(w_i)
(F2)from one of the known categories
P(c_j | w*)
w*= w_{i1} · · · w_{ik}
A document to be categorized is submitted to the system.
a w bb xyz t
The system selects a category into which the document is classified.
c_j

Figure 3: An illustrative situation assumed in the text categorization task.

(F2) In the second case, it is assumed that the k terms are selected from one of the categories in C.

The difference between $F1$ and $F2$ roughly corresponds to the difference between the existing vector and probabilistic IR models. In case $F1$, the target document is generated independently of the existing categories. Then, the objective of the categorization task is to find a category closest to the target document. In case $F2$, on the other hand, the target document comes from one of the existing categories. This time, the objective of the categorization task is to identify the category from which the document is most likely to have come. Of course in actual applications, these two formulations cannot explicitly be distinguished. Nevertheless, they require different mathematical treatments.

In case $F1$, the selection of the k terms is independent of the distribution of the terms in the existing N categories. Assuming $P(w_i)$; the distribution of the observed k terms, and $P(c_j|w_i)$, the probability of each category conditioned by w_i, are given; $P(c_j)$ is determined by

$$P(c_j) = \sum_{w_i \in W} P(w_i)P(c_j|w_i). \qquad (21)$$

Since terms in w^* are mutually independent, $P(w^*, c_j)$ $= P(c_j) \prod_{w_i \in w^+} P(w_i|c_j)^{h_i}$. Then, from the additivity property of the amount of information for independent events, the feature quantity of c_j and w^* can be expressed simply as the summation of the feature quantity of c_j and each $w_i \in w^*$:

$$\begin{aligned}\mathcal{F}(w^*, c_j) &= \sum_{w_i \in w^+} h_i \ \mathcal{F}(w_i, c_j) \\ &= \sum_{w_i \in w^+} h_i P(w_i, c_j) log \frac{P(w_i, c_j)}{P(w_i)P(c_j)} \\ &= \sum_{w_i \in w^+} h_i \ P(w_i)P(c_j|w_i) log \frac{P(c_j|w_i)}{P(c_j)}. \qquad (22)\end{aligned}$$

On the other hand, in case $F2$, the k terms are selected from the same existing category $c_j \in C$. Assuming that $P(c_j)$, the probability that c_j is selected as the origin of the k terms, and $P(w_i|c_j)$, the probability of w_i conditioned by c_j, are given, $P(w^*, c_j)$ is calculated as:

$$P(w^*, c_j) = P(c_j) \prod_{w_i \in w^+} P(w_i|c_j)^{h_i}. \qquad (23)$$

From $P(w^*) = \sum_{c_j \in C} P(w^*, c_j)$, $P(w^*)$ is given by:

$$P(w^*) = \sum_{c_j \in C} P(c_j) \prod_{w_i \in w^+} P(w_i|c_j)^{h_i}. \qquad (24)$$

Then, the feature quantity is calculated as:

$$\begin{aligned}&\mathcal{F}(w^*, c_j) \\ &= P(w^*, c_j) \ log \frac{P(w^*, c_j)}{P(w^*)P(c_j)} \\ &= P(c_j) \prod_{w_i \in w^+} P(w_i|c_j)^{h_i} \ log \frac{\prod_{w_i \in w^+} P(w_i|c_j)^{h_i}}{P(w^*)}. \qquad (25)\end{aligned}$$

Unlike the case in Eq. (22), the k terms are not independent of each other, and Eq. (25) cannot be simplified further.

Noting the different treatments of the conditional probabilities in Eqs. (22) and (25), we refer to these equations as fq-σ and fq-π, respectively.

4.2 Incorporating Different Probability Models

The interpretations of the classification tasks by $F1$ and $F2$ are based on different standpoints; that is, $F1$ uses $P(w_i)$ and $P(c_j|w_i)$ as primary distributions while $F2$ assumes $P(c_j)$ and $P(w_i|c_j)$ are given. However, if $P(w_i)$ and $P(c_j|w_i)$ are known, $P(c_j)$ and $P(w_i|c_j)$ are uniquely determined by Eq.(21) and from Bayes Theorem. The reverse is also true. This enables the comparison of $F1$ and $F2$ under the same probabilistic assumptions. In our experiments, the following three models are used for comparison.

The first model is chosen for $F1$ and referred to as $freq$. Since $F1$ assumes that the target document originated from some unknown distribution, $P(w_i)$ is set equal for all the terms, i.e., $P(w_i) = 1/M$ ($\forall w_i \in W$), where M is the total number of distinct terms. $P(c_j|w_i)$ is simply determined in proportion to the observed frequency. Denoting the frequency of w_i in category c_j as f_{ij}, the total frequency of w_i for all the categories as f_{w_i}, and $freq$ is defined as:

$$P(w_i) = \frac{1}{M}, \quad P(c_j|w_i) = \frac{f_{ij}}{f_{w_i}} \qquad (26)$$

The second model is chosen for $F2$ and referred to as *laplace*. With $F2$, consideration of unobserved events is crucial in determining the value of $P(w_i|c_j)$. The allocation policy of $freq$ does not work at all if w^* contains only a single unknown term because the probability $P(w^*, c_j)$ automatically becomes zero for all the categories. *Laplace*, given by the following equations, provides a simple solution to the zero frequency problem:

$$P(c_j) = \frac{f_{c_j}}{F}, \quad P(w_i|c_j) = \frac{1 + f_{ij}}{M + f_{c_j}} \qquad (27)$$

where f_{c_j} is the total frequency of terms in category c_j. This estimation is known as the Laplace estimator and is often used in the conventional naive Bayes approaches in the text categorization field.

Now, regardless of the convenience of the Laplace estimator, it has been widely recognized in probabilistic language model studies that the estimator does not provide a good fit compared with other dedicated discounting methods [6]. Recent studies in the text categorization field have also shown that the performance of naive Bayes categorization is sensitive to the estimation of $P(w_i|c_j)$ and that there exist cases when other event models outperform the Laplace estimator [9]. Considering these, we introduce a new probability model that can deal with both $F1$ and $F2$. The third model, referred to as *mixture*, is expressed as the mixture distribution of $P(c_j)$ in Eq. (27), and $P(c_j|w_i)$ in Eq. (26):

$$P(w_i) = \frac{f_{w_i}}{F}, \quad P(c_j|w_i) = (1-r)\frac{f_{c_j}}{F} + r\frac{f_{ij}}{f_{w_i}} \qquad (28)$$

where the mixture ratio r is determined by analysing actual corpus statistics. Specifically, we have observed that such a model provides a good fit with our corpus, but the details are still under investigation.

4.3 Correspondence with Naive Text Categorization Methods

Fq-σ given by Eq. (22) is closely related to the similarity measure widely used in the present IR systems: the normalized inner product of term vectors weighted by *tfidf*. Among many of its variations, an example of the conventional measure, which is referred to as *tfidf-cos* in this paper, is expressed as:

$$tfidf(w^*, c_j) = \sum_{w_i \in w^+} \frac{h_i \times f_{ij} \times log\frac{N}{N_i}}{||w^*||\ ||c_j||}. \qquad (29)$$

The normalization factor is given by $||w^*|| = \sum_{w_i \in w^+} {h_i}^2$ and $||c_j|| = \sum_{w_i \in W}(f_{ij} log\frac{N}{N_i})^2$, with f_{ij} being the frequency of w_i within category c_j. On the other hand, the definition of *fq-σ* with *freq* probability allocation is rewritten as:

$$\mathcal{F}(w^*, c_j) = \frac{1}{M}\sum_{w_i \in w^+} h_i \times \frac{f_{ij}}{f_{w_i}} \times log\frac{\frac{f_{ij}}{f_{w_i}}}{\sum_{w_i \in W}\frac{f_{ij}}{f_{w_i}}\frac{1}{M}} \qquad (30)$$

Comparing Eq. (29) with Eq. (30), it becomes clear that both *tfidf-cos* and *fq-σ* entail the same form of the "summation of $h_i \times f_{ij} \times log(\cdot)$". The difference is in their normalization and in their consideration of the amount of information in the log terms. Based on this, we can expect that the performance of these two categorization methods match well when the target document set conforms with $C1$ and $C2$ in the previous section.

Next, *fq-π* given by Eq. (25) has a clear correspondence with the naive Bayes method popularly used in conventional text categorization studies. The method, referred to as *n-bayes* in this paper, is based on the maximum likelihood principle that selects the category with the largest probability $P(c_j|w^*)$. Thus, the classification strategy can be expressed as:

$$nbayes(w^*, c_j) = P(c_j|w^*) = \frac{P(c_j, w^*)}{P(w^*)}$$

$$= \frac{P(c_j)\prod_{w_i \in w^+} P(w_i|c_j)^{h_i}}{P(w^*)} \qquad (31)$$

where the term $P(w^*)$ can be omitted since it is common for all the categories.

Comparing Eq. (31) with Eq. (22), it transpires that the only difference between *n-bayes* and *fq-π* is the consideration of the amount of information expressed as the *log* term in Eq. (25). The difference is most clearly demonstrated in the extreme case when $k = 0$ ($w^* = \emptyset$). Without any information available, *n-bayes* selects the most frequent category as the most plausible, while *fq-π* judges all the categories to be equally likely. However, for larger values of k, the difference between Eq. (25) and Eq. (31) is negligible and does not have significant influence on the classification results.

4.4 Text Categorization Experiments

In the following experiments, three different measures for text categorization are compared: (i) *fq-σ*, (ii) *tfidf-cos*, and (iii) *n-bayes/fq-π*. The data set used is again abstracts of academic papers extracted from the NACSIS Academic Conference Paper Database. The categorization task is to identify the society, or the class of societies, of an unknown abstract data. Since the size of each abstract is sufficiently large (the average is about 90 words per abstract), there exists no meaningful difference between the *n-bayes* and *fq-π* methods.

The training data is the same as data set $D2$ in section **3.2**, which is composed of a total of 327,880 abstracts from 24 academic societies. Each category corresponds to either

(P1) one of 24 academic societies, or
(P2) one of the two society classes; information technology related or not related.

With $P1$, the sizes of the 24 categories vary greatly, from a maximum of 7,986,568 terms to a minimum of 187,290 terms per category. With $P2$, the two society classes are of almost equal size, with the ratio about 46% to 54%. Based on these figures, we can expect that *tfidf-cos* works more consistently with *fq-σ* for $P1$ than for $P2$.

The categorization task is formulated as a multi-class problem, since each abstract belongs to one and the only society/society class. In the evaluation, a total of 10,000 abstracts are prepared that are not contained in the training data, but with the same distribution across categories. Therefore, if about 25% abstracts of the training data belong to society A, then the test data also contains 25% abstracts from society A. The performance is compared using the ratio of the correct judgments, i.e., the number of abstracts classified into the class that they originally belong to, divided by 10,000. The size of the training data set is varied as either 1,000, 10,000, or 327,880 for each of $P1$ and $P2$ so that the scalability of different categorization strategies can be compared. For all combinations, three probability models, *freq*, *laplace*, and *mixture* are tested. The results are summarized in Table 2.

From these results, we can confirm that *tfidf-cos* performs well with $P2$. In the cases with $P1$, the performance of *tfidf-cos* is considerably degraded as the size of the training data becomes large. In contrast, better performance is obtained for larger sizes of the training data for *fq-σ* and *n-bayes*. The best performance values of these two methods are almost equal. Comparing the probability models, *freq* is not applicable for *n-bayes* due to the zero frequency problem we have already mentioned,

Table 2: Result of Classification Task

(a) Results for $P1$ with 24 categories

Methods	$P(d_j\|w_i)$	the size of the training data		
		$327,880$	$10,000$	$1,000$
tfidf-cos	freq	0.6835	0.7128	0.6596
	laplace	0.6914	0.5486	0.4657
	mixture	0.6840	0.7074	0.6368
fq-σ	freq	0.8097	0.7516	0.6523
	laplace	0.6537	0.5463	0.5155
	mixture	0.8158	0.7558	0.6552
n-bayes	freq	0.0000	0.0000	0.0000
(*fq-π*)	laplace	0.7099	0.6055	0.5626
	mixture	0.7929	0.7646	0.6603

(b) Results for $P2$ with two categories

Methods	$P(d_j\|w_i)$	the size of the training data		
		$327,880$	$10,000$	$1,000$
tfidf-cos	freq	0.9709	0.9483	0.9194
	laplace	0.9704	0.9475	0.9185
	mixture	0.9701	0.9437	0.9108
fq-σ	freq	0.9595	0.9459	0.9212
	laplace	0.9575	0.9405	0.9116
	mixture	0.9625	0.9532	0.9289
n-bayes	freq	0.0009	0.0000	0.0000
(*fq-π*)	laplace	0.9658	0.9516	0.9231
	mixture	0.9689	0.9579	0.9329

while with *tfidf-cos* and *fq-σ*, *freq* seems to be reasonably good. Also, *laplace* does not work well with $P1$ for both *fq-σ* and *n-bayes*. *Mixture* performs consistently well in all the cases.

Lastly, we have so far only referred to the naive *tfidf-cos* and *n-bayes* text categorization methods. Other existing methods include *Rocchio* with Relevance feedback [2], *Prtfidf* based on probabilistic analysis [5], and a variety of machine learning methods such as LLSF, C4.5, k-NN, and Support Vector Machine [12][15]. However, we have observed that with as many as 2,300,000 feature terms, the naive methods sometimes outperform other more complicated methods. For example, the standard learning algorithms such as C4.5 or kNN can easily be applied if the feature dimension is reduced to 5,000. Another investigation using the same document set showed that the performance is around 0.7 for $P1$ [13], which is smaller than the values 0.75 to 0.80 obtained in our experiments. Of course, the difference may be overturned by the fine tuning of the learning methods. Nevertheless, we think the issue needs further investigation since the data used in our experiments is different from Reuters-21578, the standard test set for text categorization tasks today, both in its large scale and also in its specificity to the academic fields. Such phenomena may partly be explained by the fact that academic documents are highly domain specific and a considerable amount of information is carried by low frequency terms, as has been shown in Figure 2(c)(f).

5 Remarks

In this paper, we have introduced the feature quantity, a quantitative representation of specificity of textual data components. Although our investigation in this paper mainly concerns the consistency of the proposed feature quantity with, and not its superiority to, conventional statistical measures, we believe such an approach is worth trying since the information theoretic view helps us to apply the same mathematical framework to different target problems such as representative term selection, text categorization, and also automatic identification of collocations and translation pairs. In addition, such a view enables us to shed light on commonly practised heuristics such as *tfidf* or discarding of low frequency data at the pre-processing stage.

Acknowledgement

This work is part of the project "A Study on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources", funded by the Japan Society for the Promotion of Science. The authors would specially like to thank Kyo Kageura, Atsushi Takasu, and Kenro Aihara at the National Institute of Informatics for sharing experimental data and for their helpful discussions.

References

[1] G. Amati and K. van Rijsbergen. *Semantic Information Retrieval*, 189–219. Kluwer Academic Pub., 1998. (in "Information Retrieval: Uncertainty and Logics").

[2] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *SIGIR'94*, 292–300, 1994.

[3] S. A. Caraballo and E. Charniak. Determining the specificity of nouns from text. In *EMNLP'99*, 1999.

[4] W. R. Greiff. A theory of term weighting based on exploratory data analysis. In *SIGIR'98*, 11–19, 1998.

[5] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML'97*, 143–151, 1999.

[6] K. Kita. *Probabilistic Language Model.* University of Tokyo Press, Japan, 1999.

[7] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *ICML'97*, 170–178, 1997.

[8] D. Maldenić and M. Grobelnik. Feature selection for classification based on text hierarchy. In *Working notes of Learning from Text and the Web, CONALD'98*, 1998.

[9] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on learning for text categorization*, 42–49, 1998.

[10] NACSIS, editor. *NTCIR Workshop 1 - proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition.* National Center for Science Information Systems, 1999.

[11] H. Ney, S. Martin, and F. Wessel. *Statistical Language Modeling using Leaving-one-out*, 174–207. Kluwer Academic Pub., 1997. (in "Corpus-Based Methods in Language and Speech Processing").

[12] Y. Singer and D. D. Lewis. Machine learning for information retrieval: Advanced techniques. In *SIGIR'99 Tutorial*, 1999.

[13] A. Takasu and K. Aihara. Variance based classifier comparison in text categorization (poster). In *SIGIR 2000*, 2000.

[14] S. Wong and Y. Yao. An information theoretic measure of term specificity. *Journal of the American Society for Information Science*, 43(1):54–61, 1992.

[15] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR'99*, 42–49, 1999.

[16] Y. Yang and O. Pedersen. A comparative study on feature selection in text categorization. In *ICML'97*, 412–420, 1997.