PERGAMON

Information Processing and Management 39 (2003) 45–65

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

An information-theoretic perspective of tf–idf measures ☆

Akiko Aizawa *

National Institute of Informatics, 2-1-2 Hitotsubashi Chiyoda-ku, Tokyo 101-8430, Japan

Received 4 August 2001; accepted 4 January 2002

Abstract

This paper presents a mathematical definition of the "probability-weighted amount of information" (PWI), a measure of specificity of terms in documents that is based on an information-theoretic view of retrieval events. The proposed PWI is expressed as a product of the occurrence probabilities of terms and their amounts of information, and corresponds well with the conventional term frequency–inverse document frequency measures that are commonly used in today's information retrieval systems. The mathematical definition of the PWI is shown, together with some illustrative examples of the calculation.
© 2002 Elsevier Science Ltd. All rights reserved.

Keywords: tf–idf; Term weighting theories; Information theory; Text categorization

1. Introduction

"Term frequency–inverse document frequency" (tf–idf) is one of the most commonly used term weighting schemes in today's information retrieval systems. Despite its popularity, tf–idf has often been considered an empirical method, specifically from a probabilistic point of view, with many possible variations. In this paper, we first revisit the classical, but nevertheless important, question 'What is the mathematical implication of tf–idf?' in an information-theoretic framework.

In the literature, many studies relate to the problem of quantifying the significance of terms (for example, see Baeza-Yates & Ribeiro-Neto, 1988; Kageura & Umino, 1996; Manning & Schtüze, 1999). In the information retrieval field, term weights are mainly used to represent the usefulness of terms in the retrieval process; for example, frequency (Luhn, 1957), signal-to-noise ratio

☆ Parts of the results were presented at ACM SIGIR 2000.

* Tel.: +81-3-4212-2524; fax: +81-3-3556-1916.
E-mail address: akiko@nii.ac.jp (A. Aizawa).

0306-4573/03/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(02)00021-3

(Dennis, 1964; Salton & McGill, 1983), idf (Sparck-Jones, 1972), relevance weighting methods (Robertson & Sparck-Jones, 1976) and tf–idf and its variations (Salton & Buckley, 1988).

Other relevant fields include automatic term extraction in computational terminology, and also 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 purpose of which is to construct automatically a basic terminology dictionary from a corpus of a specific subject. Examples of commonly used statistical measures in term extraction include chi-squared statistics (Nagao, Mizutani, & Ikeda, 1976), pairwise mutual information (Church & Hanks, 1990), Dice coefficient (Smadja, 1993), log-likelihood ratio (Dunning, 1993) and Jaccard similarity measure (Grefenstette, 1994).

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 subsequent learning algorithms can be applied effectively, sometimes avoiding the over-fitting problem. In machine learning, such statistical measures as document frequency (Yang & Pedersen, 1997), information gain (Lewis & Ringuette, 1994; Yang & Liu, 1999), chi-squared statistics (Wiener, Pedersen, & Weighend, 1995), odds ratio (Mladenić, 1998) and expected cross entropy (Koller & Sahami, 1996, 1997) are used. There are also a great number of comparative studies in each field, from both theoretical and empirical points of view.

The present research is motivated by the assumptions that the long history of such a variety of measures is in itself evidence of the unfeasibility of determining the superiority of specific measures. It also suggests that the issue is a problem of statistical parameter estimation rather than simply a comparison of the performance of different measures. Following these assumptions, in this paper we try to clarify the statistical model on which the different measures are commonly based. We expect that such an investigation will be useful not only for comparing the performance of different measures, but also for developing a better method of selecting and verifying a statistical model in connection with conventional statistical language modelling studies.

The conventional measures of term significance are basically derived from a statistical table of co-occurrence frequencies, co-occurrences of terms either with documents in which they occur, or with other terms with which they co-occur in the same documents (or text segments). The majority of the existing probabilistic approaches initially calculate the probabilities of terms using a statistical table, and then use these values to estimate certain statistical values such as the conditional probabilities of documents, or the information gain of a specific term, depending on their objectives. However, for some heuristic measures, such a statistical calculation is not clearly defined: they are tf–idf and its variations. The problem becomes one of characterizing these measures within the same probabilistic framework.

Based on this background, we show an information-theoretic interpretation of tf–idf in this paper. By its definition, tf–idf is a metric that multiplies the two quantities tf and idf. Here, tf provides a direct estimation of the occurrence probability of a term when it is normalized by the total frequency in the document, or the document collection, depending on the scope of the calculation. Note that the normalization factor is common for all the terms in the scope, and thus can be omitted. On the other hand, idf can be interpreted as 'the amount of information' in conventional information theory (Brookes, 1972; Wong & Yao, 1992), given as the log of the inverse probability (Cover & Thomas, 1991). Bearing these 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 that it represents. Although conventional information theory does not deal explicitly with such a quantity (but uses one in the calculation of entropy, which is generally defined as the expected amount of information), we have postulated that the current popularity of the tf–idf measure suggests the usefulness of such a quantity as a measure of significance (Aizawa, 2000).

The remainder of this paper reports some of the preliminary results of our attempt to expand such ideas. The subsequent sections are organized as follows. Section 2 provides a simple overview of the conventional measures of term significance. Section 3 presents the information-theoretic interpretation of tf–idf with its extended notion of the probability-weighted amount of information (PWI). Section 4 deals with issues in selecting probabilistic models. Section 5 shows examples of the calculation of tf–idf and the newly proposed PWI using a large-scale document collection. Section 6 gives our conclusions.

2. A brief look at conventional statistical measures

2.1. Classifying the conventional measures

In this section, we provide two different views of classifying the conventional measures of term significance. Our first categorization is based on the ways that these measures are used.

(1) *Measures for term selection:* Used for selecting terms that are important in a given document set. Examples include the selection of query terms in relevance feedback and feature term selection in automatic text categorization.
(2) *Measures for term weighting:* Used for measuring the relevance of a term within a specific document. Most of the term weighting schemes in information retrieval are categorized into this group.

Such a categorization is motivated by Robertson (1990, p. 364) with the following statement: "A term weighting formula that provides appropriate weights for use in a match function for retrieval is not necessarily an appropriate measure for term selection in the first place". Although we do not specifically focus on relevance feedback as Robertson (1990) did, we follow their definition in this paper.

Our second categorization is based on the way that these measures are mathematically defined.

(a) *Measures of popularity:* Defined based on the frequencies of terms or the estimated probabilities of their occurrences. These measures employ a simple assumption that frequent terms are also significant (Luhn, 1957).
(b) *Measures of specificity:* Defined based on the amount of information or the entropy of terms. These measures quantify the deviations from randomness of the occurrences of terms. Examples of such measures include pairwise mutual information (Church & Hanks, 1990), signal-to-noise ratio (Dennis, 1964; Salton & McGill, 1983) and idf (Sparck-Jones, 1972).
(c) *Measures of discrimination:* Defined based on the contribution of terms to the performance of a specified discrimination function. These measures represent the power of distinguishing

relevant and non-relevant documents, or documents from different categories. Examples of such measures include information gain (Lewis & Ringuette, 1994; Yang & Liu, 1999) and relevance weighting (Robertson & Sparck-Jones, 1976).

(d) *Measures of representation:* Defined based on the product of term frequency and the inverse of log-scaled document frequency, i.e., tf–idf and its variations (Salton & Buckley, 1988). These measures quantify the extent of usefulness of terms in characterizing the document in which they appear.

In conventional studies of computational linguistics, it is repeatedly pointed out that (a) a simple frequency measure places too much emphasis on high frequency terms, while (b) specificity measures such as pairwise mutual information allocate too much weight to low frequency terms (for example, in Chapter 5 of Manning & Schtüze, 1999). The difficulty with selecting or weighting terms lies in establishing a good balance between popularity and specificity. Both (c) and (d) above, the measures of discrimination and representation, can be interpreted as accomplishing such balancing in some sense, although they apply different strategies.

The difference between (c) and (d) becomes clear when considering the following situation. Suppose a term exists that appears frequently in all the documents except one. Such a term is useful for distinguishing the exceptional document from the others, but hardly serves as a good index for the document without the term. Therefore, the weight of such a term regarding the exceptional document becomes relatively high with (c) and low with (d). In information retrieval, (c) is often associated with relevance feedback with query expansion, while (d) concerns more the representation in the vector space model.

Table 1 summarizes the two categorization methods and corresponding examples of conventional measures. Because measures for term selection and measures for term weighting are not necessarily distinguished in the actual application of information retrieval (Robertson, 1990), both relevance weighting and tf–idf are used in both methods.

2.2. Related theoretical studies

In the information retrieval field, measures of discrimination are closely related to the probabilistic retrieval model (Robertson & Sparck-Jones, 1976). Although these measures are theoretically well sustained, we should note here that they still require empirical adjustment in estimating certain probabilities. For example, "the probability that a term is present in a relevant document" is required in the calculation, but such a value can be estimated using only empirical techniques at the beginning.

Table 1
Examples of types of measures to represent the significance of terms

	Measures of popularity	Measures of specificity	Measures of discrimination	Measures of representation
Measures for term selection	Total term frequency	idf signal-to-noise ratio	Information gain relevance weighting	(Total term frequency) × idf
Measures for term weighting	Within-document frequency	Pairwise mutual information	Relevance weighting	(Within-document frequency) × idf

Measures of representation, on the other hand, are generally associated with the vector-space retrieval model in information retrieval (Salton & McGill, 1983). Even though tf–idf tends to be considered as a convenient heuristic, its effectiveness has been justified through the long history of information retrieval. Note also that numerous variants of tf–idf exist, and selecting an appropriate formula from these variants requires some skill when a new data set is studied.

There are many theoretical studies concerning the mathematical interpretation of idf. For example, Croft and Harper (1979) derived an equation for idf in the context of the binary independence model. Wong and Yao (1992) compared idf with signal-to-noise ratio and showed that both measures can be explained using Shannon's entropy. Church and Gale (1999) examined the gap between observed and predicted idf values. Through empirical studies, they showed that larger idf values mean larger deviations from Poisson and therefore more 'context' regarding the terms. Greiff (1998) argued the relationship between pairwise mutual information and idf, from which the efficacy of idf was theoretically justified. Note that these theoretical or experimental results are targeted to idf, not tf–idf.

As for the mathematical consideration of tf–idf, Joachims (1997) investigated tf–idf with a probabilistic framework, and proposed a new measure, called PrTFIDF, for text categorization. In their formulation, PrTFIDF was defined as the posterior probability of each category using the retrieval with probability indexing (RPI) model proposed by Fuhr (1989). Note that PrTFIDF is similar to, but does not exactly correspond to, tf–idf, and thus their formulation does not directly explain tf–idf in a probabilistic framework. Recently, Hiemstra (2000) formulated tf–idf as the conditional probability of each document for given query terms. Their explanation used two distributions: one defined using document frequencies of terms, and the other using within-document term frequencies. Although our approach in this paper is based on a different starting point, we used a similar combination of distributions in our experiment.

Although the probabilistic and vector-space models have long been conceived as alternatives, a few studies have mentioned the duality of the two models. Robertson (1994), considering the symmetric representation of queries and documents in information retrieval systems, discussed the potential advantages of, and the objections to, the dual models. Amati and Van Rijsbergen (1998) formulated a duality theory in which the probabilistic and vector-space models were defined as two probabilistic models, each the dual of the other. Because their view of the problem matches well with our formulation, we will refer to their work again.

3. Extending the notion of tf–idf

3.1. Basic formulae of information theory

We first introduce some of the basic formulae of information theory (Cover & Thomas, 1991) that we use in our theoretical development. Let x_i and y_j be two distinct events from finite event spaces X and Y. Assume a joint probability distribution $P(x_i, y_j)$ is given for $x_i \in X$ and $y_j \in Y$. Using $P(x_i, y_j)$ and the definition of the marginal distribution, it immediately follows that the probability that x_i is observed is

$$P(x_i) = \sum_{y_j \in Y} P(x_i, y_j), \tag{1}$$

and the probability that y_j is observed is

$$P(y_j) = \sum_{x_i \in X} P(x_i, y_j). \tag{2}$$

The above equations simply mean that if x_i is observed, it is always observed with some $y_i \in Y$ and vice versa.

The basic quantity in information theory is *the amount of information*, which is defined as the log of the inverse of the probability, i.e., $\log(1/P(x_i)) = -\log P(x_i)$. Now, let $\mathscr{X}$ and $\mathscr{Y}$ be random variables representing distinct events in X and Y, which occur with certain probabilities. The amount of information expected for $\mathscr{X}$ or $\mathscr{Y}$ is called the *self-entropy* and is denoted as $\mathscr{H}(\mathscr{X})$ or $\mathscr{H}(\mathscr{Y})$ in this paper. By the general definition of information theory, the self-entropy of $\mathscr{X}$ is calculated as

$$\mathscr{H}(\mathscr{X}) = -\sum_{x_i \in X} P(x_i) \log P(x_i), \tag{3}$$

and the self-entropy of $\mathscr{Y}$ is calculated as

$$\mathscr{H}(\mathscr{Y}) = -\sum_{y_j \in Y} P(y_j) \log P(y_j). \tag{4}$$

The self-entropy expresses the degree of uncertainty about which an event will occur in a future observation. Naturally, the amount becomes higher for larger numbers of events with equally likely probabilities.

The *pairwise mutual information* between x_i and y_j is the difference between the amounts of information based on (i) the actual joint probability, $P(x_i, y_j)$, and (ii) the expected probability when the independence of the two events are assumed, $P(x_i)P(y_j)$. Denoting the pairwise mutual information as $\mathscr{M}(x_i, y_j)$, the definition is given as

$$\mathscr{M}(x_i, y_j) = \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \tag{5}$$

The pairwise mutual information is sometimes used as a measure for extracting collocations or other lexical tiers in computational linguistics (for example, Church & Hanks, 1990).

On the other hand, the *expected mutual information*, or simply the *mutual information* between $\mathscr{X}$ and $\mathscr{Y}$, represents the reduction of uncertainty about either $\mathscr{X}$ or $\mathscr{Y}$ when the other is known. Denoting the mutual information $\mathscr{I}(\mathscr{X}; \mathscr{Y})$, the definition is given as

$$\begin{aligned} \mathscr{I}(\mathscr{X}; \mathscr{Y}) &= \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \mathscr{M}(x_i, y_j) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \\ &= \mathscr{H}(\mathscr{X}) - \mathscr{H}(\mathscr{X}|\mathscr{Y}) = \mathscr{H}(\mathscr{X}) + \mathscr{H}(\mathscr{Y}) - \mathscr{H}(\mathscr{X}\mathscr{Y}). \end{aligned} \tag{6}$$

The mutual information is a measure of interactions between the two random variables and is sometimes used, in the information retrieval field, to exploit statistical dependencies between terms (Crestani, 2000; Van Rijsbergen, Happer, & Porter, 1981). Note that by its definition, the mutual information is symmetric in $\mathscr{X}$ and $\mathscr{Y}$, i.e., $\mathscr{I}(\mathscr{X}; \mathscr{Y}) = \mathscr{I}(\mathscr{Y}; \mathscr{X})$.

The reduction of uncertainty of $\mathscr{Y}$ after observing a specific event x_i can be expressed using the *Kullback–Leibler information*, which is also used as a distance measure of two probability dis-

tributions. The Kullback–Leibler information between $P(\mathscr{Y}|x_i)$ and $P(\mathscr{Y})$, denoted as $\mathscr{K}(P(\mathscr{Y}|x_i)\|P(\mathscr{Y}))$ in this paper, [1] is calculated as

$$\mathscr{K}(P(\mathscr{Y}|x_i)\|P(\mathscr{Y})) = \sum_{y_j \in Y} P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)}. \tag{7}$$

Similarly, the reduction of uncertainty of $\mathscr{X}$ after observing a specific event y_j is given by the Kullback–Leibler information between $P(\mathscr{X}|y_j)$ and $P(\mathscr{X})$:

$$\mathscr{K}(P(\mathscr{X}|y_j)\|P(\mathscr{X})) = \sum_{x_i \in X} P(x_i|y_j) \log \frac{P(x_i|y_j)}{P(x_i)}. \tag{8}$$

Applying the general property of conditional probability, $P(x_i, y_j) = P(y_j|x_i)P(x_i) = P(x_i|y_j)P(y_j)$, to Eqs. (6)–(8), it is straightforward to show that the following relationships hold between the expected mutual information and the Kullback–Leibler information:

$$\mathscr{I}(\mathscr{X};\mathscr{Y}) = \sum_{x_i \in X} P(x_i)\mathscr{K}(P(\mathscr{Y}|x_i)\|P(\mathscr{Y})) = \sum_{y_j \in Y} P(y_j)\mathscr{K}(P(\mathscr{X}|y_j)\|P(\mathscr{X})). \tag{9}$$

If the occurrence of x_i is totally non-informative for $\mathscr{Y}$, then $P(\mathscr{Y}|x_i) = P(\mathscr{Y})$ and $\mathscr{K}(P(\mathscr{Y}|x_i)\|P(\mathscr{Y}))$ becomes zero. As each x_i provides more information for $\mathscr{Y}$, the values of $\mathscr{K}(P(\mathscr{Y}|x_i)\|P(\mathscr{Y}))$ and $\mathscr{I}(\mathscr{X};\mathscr{Y})$ become greater and vice versa. The correspondence becomes even clearer when the definition of the mutual information is rewritten as $\mathscr{I}(\mathscr{X};\mathscr{Y}) = \mathscr{K}(P(\mathscr{X},\mathscr{Y})\|P(\mathscr{X})P(\mathscr{Y}))$.

3.2. An information-theoretic view of tf–idf

Based on the above definitions, we next show a possible interpretation of tf–idf in view of conventional information theory. First, it is assumed that a document is given as an unordered set of terms. Let $D = \{d_1, \ldots, d_N\}$ be a set of documents and $W = \{w_1, \ldots, w_M\}$ 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 notation 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 $\mathscr{D}$ and $\mathscr{W}$ be random variables defined over the events $\{d_1, \ldots, d_N\}$ and $\{w_1, \ldots, w_M\}$, respectively. Introducing the random variables, we depict a situation where a query is submitted as a probability distribution over D, and the retrieval result is returned as a distribution over D. Our objective here is to calculate the expected mutual information between $\mathscr{D}$ and $\mathscr{W}$ to see how well documents are specified by the submitted queries (Fig. 1).

Assuming that all documents are equally likely candidates 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 $\mathscr{D}$ is

[1] In information theory, Kullback–Leibler information, or divergence, is usually denoted as $\mathscr{D}$. In this paper, we use the notation $\mathscr{K}$ because $\mathscr{D}$ is more familiar as *documents* in the information retrieval field.

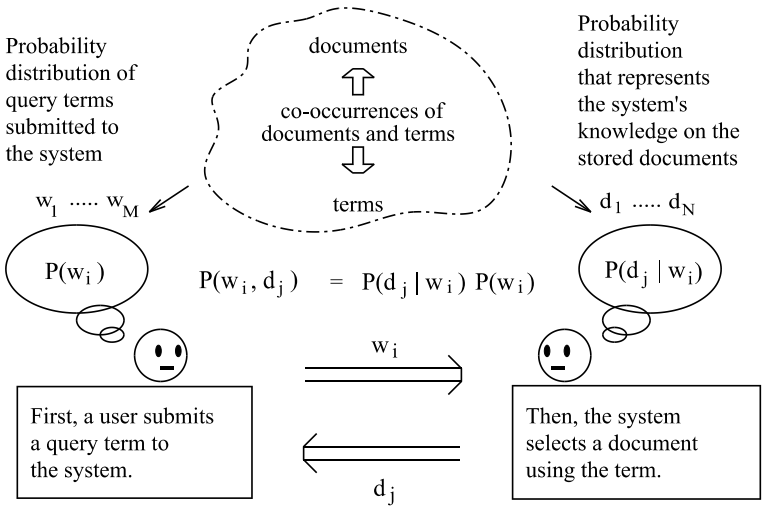


Fig. 1. An illustrative situation assumed in the calculation of the expected mutual information.

$$\mathscr{H}(\mathscr{D}) = -\sum_{d_j \in D} P(d_j) \log P(d_j) = -N \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N}. \tag{10}$$

Next, consider a situation where a subset of specified documents that contain w_i ($\in W$) is 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 $\mathscr{D}$ given w_i becomes

$$\mathscr{H}(\mathscr{D}|w_i) = -\sum_{d_j \in D} P(d_j|w_i) \log P(d_j|w_i) = -N_i \frac{1}{N_i} \log \frac{1}{N_i} = -\log \frac{1}{N_i}. \tag{11}$$

As we have assumed that documents without w_i occur with probability zero in the selected subset, 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 we randomly select a query term w_i 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 a specific w_i is selected is $\sum_j f_{ij}/F = f_{w_i}/F$. Then, the expected mutual information is calculated as

$$\begin{aligned} \mathscr{I}(\mathscr{D}; \mathscr{W}) &= H(\mathscr{D}) - H(\mathscr{D}|\mathscr{W}) = \sum_{w_i \in W} P(w_i)(H(\mathscr{D}) - H(\mathscr{D}|w_i)) \\ &= \sum_{w_i \in W} \frac{f_{w_i}}{F} \left(-\log \frac{1}{N} + \log \frac{1}{N_i} \right) = \sum_{w_i \in W} \frac{f_{w_i}}{F} \log \frac{N}{N_i} \qquad (12) \\ &= \sum_{w_i \in W} \sum_{d_j \in D} \frac{f_{ij}}{F} \log \frac{N}{N_i}. \qquad (13) \end{aligned}$$

Eqs. (12) and (13) represent the sum of the products of the tf, either in the form of f_{w_i} or f_{ij}, and the idf divided by a constant factor F. Hence, we conclude that from an information-theoretic point of view, tf–idf can be interpreted as the quantity required for the calculation of the expected mutual information that is given by Eq. (6). The idf factor expresses the change in the amount of information after observing a specific term, and the tf factor expresses the probability estimation that the term is actually observed. Note that Eqs. (12) and (13) provide two different views of tf–idf. When tf refers to f_{w_i}, tf–idf is considered as a measure for term selection, and Eq. (12), summarizing the value over all the existing words, represents the specificity of entire documents in the retrieval system. When tf refers to f_{ij}, tf–idf is considered as a measure for term weighting, and Eq. (13), summarizing the value over all the combinations of terms and documents, represents the decrease of uncertainty about the relevant documents as a result of the submitted query.

It should be noted that in the above derivation, some restrictive assumptions are used for consistency. First, it was assumed that the distribution of query terms is proportional to the observed frequency of terms in the target document. Second, it was implicitly assumed that

$$P(d_j) = \sum_{W(d_j)} \frac{f_{w_i}}{F} \frac{1}{N_i} \approx \frac{1}{N}, \tag{14}$$

and

$$P(w_i, d_j) = \frac{f_{w_i}}{F} \frac{1}{N_i} \approx \frac{f_{ij}}{F}, \tag{15}$$

where $W(d_j)$ is the set of distinct terms contained in d_j. In our view, these specific assumptions themselves represent the heuristic that tf–idf employs. The next question then becomes whether it is possible to extend the above definition of tf–idf into a more general form by applying the same information-theoretic framework. For this purpose, we now introduce a concept of PWI.

3.3. Definition of the probability-weighted amount of information

The PWI, as formulated in this section, is defined as the contribution of a specific co-occurrence event to the overall entropy calculation. In the following, we use the notation '$\delta\mathscr{I}$' to represent the PWI. From Eq. (6), the PWI value of the occurrence of w_i and d_j is defined as

$$\delta\mathscr{I}(w_i, d_j) = P(w_i, d_j)\mathscr{M}(w_i, d_j). \tag{16}$$

Similarly, from Eq. (9), the PWI of the occurrence of w_i is defined as

$$\delta\mathscr{I}(w_i; \mathscr{D}) = P(w_i)\mathscr{K}(P(\mathscr{D}|w_i), P(\mathscr{D})), \tag{17}$$

and the PWI of the occurrence of d_j as

$$\delta\mathscr{I}(d_j; \mathscr{W}) = P(d_j)\mathscr{K}(P(\mathscr{W}|d_j), P(\mathscr{W})). \tag{18}$$

In all cases, the PWI is expressed as a product of probability and information, the latter being either pairwise mutual information, in Eq. (16), or the Kullback–Leibler information, in Eqs. (17) and (18).

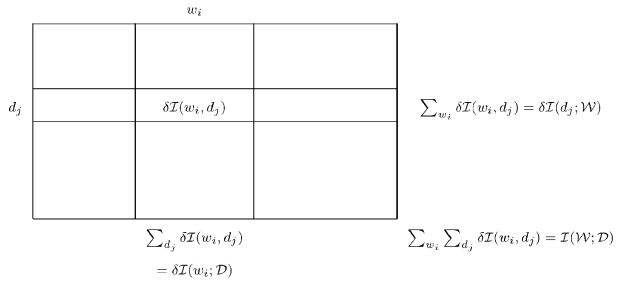


Fig. 2. The probability-weighted amount of information.

By definition, the mutual information of all co-occurrences is simply expressed as the summation of PWI values of each case:

$$\mathcal{I}(\mathscr{D}; \mathscr{W}) = \sum_{w_i \in W} \sum_{d_j \in D} \delta\mathcal{I}(w_i, d_j) = \sum_{w_i \in W} \delta\mathcal{I}(w_i; \mathscr{D}) = \sum_{d_j \in D} \delta\mathcal{I}(d_j; \mathscr{W}). \tag{19}$$

Because both tf–idf and PWI represent quantities such that their summation over all the event space equals the mutual information, we consider $\delta\mathcal{I}(w_i, d_j)$ and $\delta\mathcal{I}(w_i, \mathscr{D})$ as generalized definitions of tf–idf for term weighting and tf–idf for term selection, respectively.

The calculation is illustrated in Fig. 2. The table in the figure is similar to the contingency table of terms and documents, except that each cell represents not the frequency of the co-occurrences of a term and a document but the PWI value corresponding to the co-occurrence event. As is easily seen, the cell values in the same row or column sum to the PWI of the corresponding term or document, respectively, and the total sum of all the cells represents the mutual information between terms and documents. Note that the above formulation maintains duality regarding documents and terms. In addition, the 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.

As far as we know, there has not been such an extension of tf–idf in the conventional information retrieval field. However, in computational linguistics, a few studies have dealt with measures similar to PWI. For example, the *weighted mutual information* proposed by Fung and McKeown (1996) follows the same definition as $\delta\mathcal{I}(w_i, d_j)$ in Eq. (16). In their experiments to extract translation pairs automatically from noisy parallel corpora, they showed that the proposed weighted mutual information outperformed the Dice coefficient and also mutual information.

4. Issues in selecting probabilistic models

4.1. Probabilistic models for tf–idf and its variations

One important implication of the formulation presented in the previous section is that the two probability distributions, $P(w_i)$ and $P(d_j|w_i)$, shown in Fig. 1, can be determined independently. In the figure, $P(w_i)$ represents the probability distribution of the query terms submitted to the system

(or the relevance of the query term to the query subject), while $P(d_j|w_i)$ is the conditional probability distribution of documents (or the posterior belief that the document is relevant to the subject), given the query term. In other words, it can be considered that $P(w_i)$ serves as a model of the user and $P(d_j|w_i)$ as a model of the retrieved documents.

Comparing Eq. (13) with Eq. (19), we can assume that the following estimation is used in the classical definition of tf–idf:

[*P*-tfidf estimation]

$$\widehat{P}(d_j|w_i) = \frac{1}{N_i}, \quad \widehat{P}(w_i) = \sum_{d_j} \frac{f_{ij}}{F}. \tag{20}$$

We refer to such an estimation as *P*-tfidf in the following. As we have already seen in Eqs. (14) and (15), the values of $P(w_i, d_j)$ and $P(d_j)$ are calculated using the above estimation as $\widehat{P}(w_i, d_j) \approx f_{ij}/F$ and $\widehat{P}(d_j) \approx 1/N$, respectively. Then, using Eqs. (16) and (17), tf–idf for term weighting becomes

$$\delta\widehat{\mathscr{I}}_{\text{tfidf}}(w_i, d_j) = \widehat{P}(w_i, d_j) \log \frac{\widehat{P}(d_j|w_i)}{\widehat{P}(d_j)} = \frac{f_{ij}}{F} \log \frac{N}{N_i}, \tag{21}$$

and tf–idf for term selection becomes

$$\delta\widehat{\mathscr{I}}_{\text{tfidf}}(w_i, \mathscr{D}) = \widehat{P}(w_i) \sum_{d_j} \widehat{P}(d_j|w_i) \log \frac{\widehat{P}(d_j|w_i)}{\widehat{P}(d_j)} = \frac{f_{w_i}}{F} \log \frac{N}{N_i}. \tag{22}$$

Now, the existing variations of tf–idf adopt either (i) non-linear scaling of the tf factor, (ii) adjustment of the idf values specifically for low frequency terms, or (iii) both (i) and (ii). The PWI formulation allows us to interpret these heuristic variations as variations of probabilistic estimation methods.

4.1.1. Variations of idf: the estimation of $P(d_j|w_i)$

The self-entropy reaches a maximum when all the composing events have equal probabilities. Therefore, in the information-theoretic view, tf–idf can be interpreted as employing a strategy in its estimation of $P(d_j|w_i)$ that maximizes the entropy of $\mathscr{D}$ (uncertainty about the documents) under the restriction that only the N_i documents with w_i have non-zero probabilities. On the other hand, another extreme case also exists in which the observed frequency is directly used as the estimation for the real probability of w_i:

[*P*-exact estimation]

$$\widehat{P}(d_j|w_i) = \frac{f_{ij}}{f_{w_i}}, \quad \widehat{P}(w_i) = \sum_{d_j} \frac{f_{ij}}{F}. \tag{23}$$

In this case, it is true without any specific assumptions that $\widehat{P}(w_i, d_j) = f_{ij}/F$ and $\widehat{P}(d_j) = f_{d_j}/F$. Then, using Eqs. (16) and (17), the PWI for term weighting becomes

$$\delta\widehat{\mathscr{I}}_{\text{exact}}(w_i, d_j) = \frac{f_{ij}}{F} \log \frac{F f_{ij}}{f_{w_i} f_{d_j}}, \tag{24}$$

and the PWI for term selection becomes

$$\delta\widehat{\mathscr{I}}_{\text{exact}}(w_i, \mathscr{D}) = \frac{f_{w_i}}{F} \sum_j \frac{f_{ij}}{f_{w_i}} \log \frac{Ff_{ij}}{f_{w_i} f_{d_j}}. \tag{25}$$

As *P*-tfidf and *P*-exact represent two extreme cases, the optimal allocation, if any, should be somewhere in the middle. Further, the two quantities match when the following conditions are satisfied:

(C1) $$\frac{f_{ij}}{f_{w_i}} \approx \frac{1}{N_i},$$

(C2) $$\frac{f_{d_j}}{F} \approx \frac{1}{N}.$$

(C2) comes from the implicit assumptions of Eq. (14). (C1) indicates that the occurrence of a term does not differ much across the documents. (C2) means that all the documents have almost equal sizes. For example, these conditions naturally hold when the document set under consideration is a collection of relatively short articles. In addition, (C1) is automatically satisfied when f_{ij} is given as a Boolean value, i.e., either 1 (occurs) or 0 (does not occur).

In conclusion, we can expect that the PWI values calculated using *P*-tfidf and *P*-exact are highly correlated for a relatively homogeneous document set in which (C1) and (C2) are satisfied, when idf provides a simple but robust estimate of information. On the other hand, the two values may differ greatly for a data set composed of heterogeneous documents, when a more specific probability estimation method (for example, refer to Baayen, 2001) is required. These assumptions are confirmed by experiments using an actual corpus in the next section.

4.1.2. Non-linear scaling of tf: the estimation of $P(w_i)$

It is widely recognized that linear scaling in term frequency with the classical definition of tf–idf lays too much stress on high frequency terms. Consequently, many non-linear scaling methods exist; for example, where the tf factors are proportional to the square root of term frequency or to the log of the frequency. Although these variations result in considerable difference in their calculated results, selection of a method seems to be left to the implementer in current retrieval systems.

The PWI formulation suggests that the tf factor in term weighting is given by the product of $P(d_j|w_i)$ and $P(w_i)$, the posterior probability of a document given a query term and the relevance of the query term in the context of the retrieval task. This interpretation is somewhat contradictory to the traditional view that considers the tf factor as intra-document characterization and the idf factor as inter-document characterization (for example, summarized in Baeza-Yates & Ribeiro-Neto, 1988). However, our interpretation seems to provide tf–idf variations with more flexibility in that 'terms that have appeared in the retrieved documents' and 'terms that will be used in queries' do not necessarily follow the same distribution. At the same time, this brings us another difficulty in scaling the tf factor, as we now require some prior expectation about the queries submitted to retrieval systems. The issue is beyond the scope of this paper and further investigation will be required to analyse the query statistics and also to establish how the intention of the user is reflected in the submitted query terms.

4.2. Probabilistic models to calculate the PWI of term sequences

So far, we have considered only a single term as a submitted query. The following discussion will show two different formulations to calculate the PWI values of documents associated with a sequence of query terms.

Let $w^* = w_{i_1}, \ldots, w_{i_k}$ be a sequence of k terms given to a system. The objective is to identify documents related to w^*. 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 q_i. The strategy for evaluating the relevance of document $d_j (\in D)$ is to calculate the PWI value of d_j, given w^* (Fig. 3).

For selecting w^*, two different formulations are considered:

(F1) In the first case, it is assumed that the k terms are selected from some unknown distribution.
(F2) In the second case, it is assumed that the k terms follow the same distribution as one of the documents in D.

In case (F1), the query terms are generated independently of the existing documents. The objective of the retrieval task is then to find the document closest to the submitted query terms. In case (F2), on the other hand, it is assumed that the query terms originate in one of the existing documents. The objective of the retrieval task is now to identify the document from which the submitted query terms are most likely to have come. Of course, in actual applications, these two formulations cannot explicitly be distinguished. Nevertheless, our formulation shows that they require different mathematical treatments.

In case (F1), the occurrences of the k terms in w^* are mutually independent. As before, let us assume that $P(w_i)$, the probability that the query term is relevant, and $P(d_j|w_i)$, the probability of

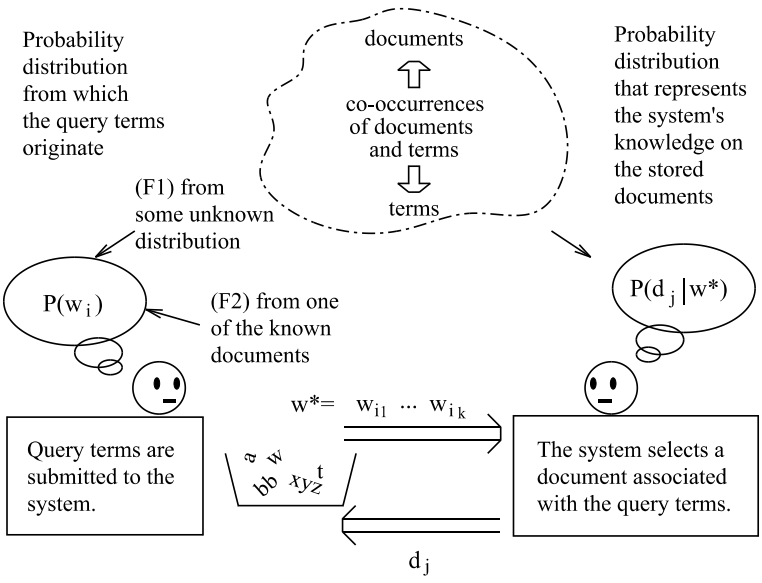


Fig. 3. An illustrative situation assumed in document retrieval or test categorization tasks.

each document conditioned by w_i, are given. Then, $P(d_j)$ is immediately calculated as $\sum_{w_i} P(d_j|w_i)P(w_i)$. Noting the additivity property of the amount of information for independent events, we simply use the summation of the PWI values of d_j given $w_i \in w^*$ to measure the relevance of d_j:

$$\delta\mathscr{I}(w^*, d_j) = \sum_{w_{i_1},\ldots,w_{i_k}} \delta\mathscr{I}(w_i, d_j) = \sum_{w_i \in w^+} q_i P(w_i) P(d_j|w_i) \log \frac{P(d_j|w_i)}{P(d_j)}. \tag{26}$$

In the above equation, $P(w_i)$ can be omitted if we assume that all the terms have equal probabilities in the originating unknown distributions. Eq. (26) is closely related to the vector-space model where the normalized inner product of term vectors weighted by tf–idf is used as the similarity measure. Assuming *P*-exact estimation where $P(d_j|w_i) \propto f_{ij}$, both methods entail the same form of "summation of $q_i f_{ij} \log(\cdot)$ with some normalization". The difference is in their normalization of document sizes and also in their consideration of the amount of information in the log terms.

On the other hand, in case (F2), the k terms are selected from some existing document $d_j \in D$. This time, we assume that $P(w_i|d_j)$, the probability of w_i conditioned by d_j, is given. Then, $P(w^*|d_j)$ is calculated as $P(w^*|d_j) = \prod_{w_i \in w^+} P(w_i|d_j)^{q_i}$, and $P(w^*)$ is immediately determined by $P(w^*) = \sum_{d_j \in D} P(w^*|d_j)P(d_j)$. Now, the PWI value between w^* and d_j is calculated as

$$\delta\mathscr{I}(w^*, d_j) = P(d_j)P(w^*|d_j) \log \frac{P(w^*|d_j)}{P(w^*)} = P(d_j) \prod_{w_i \in w^+} P(w_i|d_j)^{q_i} \log \frac{P(w^*|d_j)}{P(w^*)}. \tag{27}$$

Unlike the case in Eq. (26), the k terms, being assumed to come from the same distribution, are not independent of each other, and Eq. (27) cannot be simplified further. Although we have treated d_j as an individual document in the above equation, d_j can be any collection of terms originating from the same distribution, provided that examples of the distribution include a subset of relevant documents, or documents from the same category. In the last case, Eq. (27) has a clear correspondence with the naive Bayesian method popularly used in conventional text categorization studies. Because the Bayesian method selects the document (in this case, a category) with the largest probability $P(w^*|d_j)P(d_j)$ based on the maximum likelihood principle, the only difference between the two methods is the consideration of the amount of information expressed as the log term in Eq. (27), which is usually negligible.

Although existing retrieval theories adopt more complex forms of calculation, the difference between (F1) and (F2) is somewhat similar to the difference between the existing vector and probabilistic views of information retrieval. In our formulation, (F1) and (F2) are based on different standpoints: (F1) uses $P(w_i)$ and $P(d_j|w_i)$ as primary distributions while (F2) assumes $P(d_j)$ and $P(w_i|d_j)$ are given. Such an interpretation matches well with the results of Amati and Van Rijsbergen (1998), where the vector-space model considers terms as a basic event space and calculates the conditional expectation of each document given query terms, while the probabilistic model considers documents as a basic event space and calculates the conditional expectation of each term given its relevance. Note that if $P(w_i)$ and $P(d_j|w_i)$ are known, $P(d_j)$ and $P(w_i|d_j)$ are uniquely determined from Bayes' theorem. The reverse is also true. This enables the comparison of (F1) and (F2) under the same probabilistic assumptions.

5. Experiments and results

5.1. Objective of the experiments

In the previous section, we have extended the notion of tf–idf and defined a general measure, which we call in this paper the PWI. Although tf–idf has been widely recognized as a within-document term weighting scheme in conventional studies, our theoretical results suggest that the proposed PWI measure can be readily used as a measure for representative terms selection and also as a criterion for automatic text categorization. Moreover, these two applications of PWI are specifically good fits for our purpose because we can reasonably assume that the distribution of the extracted or submitted terms is similar to one of the indexed documents and thus can be easily estimated. In the following, we focus on these aspects of PWI and show some illustrative examples of the calculation results.

5.2. PWI in representative terms selection

In our first experiment, we compare the PWI values calculated using the two probability estimation methods: *P*-tfidf and *P*-exact given by Eqs. (20) and (23). The purpose of the experiment is to verify the assumption that tf–idf actually is a special case of PWI when the *P*-tfidf estimation is employed, and that tf–idf provides a simple but robust method to calculate the PWI value for a relatively uniform document set. In the experiment, the following two data sets were extracted from NTCIR-J1 (NACSIS, 1999):

(D1) 2106 abstracts of academic conference papers registered by the Japanese Society of Artificial Intelligence (JSAI), and
(D2) 24 groups of abstracts of academic conference papers, in total 309 999, with each group corresponding to a different academic society.

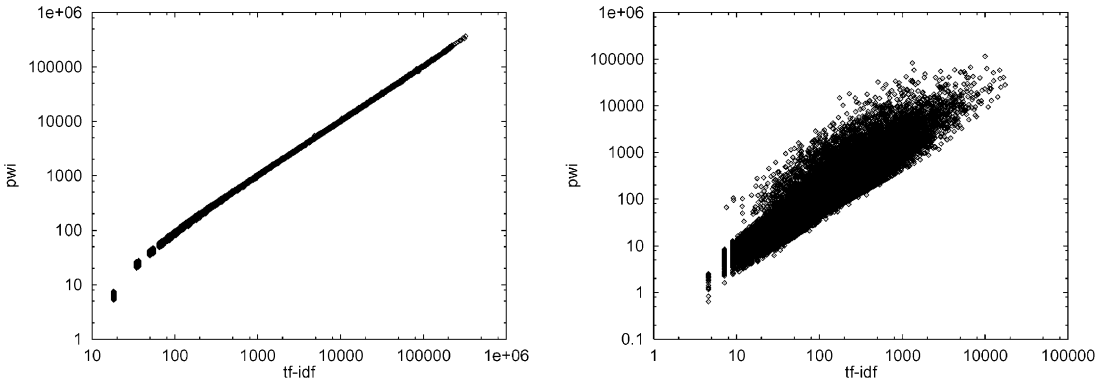
Abstracts were pre-processed by a Japanese morphological analyser ChaSen Ver. 2.02 (Matsumoto et al., 1999) to extract index terms.

For data set (D1), we treated each abstract as a separate document. The average size of a document was 69.5 words with the standard deviation being 24.4 words, which indicates that conditions (C1) and (C2) were satisfied in this case. For data set (D2), a group of abstracts presented at the same academic society was considered to be a single 'document', that is, a collection of terms originating from the distribution uniquely determined by the society. In this case, the size variation between 'documents' was extremely large: while the largest document contained about 26% of the total terms, the smallest one contained only 0.6% of the total. This implies that the conditions (C1) and (C2) no longer hold. (D2) is also used later in automatic text categorization tasks.

Table 2 compares the correlation coefficients between the tf–idf values and the PWI values calculated using *P*-exact (denoted simply as pwi). The correlations between the calculation results are also shown in Fig. 4, where the X and Y axes represent the values of tf–idf and pwi, respectively. Based on this result, we can confirm that these two values are almost identical for (D1), but differ considerably for (D2).

Table 2
Correlations between tf–idf and pwi values

Document set	Standard deviation of f_{d_j} values	Average of f_{ij} deviations	tf–idf and pwi correaltion
(D1): Homogeneous	0.37	0.18	1.00
(D2): Heterogeneous	1.44	6.27	0.52



(a) Calculation results for $D1$. (b) Calculation results for $D2$.

Fig. 4. Results of tf–idf and pwi calculations: (a) for (D1) and (b) for (D2).

In our next experiment, we investigated the advantage of the proposed PWI by applying these different measures to an automatic term extraction task. The goal of the extraction task was to identify the terminology related to the subject of artificial intelligence, given the same NTCIR data (D1) and (D2).

The result is shown in Fig. 5. The X axis is the top N ranking of the automatically extracted terms using either (i) the PWI with P-exact, (ii) information gain that is commonly used for term selection in text categorization studies, or (iii) tf–idf with tf log scaling, as a measure for selection.

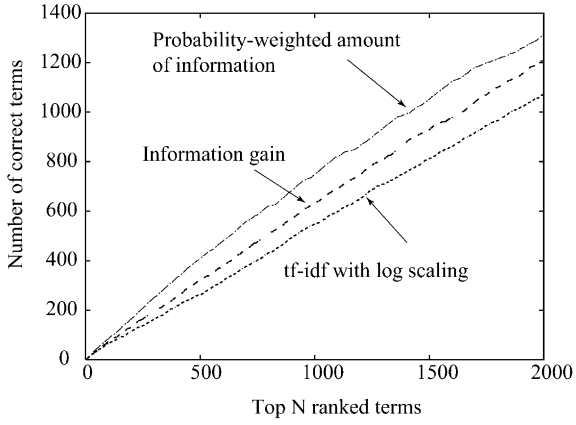


Fig. 5. Comparison of TMREC performance.

The Y axis is the number of 'correctly' extracted terms, where the 'correctness' was judged using the reference terminology set manually extracted from the same data set by the TMREC organizers (Kageura et al., 1999).

As can be seen in the figure, the PWI outperformed the other two methods and was quite effective in selecting representative terms. The reasons may be as follows. First, the original tf–idf definition basically works with the individual document bases and is unable to exploit information of non-relevant societies. The PWI extension makes it possible to apply the notion of tf–idf to binary groups of documents. Second, information gain sometimes selected terms that were particularly rare in the artificial intelligence field, while the objective of the extraction task was to select 'representative' terms rather than 'discriminating' ones. Although not being directly connected to tf–idf, the same definition as PWI is sometimes used for term selection in text categorization studies (for example, Koller & Sahami, 1997; Slonim & Tishby, 2000).

5.3. PWI in automatic text categorization

In our second experiment, we applied the proposed PWI calculation to a text categorization problem extracted from NTCIR-J1. The purpose of the experiment was to show that the vector-space-oriented calculation given by (F1) and the probabilistic one given by (F2) can be compared under the same probabilistic assumption within the unified framework of the PWI.

In the experiments, the data set (D2) from the first experiment was used as training data, where the size was varied as 1000, 10 000, 50 000 and the maximum 309 999. The 24 societies were treated as distinct categories, each of which was expressed as a single collection of terms rather than a collection of independent documents. As each abstract belonged to exactly one society, the categorization task was formulated as a multi-class problem. In the evaluation, a total of 10 000 abstracts were prepared that were not contained in the training data, but with the same distribution across categories. Therefore, if about 25% of the abstracts of the training data belonged to society A, then the test data also contained about 25% abstracts from society A. The performance was compared using the ratio of the correct judgements, i.e., the number of abstracts classified into the class to which they originally belonged, divided by 10 000.

The following categorization methods were compared in the experiments: (i) tfidf-cos, the cosine similarity between the submitted document and the target class term vectors with tf–idf weighting; (ii) pwi-vec, the PWI with (F1) formulation given by Eq. (26); and (iii) nbayes, the conventional naive Bayesian method. Note that no meaningful difference existed between the PWI with (F2) formulation given by Eq. (27) and nbayes when the data size was sufficiently large. For each of the categorization methods, the following three probability estimation methods, P-freq, P-laplace and P-mixture, were tested.

The first model was chosen for tfidf-cos and pwi-vec, and is referred to as P-freq. Denoting the frequency of w_i in category c_j as f_{ij} and the total frequency of w_i for all the categories as f_{w_i}, P-freq is given by

[P-freq estimation]

$$\widehat{P}(c_j|w_i) = \frac{f_{ij}}{f_{w_i}}, \quad \widehat{P}(w_i) = \frac{1}{M}. \tag{28}$$

The second model was chosen for nbayes, and is referred to as *P*-laplace. Denoting the total frequency of terms in category c_j as f_{c_j}, *P*-laplace is given by

[*P*-laplace estimation]

$$\widehat{P}(w_i|c_j) = \frac{1 + f_{ij}}{M + f_{c_j}}, \quad \widehat{P}(c_j) = \frac{f_{c_j}}{F}. \tag{29}$$

Note that $\widehat{P}(w_i|c_j) \neq 0$ even when $f_{ij} = 0$. Such consideration of unobserved events is crucial with the nbayes method, as otherwise the probability $\widehat{P}(w^*, c_j)$ automatically becomes zero for all categories if w^* contains only a single unknown term. *P*-laplace, known as the Laplace estimator, provides a simple way to deal with the zero frequency problem and is often used in conventional naive Bayesian approaches in the text categorization field. The third model, referred to as *P*-mixture, is expressed as the mixture distribution of $\widehat{P}(c_j)$ in Eq. (29) and $\widehat{P}(c_j|w_i)$ in Eq. (28):

[*P*-mixture estimation]

$$\widehat{P}(c_j|w_i) = (1 - r_i)\frac{f_{c_j}}{F} + r_i\frac{f_{ij}}{f_{w_i}}, \quad \widehat{P}(w_i) = \frac{1}{M}. \tag{30}$$

The mixture ratio r_i is determined as $r_i = (f_{w_i} - \delta)/f_{w_i}$ where δ is a discounting coefficient common for all the terms. Using the formula for absolute discounting in the probabilistic language modelling theory (for example, Kita, 1999), the value is determined as $\delta =$ (number of singletons)/(total frequency). This model is motivated by recent studies in the text categorization field that have shown that the performance of naive Bayesian categorization is sensitive to the estimation of $P(w_i|c_j)$ (McCallum & Nigam, 1998). It has been widely recognized in probabilistic language model studies that, despite the convenience of the Laplace estimator, it does not provide a good fit compared with other dedicated discounting methods. We have observed that the newly introduced *P*-mixture model specifically provides a good fit with our corpus.

The results are summarized in Table 3. Comparing the probability estimation methods, we can confirm that the *P*-freq estimation works reasonably well with tfidf-cos and pwi-vec, while the estimation is not applicable to nbayes, because of the zero frequency problem we have already mentioned. *P*-laplace showed a specifically good fit to nbayes, but does not work well for pwi-vec. *P*-mixture seems to be consistently good for all of the categorization methods. Comparing the categorization methods, we can observe that the performance of tfidf-cos is degraded as the size of the training data becomes large, while better performance is observed for larger sizes of the training data with pwi-vec and nbayes. The reason may be that the tfidf-cos method requires careful adjustment of the scaling parameter to adapt to different sizes of data. The performance of

Table 3
Results of text categorization experiments

	$\|D\| = 1000$			$\|D\| = 10000$			$\|D\| = 50000$			$\|D\| = 309999$		
	Freq	Laplace	Mixture	Freq	Laplace	Mixture	Freq	Laplace	Mixture	Freq	Laplace	Mixture
tfidf-cos	0.6292	0.4885	0.6246	0.6837	0.6758	0.6808	0.6539	0.6542	0.6526	0.6472	0.6480	0.6466
nbayes	0.1718	0.5967	0.6520	0.2660	0.6833	0.7752	0.3655	0.7244	0.7767	0.4154	0.7583	0.7884
pwi-vec	0.6398	0.5628	0.6454	0.7778	0.6172	0.7711	0.7920	0.6633	0.7855	0.8149	0.7146	0.8084

pwi-vec is slightly better than that of nbayes, possibly because of the sensitivity of the latter to errors in probability estimation.

Based on the above observations, we conclude that categorization performance is considerably influenced by the strategies used for probability estimation, and these two should be considered in combination. The proposed pwi-vec method seems to be promising because the method avoids the zero frequency problem of the naive Bayes method while maintaining a well-formulated probabilistic background. By applying an information-theoretic view, the method considers the occurrence of unobserved terms simply as non-informative. The performance of the pwi-vec method is further studied in Aizawa (2001) in comparison with the performance of support vector machine.

6. Conclusion

In this paper, we have investigated an information-theoretic interpretation of tf–idf, and provided a view of tf–idf as the amount of information of a term weighted by its occurrence probability. Such a perspective enables us to extend the notion of tf–idf into a more general formula of the PWI. By calculating the PWI values of a sequence of query terms under different probability assumptions, we have shown that the vector-space-oriented view of the original tf–idf can successfully be related to probability-oriented views. An illustrative example was shown in which the proposed calculation was applied to an actual text collection and the correlations between the tf–idf and the PWI values were calculated. The effect of different probability estimation methods on a real-scale text categorization problem was also examined in the experiment.

Although our investigation in this paper mainly concerns the consistency of the proposed PWI with, and not its superiority to, conventional statistical measures, we believe that such an approach is worthwhile. This approach not only leads us to better understandings of the commonly practiced heuristic measures, but also enables us to propose and verify different heuristics, in connection with other research fields such as probabilistic language modelling.

Acknowledgements

The author would like to thank Kyo Kageura, Atsushi Takasu and other colleagues at the National Institute of Informatics (NII) for their helpful discussions. The author is also grateful to Noriko Kando at NII for encouraging and assisting us in this study as one of the organizers of NTCIR.

References

Aizawa, A. (2000). The feature quantity: an information-theoretic perspective of tfidf-like measures. In *Proceedings of the 23rd ACM SIGIR conference on research and development in information retrieval* (pp. 104–111).

Aizawa, A. (2001). Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings of the sixth natural language processing Pacific rim symposium (NLPRS 2001)* (pp. 307–314).

Amati, G., & Van Rijsbergen, K. (1998). Semantic information retrieval. In F. Crestani, M. Lalmas, & C. J. Van Rijsbergen (Eds.), *Information retrieval: uncertainty and logics* (pp. 189–219). Boston: Kluwer Academic Press.

Baayen, R. H. (2001). *Word frequency distribution*. Boston: Kluwer Academic Publishers.
Baeza-Yates, R., & Ribeiro-Neto, B. (1988). *Modern information retrieval*. New York: Addison-Wesley.
Brookes, B. C. (1972). The Shannon model of IR systems. *Journal of Documentation, 28*, 160–162.
Church, K. W., & Gale, W. (1999). Inverse document frequency (IDF): a measure of deviations from Poisson. In *Natural language processing using very large corpora* (pp. 283–295). Boston: Kluwer Academic Press.
Church, K. W., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics, 6*(1), 22–29.
Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley and Sons Inc.
Crestani, F. (2000). Exploiting the similarity of non-matching terms at retrieval time. *Journal of Information Retrieval, 2*(1), 23–43.
Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285–295.
Dennis, S. F. (1964). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation, symposium proceedings (Miscellaneous publication 269)*. Washington, DC: National Bureau of Standards.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.
Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information Processing and Management, 25*(1), 55–72.
Fung, P., & McKeown, K. (1996). A technical word and term translation aid using noisy parallel corpora across language groups. *The Machine Translation Journal, 12*(1–2), 53–87.
Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston: Kluwer Academic Publishers.
Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In *Proceedings of the 21st international conference on research and development in information retrieval (SIGIR'98)* (pp. 11–18).
Hiemstra, D. (2000). A probabilistic justification for using tf × idf term weighting in information retrieval. *International Journal on Digital Libraries, 3*(2), 131–139.
Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the 14th international conference on machine learning (ICML'97)* (pp. 143–151).
Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology, 3*(2), 259–289.
Kageura, K., Yoshioka, M., Tsujii, K., Yoshikane, F., Takeuchi, K., & Koyama, T. (1999). Evaluation of the term recognition task. In *Proceedings of the 1st NTCIR workshop on research in Japanese text retrieval and term recognition* (pp. 42–49).
Kita, K. (1999). *Probabilistic language models*. Tokyo: University of Tokyo Press.
Koller, D., & Sahami, M. (1996). Toward optimal feature selection. In *Proceedings of the international conference on machine learning (ICML'96)* (pp. 284–292).
Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the international conference on machine learning (ICML'97)* (pp. 170–178).
Lewis, D. D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the 3rd annual symposium on document analysis and information retrieval (SDAIR'94)* (pp. 81–93).
Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1*(4), 309–317.
Manning, C. D., & Schtüze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.
Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, K., & Asahara, M. (1999). *Morphological analysis system ChaSen 2.0.2 users manual. NAIST Technical Report, NAIST-IS-TR99012*. Nara Institute of Science and Technology.
McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive Bayes text classification. Learning for text categorization, Technical Report WS-98-05* (pp. 41–48). Menlo Park: The AAAI Press.
Mladenić, D. (1998). Feature subset selection in text-learning. In *Proceedings of the 10th European conference on machine learning (ECML'98)* (pp. 95–100).
Nagao, M., Mizutani, M., & Ikeda, H. (1976). An automated method for the extraction of important words from Japanese scientific documents. *Transactions of Information Processing Society of Japan, 17*(2), 110–117.

National Center for Science Information Systems (1999). *NTCIR workshop 1—Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*. http://research.nii.ac.jp/~ntcadm/workshop/OnlineProceedings/.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation, 46*(4), 359–364.

Robertson, S. E. (1994). Query-document symmetry and dual models. *Journal of Documentation, 50*(3), 233–238.

Robertson, S. E., & Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science, 27*, 129–146.

Salton, G., & Buckley, C. (1988). Weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd international conference on research and development in information retrieval (SIGIR 2000)* (pp. 208–215).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*(1), 143–178.

Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Van Rijsbergen, C. J., Happer, D. J., & Porter, M. F. (1981). The selection of good search terms. *Information Processing and Management, 17*, 77–91.

Wiener, E., Pedersen, J. O., & Weighend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the 4th annual symposium on document analysis and information retrieval (SDAIR'95)* (pp. 317–332).

Wong, S. K. M., & Yao, Y. Y. (1992). An information-theoretic measure of term specificity. *Journal of the American Society for Information Science, 43*(1), 54–61.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd international conference on research and development in information retrieval (SIGIR'99)* (pp. 42–49).

Yang, Y., & Pedersen, O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning (ICML'97)* (pp. 412–420).