# A Simple Method for Estimating Term Mutual Information

D. Cai and T.L. McCluskey

**Abstract**—The ability to formally analyze and automatically measure statistical dependence of terms is a core problem in many areas of science. One of the commonly used tools for this is the expected mutual information (MI) measure. However, it seems that MI methods have not achieved their potential. The main problem in using MI of terms is to obtain actual probability distributions estimated from training data, as the true distributions are invariably not known. This study focuses on the problem and proposes a novel but simple method for estimating probability distributions. Estimation functions are introduced; mathematical meaning of the functions is interpreted and the verification conditions are discussed. Examples are provided to illustrate the possibility of failure of applying the method if the verification conditions are not satisfied. An extension of the method is considered.

**Index Terms**—Information analysis and extraction, dependence and relatedness of terms, statistical semantic analysis.

## 1 Introduction

THE ability to formally analyze and automatically measure statistical dependence (relatedness, proximity, association, similarity) of terms in textual documents is a core problem in many areas of sciences, such as, feature extraction and selection, concept learning and clustering, document representation and query formulation, text analysis and data mining. Solution of the problem has been a technical barrier for a variety of practical mathematical applications. One of the commonly used tools of analysis and measurement is the expected mutual information (MI) measure drawn from information theory [10], [16]. Many studies have used the measure for a variety of tasks in, for instance, feature selection [2], [1], [11], [15], document classification [18], face image clustering [14], multimodality image registration [13], information retrieval [6], [7], [8], [9], [14]. However, it seems that MI methods have not achieved their potential. The main problem we face in using the expected MI measure is obtaining actual probability distributions, as the true distributions are invariably not known, and we have to estimate them from training data. This work explores techniques of estimation.

To address this study clearly, let us first introduce the concept of a term *state value* distribution. A term is usually thought of as having *states* "present" or "absent" in a document. Thus, for an arbitrary term $t$, it will be convenient to introduce a variable $\delta$ taking values from set $\Omega = \{1, 0\}$, where $\delta = 1$ expresses that $t$ is present and $\delta = 0$ expresses that $t$ is absent. Denote $t^\delta = t, \bar{t}$ when $\delta = 1, 0$, respectively. We call $\Omega$ a *state value space*, and each element in $\Omega$ a *state value*, of $t$. Similarly, for an arbitrary term pair $(t_i, t_j)$, we introduce a variable pair $(\delta_i, \delta_j)$ taking values from set $\Omega \times \Omega =$ {(1,1), (1,0), (0,1), (0,0)}. We call $\Omega \times \Omega$ a state value space, and each element in $\Omega \times \Omega$ a state value pair, of $(t_i, t_j)$.

Let $D$ be a *collection* of documents (training data), and $V$ a *vocabulary* of terms used to index individual documents in $D$. Denote $V_d \subseteq V$ as the set of terms occurring in document $d \in D$. Thus, for each term $t$ occurring in $d$, its state value distribution is

$$P_d(\delta) = P(t^\delta | d) \quad (\delta \in \Omega)$$

Obviously, each term $t \in V_d$ is matched to a state value distribution and there are totally $|V_d|$ state value distributions for document $d$.

There exists statistical dependence between two terms, $t_i$ and $t_j$, if the state value of one of them provides mutual information about the probability of the state value of another. Losee [12] showed that there is a relationship between the frequencies (or probabilities) of terms and MI of terms. Therefore, term $t_i$ taking some state value $\delta_i$ (say $\delta_i = 1$) should be looked upon as complex because another state value (say $\delta_i = 0$) of $t_i$, and state values of many other terms (i.e., all terms $t_j \in V_d - \{t_i\}$), may be dependent on this $\delta_i$.

Mathematically, for two arbitrary terms $t_i, t_j \in V_d$, the *expected mutual information* [10] about the probabilities of the state value pair $(\delta_i, \delta_j)$ of term pair $(t_i, t_j)$ can be expressed by:

$$I_d(\delta_i, \delta_j) = H(\delta_i) - H(\delta_i | \delta_j) = H(\delta_j) - H(\delta_j | \delta_i)$$
$$= \sum_{\delta_i, \delta_j = 1,0} P_d(\delta_i, \delta_j) \log \frac{P_d(\delta_i, \delta_j)}{P_d(\delta_i) P_d(\delta_j)}$$

- *D. Cai is with the School of Computing and Engineering, University of Huddersfield, UK, HD1 3BE.*
- *T.L. McCluskey is with the School of Computing and Engineering, University of Huddersfield, UK, HD1 3BE.*

where $H(\delta_i)$ is entropy of $\delta_i$, measuring uncertainty on $\delta_i$; $H(\delta_i|\delta_j)$ is conditional entropy of $\delta_i$, measuring uncertainty on $\delta_i$ given knowing $\delta_j$. Thus, $I_d(\delta_i, \delta_j)$ measures the amount of information that $\delta_j$ provides about $\delta_i$, and *vice versa*.

The estimation of distributions $P_d(\delta)$ and $P_d(\delta_i, \delta_j)$ required in $I_d(\delta_i, \delta_j)$ is crucial and remains an open issue for effectively distinguishing potentially dependent term pairs from many others and, therefore, the main concern of our current study. In Section 2, we introduce estimation functions, interpret mathematical meaning of the functions and discuss verification conditions. In Section 3, we provide examples to clarify the idea of our method. Section 4 considers an extension of our method and conclusions are drawn in Section 5. Some mathematical details are given in the Appendix.

## 2 ESTIMATION

In practical application, the state value distributions may be estimated from training data. The estimation of the joint state value distribution, $P_d(\delta_i, \delta_j)$, is a more complicated task, which is thus the main concern of this section.

Let us start with considering a given document. Say we have a document $d \in D$ with $V_d = \{t_{i_1}, t_{i_2}, \dots, t_{i_s}\} \subseteq \{t_1, t_2, \dots, t_n\} = V$, where $1 \le i_1 < i_2 < \cdots < i_s \le n$. In this study, we always assume that $2 < s = |V_d| \le n$ (namely, each document has at least three distinct terms).

Generally, if we denote $f_d(t)$ as the frequency of term $t$ in $d$ and $\|d\|$ as the *length* of $d$ then, for a given document $d$, the term *occurrence frequency* distribution is given by

$$p_d(t) = p(t|d) = \frac{f_d(t)}{\sum_{t' \in V_d} f_d(t')} = \frac{f_d(t)}{\|d\|} \quad (t \in V_d)$$

which should not be confused with the term *state value* distribution $P_d(\delta)$.

In order to constitute the state value distributions, for arbitrary terms $t, t_i, t_j \in V_d$, let we introduce two *estimation functions*:

$$\psi_d(t) = \frac{f_d(t)}{\sum_{t' \in V_d} f_d(t')} \quad (1)$$

$$\gamma_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\sum_{i' < j';\, t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'})} \quad (2)$$

Clearly, $0 < \psi_d(t), \gamma_d(t_i, t_j) < 1$ for arbitrary $t, t_i, t_j \in V_d$.

Then, for each term $t \in V_d$, from the function $\psi_d(t)$, define $P_d(\delta)$ by

$$P_d(\delta = 1) = P_d(t) = \psi_d(t)$$
$$P_d(\delta = 0) = P_d(\bar{t}) = 1 - \psi_d(t) \quad (3)$$

which is a probability distribution over $\Omega$.

To constitute $P_d(\delta_i, \delta_j)$ from the function $\gamma_d(t_i, t_j)$, for given terms $t_i, t_j \in V_d$, define

$$\varphi_d(\delta_i = 1, \delta_j = 1) = \gamma_d(t_i, t_j)$$
$$\varphi_d(\delta_i = 1, \delta_j = 0) = \psi_d(t_i) - \gamma_d(t_i, t_j)$$
$$\varphi_d(\delta_i = 0, \delta_j = 1) = \psi_d(t_j) - \gamma_d(t_i, t_j)$$
$$\varphi_d(\delta_i = 0, \delta_j = 0) = 1 - \psi_d(t_i) - \psi_d(t_j) + \gamma_d(t_i, t_j)$$

Note that $\varphi_d(\delta_i, \delta_j)$ may not constitute a probability distribution.

Next, we need to prove that, under some conditions, $\varphi_d(\delta_i, \delta_j)$ can be a probability distribution by Theorem 1 below. For doing so, here, and throughout this study, we denote the denominator of $\gamma_d(t_i, t_j)$ by

$$\varpi = \sum_{i' < j';\, t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'})$$

and, for an arbitrary term $t \in V_d$, denote

$$\varpi_t = \sum_{i' < j';\, t_{i'}, t_{j'} \in V_d - \{t\}} f_d(t_{i'}) f_d(t_{j'})$$

Clearly $\varpi > \varpi_t \ge 1$ as $|V_d| > 2$.

To prove Theorem 1, we need to introduce two lemmas. Detailed proofs are given in the Appendix.

**Lemma 1.** *For an arbitrary term $t \in V_d$, we have*

$$\varpi = \|d\| f_d(t) - f_d^2(t) + \varpi_t$$

**Lemma 2.** *For the functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ given in (1) and (2), respectively, we have:*

*(a) $\varpi_{t_i} \ge f_d^2(t_i)$ if and only if $\psi_d(t_j) \ge \gamma_d(t_i, t_j)$;*

*(b) $\varpi_{t_j} \ge f_d^2(t_j)$ if and only if $\psi_d(t_i) \ge \gamma_d(t_i, t_j)$.*

We are now ready to introduce Theorem 1 below. Detailed proof is given in the Appendix.

**Theorem 1.** *For arbitrary terms $t_i, t_j \in V_d$, expression*

$$P_d(\delta_i, \delta_j) = \varphi_d(\delta_i, \delta_j) \quad (4)$$

*is a probability distribution* over $\Omega \times \Omega$ *if it satisfies two inequalities: a) $\varpi_{t_i} \ge f_d^2(t_i)$ and b) $\varpi_{t_j} \ge f_d^2(t_j)$.*

Thus, by the above expression $\varphi_d(\delta_i, \delta_j)$ and Theorem 1, we have, for instance,

$$P_d(\delta_i = 1, \delta_j = 1) = P_d(t_i, t_j) = \gamma_d(t_i, t_j)$$
$$P_d(\delta_i = 1, \delta_j = 0) = P_d(t_i, \bar{t}_j) = \psi_d(t_i) - \gamma_d(t_i, t_j)$$

The first lemma tells us that there exists a relationship between $\varpi$ and $\varpi_t$. The second lemma tells us how to

verify two inequalities (conditions) required in Theorem 1: $\varpi_{t_i} \geq f_d^2(t_i)$ and $\varpi_{t_j} \geq f_d^2(t_j)$ may be simply verified by $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$ and $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$, respectively.

The reasoning behind the estimate, $P(\delta_i, \delta_j)$, is rather intuitive. $P(\delta_i = 1, \delta_j = 0)$ and $P(\delta_i = 0, \delta_j = 1)$ are derived by two constraints:

$$P(\delta_i = 1, \delta_j = 0) + P(\delta_i = 1, \delta_j = 1) = P(\delta_i = 1);$$
$$P(\delta_i = 0, \delta_j = 1) + P(\delta_i = 1, \delta_j = 1) = P(\delta_j = 1);$$

which ensure that both $P(\delta_i)$ and $P(\delta_j)$ are marginal distributions of $P(\delta_i, \delta_j)$. $P(\delta_i = 0, \delta_j = 0)$ is derived by another constraint

$$\sum_{\delta_i, \delta_j = 1,0} P_d(\delta_i, \delta_j) = 1$$

It is worth explaining the derivation of $P(\delta_i = 1, \delta_j = 1) = \gamma(t_i, t_j)$ in more detail. In practice, the estimation functions should be considered carefully and introduced meaningfully according to a specific application problem. Let us now explain the meaning of $\gamma(t_i, t_j)$ given in (1).

It may be easier to make the explanation through an $n \times n$ matrix. Suppose we are given a document $d$ represented by a (frequency) $1 \times n$ matrix

$$\boldsymbol{m}_d = [f_d(t_1), f_d(t_2), \dots, f_d(t_n)] = [f_d(t)]_{1\times n}$$

in which, each element is a frequency satisfying $f_d(t) \geq 1$ when $t \in V_d$ and $f_d(t) = 0$ when $t \in V - V_d$. The matrix product can be written by

$$\begin{aligned}
\boldsymbol{m}_d' \times \boldsymbol{m}_d &= \begin{bmatrix} f_d(t_1) \\ \dots \\ f_d(t_n) \end{bmatrix} \times [f_d(t_1) \quad \dots \quad f_d(t_n)] \\
&= \begin{bmatrix} f_d(t_1)f_d(t_1) & \dots & f_d(t_1)f_d(t_n) \\ \dots & \dots & \dots \\ f_d(t_n)f_d(t_1) & \dots & f_d(t_n)f_d(t_n) \end{bmatrix} \\
&= \left[f_d(t_i)f_d(t_j)\right]_{n\times n} \\
&= \varpi \left[\frac{1}{\varpi} f_d(t_i)f_d(t_j)\right]_{n\times n} \\
&= \varpi \left[P_d(\delta_i = 1, \delta_j = 1)\right]_{n\times n}
\end{aligned}$$

Generally, $\left[f_d(t_i)f_d(t_j)\right]_{n\times n}$, which is symmetric, is called the *co-occurrence frequency matrix* of terms concerning $d$. Hence,

$$\begin{aligned}
\left[P_d(\delta_i = 1, \delta_j = 1)\right]_{n\times n} &= \left[\frac{1}{\varpi} f_d(t_i)f_d(t_j)\right]_{n\times n} \\
&= \left[\gamma_d(t_i, t_j)\right]_{n\times n}
\end{aligned}$$

can be referred to as the *normalized* co-occurrence frequency matrix of terms concerning $d$. Consequently, $P_d(\delta_i = 1, \delta_j = 1) = \gamma_d(t_i, t_j)$, for $i, j = 1, \dots, n$, can be represented by an $n \times n$ matrix: its numerator, $f_d(t_i)f_d(t_j)$, characterizes the co-occurrence frequencies of $t_i$ and $t_j$ in document $d$; its denominator, $\varpi$, the sum of all possible numerators $f_d(t_i)f_d(t_j)$ for $i < j$; $i, j = 1, \dots, n$, is a *normalization factor* for the characterization. Clearly, $\varpi$ is a constant for all term pairs, $(t_i, t_j)$, occurring in a given document.

Note that assumption $|V_d| > 2$ ensures that there exists more than one non-zero element in the matrix, such that $\left[P_d(\delta_i = 1, \delta_j = 1)\right]_{n\times n} \neq [0]_{n\times n}$. Notice also that, because no two components of $(t_i, t_j)$ can be the same, the elements where $i = j$, corresponding to $P_d(\delta_i = 1, \delta_j = 1)$ for $i = 1, \dots, n$, should not be considered in our context. However, it is only for notational convenience that these elements are included in the matrix.

## 3 Discussion

It should be emphasized, in order to speak of the MI of terms, that we must verify two arguments of $I(\delta_i, \delta_i)$ are probability distributions. For instance, in our method, they should satisfy the two inequalities given in Theorem 1. Let us look at examples below, which will help to clarify the idea and make understandable the computation involved in all the above formulae.

**Example 1.** Suppose $d_1 = \{t_1, t_2, t_2, t_3, t_4, t_6, t_9, t_9\}$, then we have $V_{d_1} = \{t_1, t_2, t_3, t_4, t_6, t_9\}$, and $\varpi = 26$. Thus, for term pair $(t_1, t_2)$, from (2) and (4), we have

$$P_{d_1}(\delta_1 = 1, \delta_2 = 1) = \frac{1\times 2}{26} = \frac{8}{104}$$
$$P_{d_1}(\delta_1 = 1, \delta_2 = 0) = \frac{1}{8} - \frac{2}{26} = \frac{5}{104} > 0$$
$$P_{d_1}(\delta_1 = 0, \delta_2 = 1) = \frac{2}{8} - \frac{2}{26} = \frac{18}{104} > 0$$
$$P_{d_1}(\delta_1 = 0, \delta_2 = 0) = 1 - \frac{1}{8} - \frac{2}{8} + \frac{2}{26} = \frac{73}{104}$$

Then, it follows immediately (using natural logarithms)

$$\begin{aligned}
I_{d_1}(\delta_1, \delta_2) = \frac{8}{104}\log\frac{\frac{8}{104}}{\frac{1}{8}\cdot\frac{2}{8}} + \frac{5}{104}\log\frac{\frac{5}{104}}{\frac{1}{8}\cdot\frac{6}{8}} \\
+ \frac{18}{104}\log\frac{\frac{18}{104}}{\frac{7}{8}\cdot\frac{2}{8}} + \frac{73}{104}\log\frac{\frac{73}{104}}{\frac{7}{8}\cdot\frac{6}{8}} \\
\approx 0.0693 - 0.0321 - 0.0405 + 0.0472 \\
= 0.0439
\end{aligned}$$

**Example 2.** Suppose that we are given a document $d_2 = \{t_1, t_2, t_2, t_2, t_3, t_4\}$. From which we have

$$\begin{aligned}
\varpi &= \sum_{i'<j'; t_{i'}, t_{j'} \in V_{d_2}} f_{d_2}(t_{i'})\, f_{d_2}(t_{j'}) \\
&= f_{d_2}(t_1)f_{d_2}(t_2) + f_{d_2}(t_1)f_{d_2}(t_3) + f_{d_2}(t_1)f_{d_2}(t_4) + \\
&\quad f_{d_2}(t_2)f_{d_2}(t_3) + f_{d_2}(t_2)f_{d_2}(t_4) + f_{d_2}(t_3)f_{d_2}(t_4) \\
&= 1\times 3 + 1\times 1 + 1\times 1 + 3\times 1 + 3\times 1 + 1\times 1 = 12
\end{aligned}$$

Thus, for instance, for term pair $(t_1, t_2)$, we have $\gamma_{d_2}(t_1, t_2) = \frac{1\times 3}{12}$, and

$$\begin{aligned}
P_{d_2}(\delta_1 = 1, \delta_2 = 0) &= \psi_{d_2}(t_1) - \gamma_{d_2}(t_1, t_2) \\
&= 1/6 - 3/12 = -1/12 < 0
\end{aligned}$$

from which we can conclude that $P_{d_2}(\delta_i, \delta_j)$ is not a

probability distribution since $\varpi_{d_2}(t_1) - \gamma_{d_2}(t_1, t_2) < 0$. Also, we can verify this in an alternative way:

$$\begin{aligned}\varpi_{t_2} &= \textstyle\sum_{i'<j';\, t_{i'}, t_{j'} \in V_{d_2} - \{t_2\}} f_{d_2}(t_{i'}) f_{d_2}(t_{j'}) \\ &= f_{d_2}(t_1) f_{d_2}(t_3) + f_{d_2}(t_1) f_{d_2}(t_4) \\ &\quad + f_{d_2}(t_3) f_{d_2}(t_4) \\ &= 1 + 1 + 1 < 9 = f_{d_2}(t_2) f_{d_2}(t_2)\end{aligned}$$

That is, the first inequality given in Lemma 2 is not satisfied.

The above Example 2 is a specific instance of failing to apply the estimation given in (1)-(4). From the above two examples, we can see:

(i) In order to compute term dependence, we must verify both $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$ and $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$, or equivalently to verify both $\varpi_{t_i} \geq f_d^2(t_i)$ and $\varpi_{t_j} \geq f_d^2(t_j)$, satisfied simultaneously, for each term pair considered.

(ii) $\gamma_d(t_i, t_j)$ becomes smaller rapidly as documents become longer, it should thus not be a problem to satisfy the above two inequalities in practical application.

In a practical application, we generally concentrate on the statistics of co-occurrence of terms. That is, the dependence with which we are really concerned is state value $(\delta_i, \delta_j) = (1, 1)$ of term pair $(t_i, t_j)$. In this case, what we need is to apply only the first item of $I(\delta_i, \delta_j)$ and to verify the second condition given in Theorem 1:

$$\begin{aligned}P_d(\delta_i = 1, \delta_j = 1) &= \frac{f_d(t_i) f_d(t_j)}{\varpi} \\ &= \gamma_d(t_i, t_j) > \psi_d(t_i) \psi_d(t_j) \\ &= \frac{f_d(t_i)}{\|d\|} \cdot \frac{f_d(t_j)}{\|d\|}\end{aligned}$$

to ensure that $t_i$ and $t_j$ are highly dependent under their co-occurrence.

## 4 EXTENSION

The method proposed in this study may be applicable to any quantitative document representations. That is, the estimation functions given in (1) and (2) can be applied to document representations not only for the frequency matrix, but also for a more general case, where each matrix element is a real number.

More particularly, suppose each $d \in D$ can be expressed by a $1 \times n$ (weight) matrix

$$\boldsymbol{m}_d = [w_d(t_1), w_d(t_2), \ldots, w_d(t_n)] = [w_d(t)]_{1 \times n}$$

satisfying $w_d(t) > 0$ when $t \in V_d$ and $w_d(t) = 0$ when $t \in V - V_d$. The $w_d(t)$ is called a *weighting function*, indicating the importance of term $t$ in representing document $d$. For instance, a widely used weighting function would be $w_d(t) = f_d(t) \times \log \frac{|D|}{n_D(t)}$, where $n_D(t)$ is the number of documents in $D$ in which $t$ occurs. Also, the method described in previous sections is a special case where, $w_d(t) = f_d(t)$ for $t \in V_d$.

With document representation by $w_d(t)$, let us continue to denote the "length" of document $d$ by

$$\|d\| = \sum_{t' \in V_d} w_d(t')$$

and denote

$$\begin{aligned}\varpi_t &= \textstyle\sum_{i'<j';\, t_{i'}, t_{j'} \in V_d - \{t\}} w_d(t_{i'})\, w_d(t_{j'}) \\ &< \textstyle\sum_{i'<j';\, t_{i'}, t_{j'} \in V_d} w_d(t_{i'})\, w_d(t_{j'}) = \varpi\end{aligned}$$

Then, for arbitrary terms $t, t_i, t_j \in V_{\mathrm{d}}$, similar to the expressions given in (3) and (4), we may write the corresponding marginal distribution:

$$\begin{aligned}P_d(\delta = 1) &= \frac{w_d(t)}{\|d\|} = p_d(t) = \psi_d(t) \\ P_d(\delta = 0) &= 1 - \psi_d(t)\end{aligned} \qquad (5)$$

and joint distribution

$$\begin{aligned}P_d(\delta_i = 1, \delta_j = 1) &= \frac{w_d(t_i) w_d(t_j)}{\varpi} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{w_d(t_i)}{\|d\|} - \frac{w_d(t_i) w_d(t_j)}{\varpi} \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{w_d(t_j)}{\|d\|} - \frac{w_d(t_i) w_d(t_j)}{\varpi} \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j)\end{aligned} \qquad (6)$$

Also, the verification conditions are given by the following lemmas and theorem. Proofs of Lemmas 3-4 and Theorem 2 are here omitted as they are similar to the respective proofs of Lemmas 1-2 and Theorem 1.

**Lemma 3.** *For an arbitrary term $t \in V_d$, we have*

$$\varpi = \|d\| w_d(t) - w_d^2(t) + \varpi_t$$

**Lemma 4.** *For functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ given in (5) and (6), respectively, we have:*

*$\varpi_{t_i} \geq w_d^2(t_i)$ if and only if $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$;*

*$\varpi_{t_j} \geq w_d^2(t_j)$ if and only if $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$.*

**Theorem 2.** *$P_d(\delta_i, \delta_j)$ given in (6) is a probability distribution if it satisfies two inequalities: a) $\varpi_{t_i} \geq w_d^2(t_i)$ and b) $\varpi_{t_j} \geq w_d^2(t_j)$.*

Obviously, $w_d(t)$ is the main component of the estimation functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$. As we all know,

document representations, $w_d(t)$, play an essential role in determining effectiveness. The issue of accuracy and validity of document representation has long been a crucial and open problem. It is beyond the scope of this paper to discuss the issue in greater detail. A detailed discussion about representation techniques may be found, for instance, in study [3][4].

## 5 Conclusion

It seems that MI methods have not achieved their potential for automatically measuring statistical dependence of terms. The main problem in MI methods is to obtain actual probability distributions estimated from training data. This study concentrated on such a problem and proposed a novel but simple method for measures. We introduced estimation functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$, which may be used to capture the occurrence and co-occurrence information of terms and to define distributions $P_d(\delta)$ and $P_d(\delta_i, \delta_j)$. We interpreted mathematical meaning of the functions within practical application contexts. We discussed verification conditions in order to ensure $P_d(\delta)$ and $P_d(\delta_i, \delta_j)$ are probability distributions under the conditions. We provided examples to clarify the idea of our method, to make understandable the computation involved in all the formulae and, in particular, to illustrate the possibility of failure of applying our method if the verification conditions are not satisfied. We considered the possibility of extension of our method, indicated that it is applicable to any quantitative document representations with a weighting function. The generality of the formal discussion means our method can be applicable to many areas of science, involving statistical semantic analysis of textual data.

## Appendix

**Lemma 1.** *For an arbitrary term $t \in V_d$, we have*

$$\varpi = \|d\| f_d(t) - f_d^2(t) + \varpi_t$$

**Proof.** Without losing generality, suppose $t = t_{i_1}$. (Otherwise, let $t = t_{i_h}$. Notice that the order of the elements in the set is unnecessary, so we can rewrite $V_d = \{t_{i_h}, t_{i_1}, \dots, t_{i_{h-1}}, t_{i_{h+1}}, \dots, t_{i_s}\}$, and thus $V_d - \{t\} = \{t_{i_1}, \dots, t_{i_{h-1}}, t_{i_{h+1}}, \dots, t_{i_s}\}$ with $1 \le i_1 < \cdots < i_{h-1} < i_{h-2} < \cdots < i_s \le n$. So our discussion still holds.) Thus we have

$$\begin{aligned}\varpi &= f_d(t_{i_1})[f_d(t_{i_2}) + \cdots + f_d(t_{i_s})] \\ &\quad + f_d(t_{i_2})[f_d(t_{i_3}) + \cdots + f_d(t_{i_s})] \\ &\quad + \cdots + f_d(t_{i_{s-2}})[f_d(t_{i_{s-1}}) + f_d(t_{i_s})] \\ &\quad + f_d(t_{i_{s-1}})[f_d(t_{i_s})] \\ &= f_d(t)[\|d\| - f_d(t)] + f_d(t_{i_2}) \textstyle\sum_{j'=i_3,\dots,i_s} f_d(t_{j'}) \\ &\quad + \cdots + f_d(t_{i_{s-1}}) \textstyle\sum_{j'=i_s} f_d(t_{j'}) \\ &= \|d\| f_d(t) - f_d^2(t) \\ &\quad + \textstyle\sum_{i'<j';\, t_{i'}, t_{j'} \in V_d - \{t\}} f_d(t_{i'}) f_d(t_{j'}) \\ &= \|d\| f_d(t) - f_d^2(t) + \varpi_t\end{aligned}$$ □

**Lemma 2.** *Functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ given in (1) have:*

*(a) $\varpi_{t_i} \ge f_d^2(t_i)$ if and only if $\psi_d(t_j) \ge \gamma_d(t_i, t_j)$;*

*(b) $\varpi_{t_j} \ge f_d^2(t_j)$ if and only if $\psi_d(t_i) \ge \gamma_d(t_i, t_j)$.*

**Proof.** We only prove (a). The proof of (b) is similar to (a). Notice that $\varpi \ne 0$ and $\|d\| \ne 0$. Thus, by Lemma 1, we have

$$\varpi_{t_i} - f_d^2(t_i) \ge 0$$

if and only if

$$\varpi = \|d\| f_d(t_i) + [\varpi_{t_i} - f_d^2(t_i)] \ge \|d\| f_d(t_i)$$

if and only if

$$\varpi f_d(t_j) \ge \|d\| f_d(t_i) f_d(t_j)$$

if and only if

$$\psi_d(t_j) = \frac{f_d(t_j)}{\|d\|} \ge \frac{f_d(t_i) f_d(t_j)}{\varpi} = \gamma_d(t_i, t_j)$$ □

**Theorem 1.** *$P_d(\delta_i, \delta_j)$ given in (4) is a probability distribution if a) $\varpi_{t_i} \ge f_d^2(t_i)$ and b) $\varpi_{t_j} \ge f_d^2(t_j)$*

**Proof.** $P_d(\delta_i = 1, \delta_j = 1) > 0$ as $0 < \gamma_d(t_i, t_j) < 1$; $P_d(\delta_i = 1, \delta_j = 0), P_d(\delta_i = 0, \delta_j = 1) \ge 0$ as $\psi_d(t_i), \psi_d(t_j) \ge \gamma_d(t_i, t_j)$ by Lemma 2; $P_d(\delta_i = 0, \delta_j = 0) = [1 - \psi_d(t_i)] - [\psi_d(t_j) - \gamma_d(t_i, t_j)] > 0$ as $0 < \psi_d(t_i) < 1$. Finally, $\sum_{\delta_i, \delta_j = 1,0} P_d(\delta_i, \delta_j) = 1$ can easily be seen from (3). □

## References

[1] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol 3, no. 10, pp. 1168-1174, 2010.

[2] A. E. Akadi, A.E. Abdeljalil El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, pp. 116-121, 2008.

[3] D. Cai, "An information theoretic foundation for the measurement of discrimination information," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 9, pp. 1262-1273, 2010.

[4] D. Cai, "Determining semantic relatedness through the measurement of discrimination information using Jensen difference," *International Journal of Intelligent Systems*, vol. 24, no. 5, pp. 477-503, 2009.

[5] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Journal of the American Society*

*for Information Science*, vol. 16, no. 1, pp. 22-29, 1990.

[6] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," *Proc. 29th Ann. International ACM-SIGIR Conf. Research and Development in Information Retrieva*l, pp. 115-122, 2006.

[7] S. Gauch, J. Wang, and S. M. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases," *ACM Trans. Information Systems*, vol. 17, no. 3, pp. 250-269, 1999.

[8] M. Kim and K. Choi, "A comparison of collocation-based similarity measures in query expansion," *Information Processing & Management*, vol. 35, no. 1, pp. 19-30, 1999.

[9] S. Kullback, *Information Theory and Statistics*, New York: Wiley, 1959.

[10] H.-W. Liu, J.-G. Sun, L. Liu, and H.-J. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, pp. 1330-1339, 2009.

[11] R. M. Losee, Jr., "Term dependence: A basis for Luhn and Zipf models," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 12, pp. 1019-1025, 2001.

[12] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187-198, 1997.

[13] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361-378, 2000.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System and Technical Journal*, vol. 27, pp. 379-423, 623-656, 1948.

[16] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movies," *Proc. IEEE International Conf. Multimedia and Expo. (ICME'06)*, pp. 1013-1016, 2006.

[17] G. Wang, F.H. Lochovsky, and Q. Yang, "Feature selection with conditional mutual information maximin in text categorization," *Proc. 10th International Conf. Information and Knowledge Management*, pp. 342-349, 2004.

**Di Cai** received her PhD in the Department of Computing Science at the University of Glasgow in UK. She is currently a research fellow in the School of Computing and Engineering at the University of Hudersfield in UK. Her main research interests include information extraction and retrieval, document classification and summarization, text mining and opinion mining, emotion and sentiment analysis. She is a member of the IEEE and ACM.

**Thomas Leo McCluskey** is professor of software technology at the University of Huddersfield in the UK, and director of research for the University's School of Computing and Engineering. His research interests include software and knowledge engineering, domain modelling, planning and machine learning. His research group has developed a series of knowledge engineering aids which help in the formulation process of structural and heuristic planning knowledge, ranging from interactive interfaces to fully automated learning tools.