A Simple Method for Estimating Term Mutual Information

Original Citation


This version is available at http://eprints.hud.ac.uk/15984/

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

http://eprints.hud.ac.uk/
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 science, 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. This study focuses on the problem 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.

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 \( t^\delta = t, \overline{t} \) when \( \delta = 1, 0 \), respectively. We call \( \Omega \) a state value space, and each element in \( \Omega \) a state value, oft. 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:

\[
I_d(\delta_i, \delta_j) = H(\delta_i) - H(\delta_i | \delta_j) = H(\delta_j) - H(\delta_j | \delta_i) = \sum_{\delta \in \Omega} P_d(\delta) \log \frac{P_d(\delta_i) P_d(\delta_j)}{P_d(\delta_i) P_d(\delta_j)}
\]
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 \( V_d = \{ t_{i_1}, t_{i_2}, ..., t_{i_n} \} \subseteq \{ t_1, t_2, ..., t_n \} = V \), where \( 1 \leq i_1 < i_2 < \cdots < i_s \leq n \). In this study, we always assume that \( 2 < s = |V_d| \leq n \) (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

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

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 us introduce two estimation functions:

\[
\psi_d(t) = \frac{f_d(t)}{\sum_{t' \in V_d} f_d(t')}
\]

(1)

\[
\gamma_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\sum_{t' \in V_d} f_d(t_i) f_d(t_j) f_d(t')}
\]

(2)

Clearly, \( 0 < \psi_d(t), \gamma_d(t_i, t_j) \) 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

\[
P_d(\delta = 1) = P_d(t) = \psi_d(t)
\]

\[
P_d(\delta = 0) = P_d(\emptyset) = 1 - \psi_d(t)
\]

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

\[
\sigma = \sum_{t', t_j \in V_d} f_d(t_i) f_d(t_j)
\]

and, for an arbitrary term \( t \in V_d \), denote

\[
\sigma_t = \sum_{t' < t_j; t_i, t_j \in V_d} f_d(t_i) f_d(t_j)
\]

Clearly \( \sigma > \sigma_t \geq 1 \) when \( |V_d| > 2 \).

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

\[
\sigma_t = ||d|| f_d(t) - f_d^2(t) + \sigma_t
\]

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

\( a \) \( \sigma_t \geq f_d^2(t) \) if and only if \( \psi_d(t) \geq \gamma_d(t_i, t_j) \);

\( b \) \( \sigma_t \geq f_d^2(t) \) if and only if \( \psi_d(t) \geq \gamma_d(t_i, t_j) \).

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

\[
P_d(\delta_i, \delta_j) = \varphi_d(\delta_i, \delta_j)
\]

is a probability distribution over \( \Omega \times \Omega \) if it satisfies two inequalities: \( a \) \( \sigma_t \geq f_d^2(t) \) and \( b \) \( \sigma_t \geq f_d^2(t) \).

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

\[
P_d(\delta_i = 1, \delta_j = 1) = P_d(t_i, t_j) = \gamma_d(t_i, t_j)
\]

\[
P_d(\delta_i = 1, \delta_j = 0) = P_d(t_i, \emptyset) = \psi_d(t_i) - \gamma_d(t_i, t_j)
\]

The first lemma tells us that there exists a relationship between \( \sigma \) and \( \sigma_t \). The second lemma tells us how to
verify two inequalities (conditions) required in Theorem 1: \( \sigma \geq f_d(t_i) \) and \( \sigma \geq f_d(t_j) \) 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, \( \{ \delta_i, \delta_j \} \), is rather intuitive. \( \{ \delta_i = 1, \delta_j = 0 \} \) and \( \{ \delta_i = 0, \delta_j = 1 \} \) 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

\[
\sum_{\delta_i, \delta_j = 1, 0} P(\delta_i, \delta_j) = 1
\]

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

\[
m_d = \begin{bmatrix} f_d(t_1) & f_d(t_2) & \ldots & f_d(t_n) \end{bmatrix} = [f_d(t)]_{1 \times n}
\]

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

\[
m'_d \times m_d = \begin{bmatrix} f_d(t_1) & f_d(t_2) & \ldots & f_d(t_n) \\
\vdots & \vdots & \ddots & \vdots \\
f_d(t_1) & f_d(t_2) & \ldots & f_d(t_n) \\
\end{bmatrix}
\]

\[
= \begin{bmatrix} f_d(t_1) f_d(t_1) & f_d(t_1) f_d(t_2) & \ldots & f_d(t_1) f_d(t_n) \\
f_d(t_2) f_d(t_1) & f_d(t_2) f_d(t_2) & \ldots & f_d(t_2) f_d(t_n) \\
\vdots & \vdots & \ddots & \vdots \\
f_d(t_n) f_d(t_1) & f_d(t_n) f_d(t_2) & \ldots & f_d(t_n) f_d(t_n) \\
\end{bmatrix}
\]

\[
= \begin{bmatrix} \sigma f_d(t_1) f_d(t_1) & \sigma f_d(t_1) f_d(t_2) & \ldots & \sigma f_d(t_1) f_d(t_n) \\
\sigma f_d(t_2) f_d(t_1) & \sigma f_d(t_2) f_d(t_2) & \ldots & \sigma f_d(t_2) f_d(t_n) \\
\vdots & \vdots & \ddots & \vdots \\
\sigma f_d(t_n) f_d(t_1) & \sigma f_d(t_n) f_d(t_2) & \ldots & \sigma f_d(t_n) f_d(t_n) \\
\end{bmatrix}_{n \times n}
\]

\[
= \sigma \begin{bmatrix} P_d(\delta_i = 1, \delta_j = 1) \end{bmatrix}_{n \times n}
\]

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

\[
[P_d(\delta_i = 1, \delta_j = 1)]_{n \times n} = \frac{1}{\sigma} [f_d(t_i) f_d(t_j)]_{n \times n}
\]

\[
= [\gamma_d(t_i, t_j)]_{n \times n}
\]

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 \( i, j = 1, \ldots, n \), 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, \( \sigma \), the sum of all possible numerators \( f_d(t_i) f_d(t_j) \) for \( i < j \); \( i, j = 1, \ldots, n \), is a normalization factor for the characterization. Clearly, \( \sigma \) 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 \( [P_d(\delta_i = 1, \delta_j = 1)]_{n \times n} \neq [0]_{n \times n} \). 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 \( i = 1, \ldots, n \), 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 \( I(\delta_i, \delta_j) \) 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 \( d_1 = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\} \), then we have \( V_{d_1} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\} \), and \( \sigma = 26 \). Thus, for term pair \((t_1, t_2)\), from (2) and (4), we have

\[
P_{d_1}(\delta_i = 1, \delta_j = 1) = \frac{1 \times 2}{26} = \frac{8}{104}
\]

\[
P_{d_1}(\delta_i = 1, \delta_j = 0) = \frac{1}{8} - \frac{2}{26} = \frac{5}{104} > 0
\]

\[
P_{d_1}(\delta_i = 0, \delta_j = 1) = \frac{2}{8} - \frac{2}{26} = \frac{18}{104} > 0
\]

\[
P_{d_1}(\delta_i = 0, \delta_j = 0) = 1 - \frac{1}{8} - \frac{2}{26} = \frac{73}{104}
\]

Then, it follows immediately (using natural logarithms)

\[
l_{d_1}(\delta_i, \delta_j) = \log \frac{8}{104} + \log \frac{5}{104} + \log \frac{18}{104} + \log \frac{73}{104}
\]

\[
\approx 0.0693 - 0.0321 - 0.0405 + 0.0472 = 0.0439
\]

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

\[
\sigma = \sum_{(i < j) \in V_{d_2}} f_{d_2}(t_i) f_{d_2}(t_j)
\]

\[
= f_{d_2}(t_1) f_{d_2}(t_2) + f_{d_2}(t_1) f_{d_2}(t_3) + f_{d_2}(t_2) f_{d_2}(t_3) + f_{d_2}(t_2) f_{d_2}(t_4) + f_{d_2}(t_3) f_{d_2}(t_4)
\]

\[
= 1 \times 3 + 1 \times 1 + 1 \times 3 + 1 \times 3 + 1 \times 1 + 1 \times 1 = 12
\]

Thus, for instance, for term pair \((t_1, t_2)\), we have

\[
\gamma_{d_2}(t_1, t_2) = \frac{1 \times 3}{12} = \frac{1}{4}
\]

\[
P_{d_2}(\delta_i = 1, \delta_j = 0) = \psi_{d_2}(t_1) - \gamma_{d_2}(t_1, t_2)
\]

\[
= 1/6 - 3/12 = -1/12 < 0
\]

From which we can conclude that \( P_{d_2}(\delta_i, \delta_j) \) is not a
probability distribution since \( \sigma_{a_{d}}(t_1) - \gamma_{a_{d}}(t_1, t_2) < 0 \). Also, we can verify this in an alternative way:

\[
\sigma_{a_{d}} = \sum_{i' \neq i} \sum_{t_{i'}, t_{i'} \in V \setminus V_d} f_{d}(t_{i'})f_{d}(t_{i'}) \leq f_{d}(t_{i})f_{d}(t_{i}) = f_{d}(t_{i})f_{d}(t_{i}) + f_{d}(t_{i})f_{d}(t_{i}) \leq 1 + 1 + 1 < 9 = f_{d}(t_{i})f_{d}(t_{i})
\]

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 \( \psi_{a_{d}}(t_{i}) \geq \gamma_{a_{d}}(t_{i}, t_{i}) \) and \( \psi_{a_{d}}(t_{i}) \geq \gamma_{a_{d}}(t_{i}, t_{i}) \), or equivalently to verify both \( \sigma_{a_{d}} \geq f_{a_{d}}^2(t_{i}) \) and \( \sigma_{a_{d}} \geq f_{a_{d}}^2(t_{i}) \), satisfied simultaneously, for each term pair considered.

(ii) \( \gamma_{a_{d}}(t_{i}, t_{i}) \) 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 \( \delta_{i}, \delta_{j} \in \mathbb{N} \) 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:

\[
P_{d}(\delta_{i} = 1, \delta_{j} = 1) = \frac{f_{a_{d}}(t_{i})f_{a_{d}}(t_{j})}{\sigma} = \gamma_{a_{d}}(t_{i}, t_{j}) \geq \psi_{a_{d}}(t_{i}) \psi_{a_{d}}(t_{j}) = \frac{f_{a_{d}}(t_{i})}{\|d\|} \cdot \frac{f_{a_{d}}(t_{j})}{\|d\|}
\]

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

\[
m_{d} = [w_{d}(t_{1}), w_{d}(t_{2}), ..., w_{d}(t_{n})] = [w_{d}(t)]_{1 \times n}
\]
satisfying \( w_{d}(t) > 0 \) when \( t \in V \) and \( w_{d}(t) = 0 \) when \( t \in V \setminus V_d \). 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

\[
\|d\| = \sum_{t \in \mathbb{V}} w_{d}(t)
\]

and denote

\[
\sigma_{t} = \sum_{i' < i, t_{i'}, t_{i'} \in V \setminus V_d} f_{d}(t_{i'})f_{d}(t_{i'}) < \sum_{i' < i, t_{i'}, t_{i'} \in V \setminus V_d} w_{d}(t_{i'})w_{d}(t_{i'}) = \sigma
\]

Then, for arbitrary terms \( t, t_{i}, t_{j} \in \mathbb{V} \), similar to the expressions given in (3) and (4), we may write the corresponding marginal distribution:

\[
P_{d}(\delta_{i} = 1) = \frac{w_{d}(t_{i})w_{d}(t_{j})}{\|d\|} = p_{d}(t_{i}) = \psi_{d}(t_{i}) \

P_{d}(\delta_{i} = 0, \delta_{j} = 1) = \frac{w_{d}(t_{i})}{\|d\|} - \frac{w_{d}(t_{i})w_{d}(t_{j})}{\|d\|} = \overline{p_{d}(t_{i})} = \gamma_{d}(t_{i}, t_{j})
\]

and joint distribution

\[
P_{d}(\delta_{i} = 1, \delta_{j} = 1) = \frac{w_{d}(t_{i})w_{d}(t_{j})}{\|d\|} = \gamma_{d}(t_{i}, t_{j})
\]

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

\[
\sigma = \|d\|w_{d}(t) - w_{d}(t) + \sigma_{t}
\]

**Lemma 4.** For functions \( \psi_{d}(t) \) and \( \gamma_{d}(t_{i}, t_{j}) \) given in (5) and (6), respectively, we have:

\[
\sigma_{t_{i}} \geq w_{d}^{*}(t_{i}) \text{ if and only if } \psi_{d}(t_{i}) \geq \gamma_{d}(t_{i}, t_{i}) ;
\]

\[
\sigma_{t_{j}} \geq w_{d}^{*}(t_{j}) \text{ if and only if } \psi_{d}(t_{i}) \geq \gamma_{d}(t_{i}, t_{j}) .
\]

**Theorem 2.** \( P_{d}(\delta_{i}, \delta_{j}) \) given in (6) is a probability distribution if it satisfies two inequalities: a) \( \sigma_{t_{i}} \geq w_{d}^{*}(t_{i}) \) and b) \( \sigma_{t_{j}} \geq w_{d}^{*}(t_{j}) \).

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 the 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

\[
\sigma = \|d\|f_d(t) - f^2_d(t) + \sigma_t
\]

Proof. Without losing generality, suppose \( t = t_i \). (Otherwise, \( t = t_{i_{h}} \).) Notice that the order of the elements in the set is unnecessary, so we can rewrite \( V_d = \{ t_{i_{h}}, t_{i_{i_1}}, \ldots, t_{i_{i_{2}}}, t_{i_{i_{3}}}, \ldots, t_{i_{i_{s}}} \} \), and thus \( V_d - \{t\} = \{ t_{i_{s_1}}, \ldots, t_{i_{h_{1}}}, t_{i_{h_{2}}}, \ldots, t_{i_{h_{s}}} \} \) with \( 1 \leq i_{s_2} < \cdots < i_{h_{1}} < i_{h_{2}} < \cdots < i_{s_2} \leq n \). So our discussion still holds.)
Thus we have
\[
\sigma = f_d(t_i)\left[f_d(t_{i_{s_1}}) + \cdots + f_d(t_{i_{s_2}})\right]
+ f_d(t_{i_{s_1}})\left[f_d(t_{i_{s_2}}) + \cdots + f_d(t_{i_{s_3}})\right]
+ \cdots + f_d(t_{i_{h_{1}}})\left[f_d(t_{i_{h_{2}}}) + \cdots + f_d(t_{i_{h_{s}}})\right]
+ f_d(t_{i_{h_{1}}})\left[f_d(t_{i_{h_{2}}}) + \cdots + f_d(t_{i_{h_{s}}})\right]
= f_d(t_i)[\|d\| - f_d(t_i)] + f_d(t_{i_{s_1}})\sum_{i'=i_{s_3}}^{i_{s_2}} f_d(t_{i'})
+ \cdots + f_d(t_{i_{h_{1}}})\sum_{i'=i_{h_{2}}}^{i_{h_{s}}} f_d(t_{i'})
\]

Reference:


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 Huddersfield 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.