A General Framework of Generating Estimation Functions for Computing the Mutual Information of Terms

D. Cai and T.L. McCluskey
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK, HD1 3DH
Email: {d.cai; t.l.mccluskey}@hud.ac.uk

***Abstract*—Computing statistical dependence of terms in textual documents is a widely studied subject and a core problem in many areas of science. This study focuses on such a problem and explores the techniques of estimation using the expected mutual information measure. A general framework is established for tackling a variety of estimations: (i) general forms of estimation functions are introduced; (ii) a set of constraints for the estimation functions is discussed; (iii) general forms of probability distributions are defined; (iv) general forms of the measures for calculating mutual information of terms (MIT) are formalised; (v) properties of the MIT measures are studied and, (vi) relations between the MIT measures are revealed. Four estimation methods, as examples, are proposed and mathematical meanings of the individual methods are respectively interpreted. The methods may be directly applied to practical problems for computing dependence values of individual term pairs. Due to its generality, our method is applicable to various areas, involving statistical semantic analysis of textual data.**

***Index Terms*—mutual information of terms (MIT); term dependence; statistical semantic analysis; probability estimation.**

I. Introduction

Analysing and computing statistical dependence (relatedness, proximity, association, similarity) of terms (features, concepts, phrases, words) in textual documents is a widely studied subject in many areas of science. The subject has achieved importance and popularity during the past four decades or so, due chiefly to its demonstrated applications in numerous seemingly diverse areas of science. One of the commonly used tools of analysis and computation is the expected mutual information measure (EMIM) drawn from information theory [1], [2].

The issue of computing the mutual information of terms is an active research topic. A variety of methods have been developed in order to assign dependence values to individual term pairs, and then some decision is made on the basis of the values. Many studies have used the measure for a variety of tasks in, for instance, feature selection [3]–[6], document classification [7], face image clustering [8], multimodality image registration [9], information retrieval [10]–[14]. However, it seems that mutual information methods have not achieved their potential. The main problem we face in using EMIM 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.

Before introducing a series of formulae, let us first clarify the difference between a term *state value* distribution and a term *occurrence frequency* 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 δ 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 Ω a *state value space*, and each element in Ω a *state value*, of t. Similarly, for an arbitrary term pair (t_i, t_j), we introduce a variable pair (δ_i, δ_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 $D = \{d_1, d_2, ..., d_m\}$ be a *collection* of documents (training data), and $V = \{t_1, t_2, ..., t_n\}$ 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 a given d, the term occurrence frequency distribution, generally denoted by $p_d(t) = p(t|d)$, is over V, whereas for a given term t occurring in d, its state value distribution, denoted by $P_d(\delta) = P(t^\delta|d)$, is over Ω. Obviously, each term $t \in V_d$ is matched to a state value distribution and there are $|V_d|$ state value distributions in total for the 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 the other [15]. The study [16] shows that there is a relationship between the frequencies (or probabilities) of terms and the mutual information of terms. Therefore, term t_i taking some state value δ_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 $t_j \in V - \{t_i\}$), may be dependent on this δ_i [15].

Mathematically, for two arbitrary distinct terms $t_i, t_j \in V$, the *expected mutual information* [1] about the probabilities of the state value pair (δ_i, δ_j) of term pair (t_i, t_j) can be

expressed by EMIM:

$$I(\delta_i;\delta_j) = \sum_{\delta_i,\delta_j=1,0} P(\delta_i,\delta_j)\log\frac{P(\delta_i,\delta_j)}{P(\delta_i)P(\delta_j)}$$

which measures the amount of information that δ_j provides about δ_i, and vice versa.

Intuitively, a high $I(\delta_i;\delta_j)$ value indicates more of the information that one of two terms t_i and t_j carries is determined by the other and thus the terms are more dependent; a low $I(\delta_i;\delta_j)$ value on the other hand suggests that t_i and t_j are better able to provide self-information and thus are likely to be independent. However, the current study does not support this intuition and instead points out:

1) one should consider the mutual information of t_i and t_j under the individual state values (δ_i,δ_j), where $\delta_i,\delta_j = 1,0$;
2) one cannot assert that t_i and t_j are highly dependent for their co-occurrence from a high $I(\delta_i;\delta_j)$ value.

The estimation of probability distributions, $P(\delta)$ and $P(\delta_i,\delta_j)$, required in $I(\delta_i;\delta_j)$ 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. We attempt to establish a general framework for constructing estimation functions, with a set of constraints, in order to define $P(\delta)$ and $P(\delta_i,\delta_j)$ meeting some criteria. We next formalise measures for computing the mutual information of terms (MIT) under the individual state values and study corresponding properties of the MIT measures, which is an underlying basis for practical applications. We then propose four estimation methods, as examples, to clarify and illustrate our ideas described in the current study by interpreting their mathematical meanings and discussing corresponding properties. The four estimation methods may be applied directly to practical problems for assigning a dependence value to each term pair.

The remainder of the paper is organized as follows. Section II establishes a general framework for constructing estimation functions and defining probability distributions. Section III formalises the MIT measures and studies their properties. Section IV proposes four estimation methods and discusses corresponding properties. Section V addresses some key points of our study. Conclusions are drawn in Section VI.

II. A General Estimation Framework

In practical applications, the probability distributions of state values may be estimated from training data. This section establishes a general framework in order to define two arguments, $P(\delta)$ and $P(\delta_i,\delta_j)$, required in $I(\delta_i;\delta_j)$. The definition of the joint state value distribution, $P(\delta_i,\delta_j)$, is a more complicated task and the main concern of this section.

In the current study, the probability distributions are defined from estimation functions and, therefore, we need to first introduce the concept of estimation functions. Let $\Xi \subseteq D$ be the set of sample documents considered, and $V_\Xi \subseteq V$ the set of terms occurring in at least one of the documents in Ξ. We have the following definition.

Definition 2.1 For arbitrary terms $t, t_i, t_j \in V$, where $i \neq j$, we define two non-negative functions, denoted by $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$, with the form:

$$\begin{aligned}\psi_\Xi(t)&\begin{cases}>0 & t\in V_\Xi\\ =0 & t\notin V_\Xi\end{cases}\\ \gamma_\Xi(t_i,t_j)&\begin{cases}>0 & (t_i,t_j)\in V_\Xi\times V_\Xi\\ =0 & (t_i,t_j)\notin V_\Xi\times V_\Xi\end{cases}\end{aligned} \tag{1}$$

satisfying a set of *constraints*

$$0 \le \gamma_\Xi(t_i,t_j) \le \psi_\Xi(t_i),\psi_\Xi(t_j) < 1 \tag{2}$$

and call $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$ the *general forms of estimation functions*.

Definition 2.2 For arbitrary given terms $t, t_i, t_j \in V_\Xi$, where $i \neq j$, suppose $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$ are the estimation functions given in Definition 2.1. We define $P_\Xi(\delta)$:

$$\begin{aligned}P_\Xi(\delta=1)&=\psi_\Xi(t)\\ P_\Xi(\delta=0)&=1-\psi_\Xi(t)\end{aligned} \tag{3}$$

and define $P_\Xi(\delta_i,\delta_j)$:

$$\begin{aligned}P_\Xi(\delta_i=1,\delta_j=1)&=\gamma_\Xi(t_i,t_j)\\ P_\Xi(\delta_i=1,\delta_j=0)&=\psi_\Xi(t_i)-\gamma_\Xi(t_i,t_j)\\ P_\Xi(\delta_i=0,\delta_j=1)&=\psi_\Xi(t_j)-\gamma_\Xi(t_i,t_j)\\ P_\Xi(\delta_i=0,\delta_j=0)&=1-\psi_\Xi(t_i)-\psi_\Xi(t_j)+\gamma_\Xi(t_i,t_j)\end{aligned} \tag{4}$$

and call $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$ the *general forms of probability distributions* of state values of term pair (t_i,t_j).

Theorem 2.1 Suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$ are given in Definition 2.2. Then $P_\Xi(\delta)$ is a probability distribution on $\Omega=\{1,0\}$; $P_\Xi(\delta_i,\delta_j)$ is a probability distribution on $\Omega\times\Omega$; $P_\Xi(\delta_i)$ and $P_\Xi(\delta_j)$ are the marginal distributions of $P_\Xi(\delta_i,\delta_j)$.

Proof: Clearly, from the above definition and constraints given in (2), $P_\Xi(\delta)$ is a probability distribution on $\Omega=\{1,0\}$. Also, by the constraints and four expressions in (4), we have

$$P_\Xi(\delta_i,\delta_j) \ge 0$$

for $\delta_i,\delta_j = 1,0$ and

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

Thus $P_\Xi(\delta_i,\delta_j)$ is a probability distribution on $\Omega\times\Omega$. Also, it can easily be seen:

$$\begin{aligned}P_\Xi(\delta_i=1)&=\sum_{\delta_j=1,0}P_\Xi(\delta_i=1,\delta_j)=\psi_\Xi(t_i)\\ P_\Xi(\delta_i=0)&=\sum_{\delta_j=1,0}P_\Xi(\delta_i=0,\delta_j)=1-\psi_\Xi(t_i)\end{aligned}$$

Hence, $P_\Xi(\delta_i)$ is the marginal distributions of $P_\Xi(\delta_i,\delta_j)$. A similar discussion may be given for $P_\Xi(\delta_j)$. □

Let us next examine the absolute continuity of $P_\Xi(\delta_i,\delta_j)$ with respect to $P_\Xi(\delta_i)P_\Xi(\delta_j)$, or in symbols, $P_\Xi(\delta_i,\delta_j) \ll P_\Xi(\delta_i)P_\Xi(\delta_j)$. The following theorem serves this purpose.

Theorem 2.2 Suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$ are given in Definition 2.2. Then, $P_\Xi(\delta_i,\delta_j) \ll P_\Xi(\delta_i)P_\Xi(\delta_j)$ for $\delta_i,\delta_j = 1,0$.
Proof: The proof is trivial: It can be easily seen, by expressions (1) and (3), that it always has $0 < P_\Xi(\delta_i), P_\Xi(\delta_j) < 1$ for $\delta_i,\delta_j = 0,1$ if $t_i,t_j \in V_\Xi$. □

It should be emphasized that in order to speak of the mutual information of terms, we must verify the two arguments of $I(\delta_i,\delta_i)$ meeting the following three *criteria* simultaneously:

1) $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$ are probability distributions,
2) $P_\Xi(\delta_i)$ and $P_\Xi(\delta_j)$ are the marginal distributions of $P_\Xi(\delta_i,\delta_j)$,
3) $P_\Xi(\delta_i,\delta_j)$ is absolutely continuous with respect to $P_\Xi(\delta_i)P_\Xi(\delta_j)$.

Meeting these three criteria is the major premise when applying $I(\delta_i;\delta_j)$ to effectively capture the mutual information inherent among terms. We will give an example to clarify our idea here in Section V.

We thus learn from Theorems 2.1 and 2.2, under the general framework, that as long as $P_d(\delta)$ and $P_d(\delta_i,\delta_j)$ are defined from the estimation functions satisfying the constraints given in (2), they are probability distributions meeting the three criteria. Consequently, the difficulty becomes:

- to construct $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$ that can capture the occurrence and co-occurrence information of terms practically appropriate and mathematically meaningful in application contexts;
- to verify the constraints given in (2) for each term pair considered in order to ensure that the probability distributions, when defined from $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$, meeting the three criteria.

Thus, the construction of $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$ and verification of the constraints given in (2), which are relatively simple, are the core of obtaining actual probability distributions $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$. Section IV will return to this issue and provide four useful examples, after formalising the MIT measures and discussing their properties and relations in the next section.

III. The MIT Measures

Suppose we are given two arbitrary distinct terms $t_i,t_j \in V_\Xi$. In order to measure the mutual information of terms t_i and t_j, we need to consider the mutual information under each state value (δ_i,δ_j), namely, we need to measure the extent of the contribution made by the individual state values to EMIM:

$$\begin{aligned} I_\Xi(\delta_i;\delta_j) &= \sum_{\delta_i,\delta_j=1,0} P_\Xi(\delta_i,\delta_j)\log\frac{P_\Xi(\delta_i,\delta_j)}{P_\Xi(\delta_i)P_\Xi(\delta_j)} \\ &= \sum_{\delta_i,\delta_j=1,0} \mathbf{mit}_\Xi(t_i^{\delta_i},t_j^{\delta_j}) \end{aligned} \tag{5}$$

Note that the above expression can be expressed as a sum of four items. Each of four items,

$$\mathbf{mit}_\Xi(t_i^{\delta_i},t_j^{\delta_j}) = P_\Xi(\delta_i,\delta_j)\log\frac{P_\Xi(\delta_i,\delta_j)}{P_\Xi(\delta_i)P_\Xi(\delta_j)} \tag{6}$$

can be regarded as '**m**utual **i**nformation of **t**erms', t_i and t_j, in support of dependence but rejecting independence under state value (δ_i,δ_j), where $\delta_i,\delta_j = 1,0$. Thus, we can regard each item as a MIT measure, computing the extent of the contributions made by the corresponding state value to $I_\Xi(\delta_i;\delta_j)$.

Now, substituting estimates (3) and (4) into (6), corresponding to respective four state value pairs, $(1,1)$, $(1,0)$ $(0,1)$, $(0,0)$, we can formalise the general forms of the four MIT measures by a definition below:

Definition 3.1 Suppose $P_\Xi(\delta)$ and $P_\Xi(\delta_i,\delta_j)$ are the probability distributions given in Definition 2.2. Then the *general forms of four MIT measures* can be defined as follows.

$$\mathbf{mit}_\Xi(t_i,t_j) = \gamma_\Xi(t_i,t_j)\log\frac{\gamma_\Xi(t_i,t_j)}{\psi_\Xi(t_i)\psi_\Xi(t_j)}$$

which computes the dependence of terms t_i and t_j for their co-occurrence in Ξ;

$$\mathbf{mit}_\Xi(t_i,\bar{t}_j) = \big(\psi_\Xi(t_i)-\gamma_\Xi(t_i,t_j)\big)\log\frac{\psi_\Xi(t_i)-\gamma_\Xi(t_i,t_j)}{\psi_\Xi(t_i)\big(1-\psi_\Xi(t_j)\big)}$$

which computes the dependence of term t_i occurring but term t_j not occurring in Ξ;

$$\mathbf{mit}_\Xi(\bar{t}_i,t_j) = \big(\psi_\Xi(t_j)-\gamma_\Xi(t_i,t_j)\big)\log\frac{\psi_\Xi(t_j)-\gamma_\Xi(t_i,t_j)}{\big(1-\psi_\Xi(t_i)\big)\psi_\Xi(t_j)}$$

which computes the dependence of term t_i not occurring but term t_j occurring in Ξ;

$$\begin{aligned} \mathbf{mit}_\Xi(\bar{t}_i,\bar{t}_j) = \big(1-\psi_\Xi(t_i)-\psi_\Xi(t_j)+\gamma_\Xi(t_i,t_j)\big)\times \\ \log\frac{1-\psi_\Xi(t_i)-\psi_\Xi(t_j)+\gamma_\Xi(t_i,t_j)}{\big(1-\psi_\Xi(t_i)\big)\big(1-\psi_\Xi(t_j)\big)} \end{aligned}$$

which computes the dependence of both terms t_i and t_j not occurring in Ξ.

Clearly, each of the four MIT measures is uniquely determined by the estimation functions $\psi_\Xi(t)$ and $\gamma_\Xi(t_i,t_j)$.

Next, we give some interesting properties of the four MIT measures by Theorem 3.1 below. The properties derive their importance from the fact that they underpin the methods proposed in the current study and are essential for guiding practical applications.

Theorem 3.1 Suppose the four MIT measures are given in Definition 3.1. Then we have the following properties:

(a) if $\gamma_\Xi(t_i,t_j) > \psi_\Xi(t_i)\psi_\Xi(t_j)$ then

$$\begin{aligned} &\mathbf{mit}_\Xi(t_i,t_j) > 0, \quad \mathbf{mit}_\Xi(t_i,\bar{t}_j) \le 0 \\ &\mathbf{mit}_\Xi(\bar{t}_i,t_j) \le 0, \quad \mathbf{mit}_\Xi(\bar{t}_i,\bar{t}_j) > 0 \end{aligned}$$

(b) if $\gamma_\Xi(t_i,t_j) = \psi_\Xi(t_i)\psi_\Xi(t_j)$ then

$$\begin{aligned} &\mathbf{mit}_\Xi(t_i,t_j) = 0, \quad \mathbf{mit}_\Xi(t_i,\bar{t}_j) = 0 \\ &\mathbf{mit}_\Xi(\bar{t}_i,t_j) = 0, \quad \mathbf{mit}_\Xi(\bar{t}_i,\bar{t}_j) = 0 \end{aligned}$$

(c) if $\gamma_\Xi(t_i,t_j) < \psi_\Xi(t_i)\psi_\Xi(t_j)$ then

$$\begin{aligned} &\mathbf{mit}_\Xi(t_i,t_j) < 0, \quad \mathbf{mit}_\Xi(t_i,\bar{t}_j) \ge 0 \\ &\mathbf{mit}_\Xi(\bar{t}_i,t_j) \ge 0, \quad \mathbf{mit}_\Xi(\bar{t}_i,\bar{t}_j) < 0 \end{aligned}$$

Proof: The proof of (b) is obvious. Here we prove only (a), and a similar proof can be given for (c). Consider the general forms of the four MIT measures. From $\gamma_{\Xi}(t_i,t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$, we have:

$$\begin{aligned}
&\gamma_{\Xi}(t_i,t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i,t_j) < \psi_{\Xi}(t_i) - \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \psi_{\Xi}(t_i)\big(1-\psi_{\Xi}(t_j)\big)\\
&\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i,t_j) < \psi_{\Xi}(t_j) - \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \psi_{\Xi}(t_j)\big(1-\psi_{\Xi}(t_i)\big)\\
&1-\psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i,t_j)\\
&\qquad\qquad > 1-\psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\
&\qquad\qquad = \big(1-\psi_{\Xi}(t_i)\big)\big(1-\psi_{\Xi}(t_j)\big)
\end{aligned}$$

which correspond respectively to

$$\begin{aligned}
&\frac{\gamma_{\Xi}(t_i,t_j)}{\psi_{\Xi}(t_i)\psi_{\Xi}(t_j)} > 1,\\
&\frac{\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i,t_j)}{\psi_{\Xi}(t_i)\big(1-\psi_{\Xi}(t_j)\big)} < 1,\\
&\frac{\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i,t_j)}{\psi_{\Xi}(t_j)\big(1-\psi_{\Xi}(t_i)\big)} < 1,\\
&\frac{1-\psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i,t_j)}{\big(1-\psi_{\Xi}(t_i)\big)\big(1-\psi_{\Xi}(t_j)\big)} > 1
\end{aligned}$$

On the other hand, $0 < \gamma_{\Xi}(t_i,t_j) \le \psi_{\Xi}(t_i) < 1$ and $0 < \gamma_{\Xi}(t_i,t_j) \le \psi_{\Xi}(t_j) < 1$ for $t_i, t_j \in V_{\Xi}$. Thus, we have

$$\begin{aligned}
&\gamma_{\Xi}(t_i,t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j) > 0\\
&\psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i,t_j) \ge 0\\
&\psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i,t_j) \ge 0\\
&1-\psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i,t_j)\\
&\qquad = \big(1-\psi_{\Xi}(t_i)\big)\big(1-\psi_{\Xi}(t_j)\big) > 0
\end{aligned}$$

Hence, the four inequalities in (a) hold. □

The properties given in Theorem 3.1 enable us to gain an insight into the signs of the four MIT measures. That is, we have

$$\mathbf{mit}_{\Xi}(t_i,t_j), \mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j) \begin{cases} > 0 & \gamma_{\Xi}(t_i,t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\ = 0 & \gamma_{\Xi}(t_i,t_j) = \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\ < 0 & \gamma_{\Xi}(t_i,t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\end{cases}$$

$$\mathbf{mit}_{\Xi}(t_i,\bar{t}_j), \mathbf{mit}_{\Xi}(\bar{t}_i,t_j) \begin{cases} \le 0 & \gamma_{\Xi}(t_i,t_j) > \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\ = 0 & \gamma_{\Xi}(t_i,t_j) = \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\\ \ge 0 & \gamma_{\Xi}(t_i,t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)\end{cases}$$

Clearly, the relation between $\gamma_{\Xi}(t_i,t_j)$ and $\psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$ can infer all the signs of $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$. Thus, with the properties given in Theorem 3.1, we can further learn the relations of the four MIT measures from the signs:

- The signs of $\mathbf{mit}_{\Xi}(t_i,t_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j)$ are always the same, so are the signs of $\mathbf{mit}_{\Xi}(t_i,\bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)$;
- The signs of $\mathbf{mit}_{\Xi}(t_i,t_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i,\bar{t}_j)$ are always opposite to the signs of $\mathbf{mit}_{\Xi}(t_i,\bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i,t_j)$.

The relations tells us a key point of applying $I_{\Xi}(\delta_i;\delta_j)$, which we will explain in Section V.

IV. Example Estimations

As mentioned previously, the construction of the estimation functions and verification of the constraints are the core of defining actual probability distributions. This section presents four estimation methods, as examples, to illustrate our ideas described in the previous section. The first three consider the estimates in individual documents (i.e., $|\Xi| = 1$), and the last one considers the estimate in the set of documents (i.e., $|\Xi| > 1$).

In what follows, we always assume that $2 < |V_d| \le n$ (where $n = |V|$), namely, each document $d \in D$ has at least three distinct terms. Also, for an arbitrary term $t \in V$, we denote

$$p_d(t) = p(t|d) = \begin{cases} \frac{f_d(t)}{||d||} & t \in V_d\\ 0 & t \notin V_d\end{cases}$$

where $f_d(t)$ is the occurrence *frequency* of term t in d and $||d|| = \sum_{t\in V_d} f_d(t)$ as the *length* of d.

A. Estimate in a Single Document

Suppose each document d is represented by a $1 \times n$ frequency matrix

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

in which, each element in the matrix satisfies $f_d(t) > 0$ when $t \in V_d$ and $f_d(t) = 0$ when $t \in V - V_d$.

Then, for an arbitrary term $t \in V$, introduce an estimation function:

$$\psi_d(t) = \begin{cases} \frac{f_d(t)}{\sum_{t'\in V_d} f_d(t')} & t \in V_d\\ 0 & t \notin V_d\end{cases} \tag{7}$$

Clearly, we have $0 < \psi_d(t) < 1$ for every $t \in V_d \subseteq V$. Next, for an arbitrary given term $t \in V_d$, define a probability distribution by expression (3):

$$\begin{aligned} P_d(\delta = 1) &= \psi_d(t) = p_d(t)\\ P_d(\delta = 0) &= 1 - p_d(t)\end{aligned} \tag{8}$$

The function $\psi_d(t)$ and distribution $P_d(\delta)$ will be used in the three methods below.

A.1 Method One

For two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_d(t_i,t_j) = \begin{cases} \frac{f_d(t_i)f_d(t_j)}{\varpi} & (t_i,t_j) \in V_d \times V_d\\ 0 & (t_i,t_j) \notin V_d \times V_d\end{cases} \tag{9}$$

where the denominator of γ_d is,

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

which is the sum of all the possible products $f_d(t_{i'})f_d(t_{j'})$ for $i' < j'$; $i', j' \in \{1, 2, ..., n\}$, Clearly, as $|V_d| \geq 3$, we have $0 < \gamma_d(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_d \times V_d \subseteq V \times V$.

Next, for two arbitrary given terms $t_i, t_j \in V_d$ (where $i \neq j$), define a probability distribution by expression (4):

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 1) &= \frac{f_d(t_i)f_d(t_j)}{\varpi} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{f_d(t_i)}{||d||} - \gamma_d(t_i, t_j) \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{f_d(t_j)}{||d||} - \gamma_d(t_i, t_j) \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) \end{aligned} \tag{10}$$

In order to verify the constraints given in (2):

$$\gamma_d(t_i, t_j) \leq \psi_d(t_i), \psi_d(t_j)$$

for an arbitrary $t \in V_d$, let us denote

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

Study [15] has proven, for the functions $\psi_d(t)$ and $\gamma_d(t_i, t_j)$ given in (7) and (9), respectively, we have:

- $\varpi_{t_i} \geq f_d^2(t_i)$ if and only if $\psi_d(t_i) \geq \gamma_d(t_i, t_j)$;
- $\varpi_{t_j} \geq f_d^2(t_j)$ if and only if $\psi_d(t_j) \geq \gamma_d(t_i, t_j)$.

Thus we can write immediately the following theorem [15].

Theorem 4.1 The expression, $P_d(\delta_i, \delta_j)$, defined in (10) is a probability distribution if $\varpi_{t_i} \geq f_d^2(t_i)$ and $\varpi_{t_j} \geq f_d^2(t_j)$.

The above theorem tells us, when the estimation functions given in (7) and (9) are used, that $P_d(\delta_i, \delta_j)$ given in (10) is a probability distribution if two conditions $\varpi_{t_j} \geq f_d^2(t_j)$ and $\varpi_{t_i} \geq f_d^2(t_i)$ are satisfied simultaneously. The conditions can also be verified by $p_d(t_i) \geq \gamma_d(t_i, t_j)$ and $p_d(t_j) \geq \gamma_d(t_i, t_j)$, respectively, which may be easier to compute in practical application. Next, we give the property of the MIT measures by the following corollary.

Corollary 4.1 For the four MIT measures derived from expressions (8) and (10), four inequalities,

$$\begin{aligned} \mathbf{mit}_d(t_i, t_j) > 0, \quad \mathbf{mit}_d(t_i, \bar{t}_j) \leq 0 \\ \mathbf{mit}_d(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_d(\bar{t}_i, \bar{t}_j) > 0 \end{aligned}$$

always hold if $\varpi_{t_j} \geq f_d^2(t_j)$ and $\varpi_{t_i} \geq f_d^2(t_i)$.

Proof: By Theorem 4.1, $P_d(\delta_i, \delta_j)$ given in (10) is a probability distribution for terms $t_i, t_j \in V_d$. Also,

$$\begin{aligned} \varpi &= \sum_{i'<j';\ t_{i'}, t_{j'} \in V_d} f_d(t_{i'})f_d(t_{j'}) \\ &< \sum_{t_{i'}, t_{j'} \in V_d} f_d(t_{i'})f_d(t_{j'}) = ||d||^2 \end{aligned}$$

from which we have

$$\gamma_d(t_i, t_j) = \frac{f_d(t_i)f_d(t_j)}{\varpi} > \frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||} = p_d(t_i)p_d(t_j)$$

Thus, from (a) of Theorem 3.1, four inequalities hold. □

A.2 Method Two

Note that $f_d(t)$ is the number of time(s) that term t occurs in d and that $f_d(t_1) + f_d(t_2) + ... + f_d(t_n) = ||d||$. Thus, the probability that two distinct terms t_i and t_j are simultaneously found in d should be

$$\begin{aligned} &\frac{C_{f_d(t_i)}^1 C_{f_d(t_j)}^1}{C_{||d||}^2} \\ &= \frac{[f_d(t_i)]!}{1![f_d(t_i) - 1]!}\frac{[f_d(t_j)]!}{1![f_d(t_j) - 1]!} \Big/ \frac{||d||!}{2!(||d|| - 2)!} \\ &= \frac{2f_d(t_i)f_d(t_j)}{||d|| \cdot (||d|| - 1)} \end{aligned}$$

Hence, for two arbitrary distinct terms $t_i, t_j \in V$, introduce an estimation function:

$$\gamma_d(t_i, t_j) = \begin{cases} \frac{2f_d(t_i)f_d(t_j)}{||d|| \cdot (||d|| - 1)} & (t_i, t_j) \in V_d \times V_d \\ 0 & (t_i, t_j) \notin V_d \times V_d \end{cases} \tag{11}$$

which satisfies $0 < \gamma_d(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_d \times V_d \subseteq V \times V$ as $|V_d| \geq 3$.

Next, for two arbitrary given terms $t_i, t_j \in V_d$ (where $i \neq j$), define a probability distribution by (4):

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 1) &= \frac{2f_d(t_i)f_d(t_j)}{||d|| \cdot (||d|| - 1)} = \gamma_d(t_i, t_j) \\ P_d(\delta_i = 1, \delta_j = 0) &= \frac{f_d(t_i)}{||d||}\Big(1 - \frac{2f_d(t_j)}{||d|| - 1}\Big) \\ &= p_d(t_i) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 1) &= \frac{f_d(t_j)}{||d||}\Big(1 - \frac{2f_d(t_i)}{||d|| - 1}\Big) \\ &= p_d(t_j) - \gamma_d(t_i, t_j) \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) \end{aligned} \tag{12}$$

We may give two conditions of $P_d(\delta_i, \delta_j)$, such that it satisfies the constraints given in (2) by the following theorem.

Theorem 4.2 The expression, $P_d(\delta_i, \delta_j)$, defined in (12) is a probability distribution if $f_d(t_i) \leq \frac{||d||-1}{2}$ and $f_d(t_j) \leq \frac{||d||-1}{2}$.

Proof: From $f_d(t_j) \leq \frac{||d||-1}{2}$, we have $1 \geq \frac{2f_d(t_j)}{||d||-1}$, that is,

$$p_d(t_i) = \frac{f_d(t_i)}{||d||} \geq \frac{2f_d(t_i)f_d(t_j)}{||d|| \cdot (||d|| - 1)} = \gamma_d(t_i, t_j)$$

A similar proof can be applied to $p_d(t_j) \geq \gamma_d(t_i, t_j)$. □

Next, we can give the property of the MIT measures by the following corollary.

Corollary 4.2 For the four MIT measures derived from expressions (8) and (12), four inequalities,

$$\mathbf{mit}_d(t_i,t_j)>0,\quad \mathbf{mit}_d(t_i,\bar{t}_j)\le 0$$
$$\mathbf{mit}_d(\bar{t}_i,t_j)\le 0,\quad \mathbf{mit}_d(\bar{t}_i,\bar{t}_j)>0$$

always hold if $f_d(t_i)\le\frac{||d||-1}{2}$ and $f_d(t_j)\le\frac{||d||-1}{2}$.

Proof: By Theorem 4.2, $P_d(\delta_i,\delta_j)$ given in (12) is a probability distribution for terms $t_i,t_j\in V_d$. Also, we have $||d||\cdot(||d||-1)<||d||\cdot||d||$, thus,

$$\gamma_d(t_i,t_j)=\frac{2f_d(t_i)f_d(t_j)}{||d||(||d||-1)}>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-1}$$
$$>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||}=p_d(t_i)p_d(t_j)$$

Hence, from (a) of Theorem 3.1, the four inequalities hold. □

A.3 Method Three

The probability that term t_j is found in d after term t_i has been found in d, where $i\neq j$, should be

$$P_d(\delta_j=1|\delta_i=1)=\frac{f_d(t_j)}{||d||-f_d(t_i)}$$

Thus, for two arbitrary distinct terms $t_i,t_j\in V$, introduce an estimation function:

$$\gamma_d(t_i,t_j)=\begin{cases}\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-f_d(t_i)} & (t_i,t_j)\in V_d\times V_d\\ 0 & (t_i,t_j)\notin V_d\times V_d\end{cases}\quad(13)$$

which satisfies $0<\gamma_d(t_i,t_j)<1$ for every $(t_i,t_j)\in V_d\times V_d\subseteq V\times V$ as $|V_d|\ge 3$.

Next, for two arbitrary given terms $t_i,t_j\in V_d$ (where $i\neq j$), define a probability distribution by (4):

$$\begin{aligned}P_d(\delta_i=1,\delta_j=1)&=\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-f_d(t_i)}=\gamma_d(t_i,t_j)\\ P_d(\delta_i=1,\delta_j=0)&=\frac{f_d(t_i)}{||d||}\left(1-\frac{f_d(t_j)}{||d||-f_d(t_i)}\right)\\ &=p_d(t_i)-\gamma_d(t_i,t_j)\\ P_d(\delta_i=0,\delta_j=1)&=\frac{f_d(t_j)}{||d||}\left(1-\frac{f_d(t_i)}{||d||-f_d(t_j)}\right)\\ &=p_d(t_j)-\gamma_d(t_i,t_j)\\ P_d(\delta_i=0,\delta_j=0)&=1-p_d(t_i)-p_d(t_j)+\gamma_d(t_i,t_j)\end{aligned}\quad(14)$$

We need to find out if there exists any verification condition, such that $P_d(\delta_i,\delta_j)$ satisfies the constraints given in (2), by the following theorem.

Theorem 4.3 The expression, $P_d(\delta_i,\delta_j)$, defined in (14) is a probability distribution.

Proof: Notice that $f_d(t_i)+f_d(t_j)\le||d||$. Thus,

$$1\ge\frac{f_d(t_j)}{||d||-f_d(t_i)}$$

that is,

$$p_d(t_i)=\frac{f_d(t_i)}{||d||}>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-f_d(t_i)}=\gamma_d(t_i,t_j)$$

A similar proof can be applied to $p_d(t_j)>\gamma_d(t_i,t_j)$. □

It is clear, unlike Methods 1 and 2, that $P_d(\delta_i,\delta_j)$ in (14) is a probability distribution unconditionally. Next, we give the property of the MIT measures by the following corollary.

Corollary 4.3 For the four MIT measures derived from expressions (8) and (14), four inequalities

$$\mathbf{mit}_d(t_i,t_j)>0,\quad \mathbf{mit}_d(t_i,\bar{t}_j)\le 0$$
$$\mathbf{mit}_d(\bar{t}_i,t_j)\le 0,\quad \mathbf{mit}_d(\bar{t}_i,\bar{t}_j)>0$$

always hold for arbitrary terms $t_i,t_j\in V_d$.

Proof: By Theorem 4.3, $P_d(\delta_i,\delta_j)$ given in (14) is a probability distribution for terms $t_i,t_j\in V_d$. Also, from $||d||-f_d(t_i)<||d||$, we have,

$$\gamma_d(t_i,t_j)=\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-f_d(t_i)}>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||}$$
$$=p_d(t_i)p_d(t_j)$$

Hence, from (a) of Theorem 3.1, the four inequalities hold. □

B. Estimate in a Set of Documents

The above three estimation methods consistently use frequency representation for the individual documents. However, in some probabilistic methods, one would state that the *binary assumption* suffices to specify the dependence of terms. The method discussed here is under this assumption.

By 'binary' it is here meant that each document $d\in D$ is represented by a $1\times n$ matrix:

$$\mathbf{m}_d=\left[t_1^{\delta_1},t_2^{\delta_2},...,t_n^{\delta_n}\right]=\left[t^{\delta}\right]_{1\times n}$$

in which, each element in the matrix is a binary number satisfying $t^{\delta}=1$ when $t\in V_d$ and $t^{\delta}=0$ when $t\in V-V_d$.

Consider a sample set Ξ, satisfying $|\Xi|>1$. Denote $n_{\Xi}(t)$ as the number of documents in Ξ in which term t occurs, and $n_{\Xi}(t_i,t_j)$ as the number of documents in Ξ in which terms t_i and t_j co-occur. It can be easily seen $n_{\Xi}(t_i,t_j)\le n_{\Xi}(t_i),n_{\Xi}(t_j)\le|\Xi|$

Then, for an arbitrary term $t\in V$, introduce an estimation function:

$$\psi_{\Xi}(t)=\begin{cases}\frac{n_{\Xi}(t)}{|\Xi|} & t\in V_{\Xi}\\ 0 & t\notin V_{\Xi}\end{cases}\quad(15)$$

Obviously, we have $0<\psi_{\Xi}(t)<1$ for every $t\in V_{\Xi}\subseteq V$. Next, for an arbitrary given term $t\in V_d$, define a probability distribution by expression (3):

$$\begin{aligned}P_{\Xi}(\delta=1)&=\psi_{\Xi}(t)\\ P_{\Xi}(\delta=0)&=1-\psi_{\Xi}(t)\end{aligned}\quad(16)$$

The function $\psi_{\Xi}(t)$ and distribution $P_{\Xi}(\delta)$ will be used in the fourth method below.

B.1 Method Four

For two arbitrary distinct terms $t_i,t_j\in V$, introduce an estimation function:

$$\gamma_{\Xi}(t_i,t_j)=\begin{cases}\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} & (t_i,t_j)\in V_{\Xi}\times V_{\Xi}\\ 0 & (t_i,t_j)\notin V_{\Xi}\times V_{\Xi}\end{cases}\quad(17)$$

which satisfies $0 < \gamma_{\Xi}(t_i, t_j) < 1$ for every $(t_i, t_j) \in V_{\Xi} \times V_{\Xi} \subseteq V \times V$ as $|V_d| \geq 3$.

Next, for two arbitrary given terms $t_i, t_j \in V_{\Xi}$ (where $i \neq j$), define a probability distribution by expression (4):

$$\begin{aligned}
P_{\Xi}(\delta_i = 1, \delta_j = 1) &= \frac{n_{\Xi}(t_i, t_j)}{|\Xi|} = \gamma_{\Xi}(t_i, t_j) \\
P_{\Xi}(\delta_i = 1, \delta_j = 0) &= \frac{n_{\Xi}(t_i) - n_{\Xi}(t_i, t_j)}{|\Xi|} \\
&= \psi_{\Xi}(t_i) - \gamma_{\Xi}(t_i, t_j) \\
P_{\Xi}(\delta_i = 0, \delta_j = 1) &= \frac{n_{\Xi}(t_j) - n_{\Xi}(t_i, t_j)}{|\Xi|} \\
&= \psi_{\Xi}(t_j) - \gamma_{\Xi}(t_i, t_j) \\
P_{\Xi}(\delta_i = 0, \delta_j = 0) &= 1 - \psi_{\Xi}(t_i) - \psi_{\Xi}(t_j) + \gamma_{\Xi}(t_i, t_j)
\end{aligned} \tag{18}$$

It is interesting to note that $\psi_{\Xi}(t_i), \psi_{\Xi}(t_j) \geq \gamma_{\Xi}(t_i, t_j)$ as

$$\frac{n_{\Xi}(t_i)}{|\Xi|}, \frac{n_{\Xi}(t_j)}{|\Xi|} \geq \frac{n_{\Xi}(t_i, t_j)}{|\Xi|}$$

for arbitrary $t_i, t_j \in V_{\Xi}$. Hence the estimation functions given in (15) and (17) satisfy the constraints given in (2) and, thus we can give the following theorem.

Theorem 4.4 The expression, $P_d(\delta_i, \delta_j)$, defined in (18) is a probability distribution.

Like Method 3, $P_{\Xi}(\delta_i, \delta_j)$ given in (18) is a probability distribution unconditionally. From Theorem 4.4, we may give the properties of the MIT measures by the following corollary.

Corollary 4.4 For the four MIT measures derived from expressions (16) and (18),

(a') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} > \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
&\mathbf{mit}_{\Xi}(t_i, t_j) > 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \leq 0 \\
&\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \leq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) > 0
\end{aligned}$$

(b') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} = \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
&\mathbf{mit}_{\Xi}(t_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) = 0 \\
&\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) = 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) = 0
\end{aligned}$$

(c') if $\frac{n_{\Xi}(t_i,t_j)}{|\Xi|} < \frac{n_{\Xi}(t_i)}{|\Xi|}\frac{n_{\Xi}(t_j)}{|\Xi|}$ then

$$\begin{aligned}
&\mathbf{mit}_{\Xi}(t_i, t_j) < 0, \quad \mathbf{mit}_{\Xi}(t_i, \bar{t}_j) \geq 0 \\
&\mathbf{mit}_{\Xi}(\bar{t}_i, t_j) \geq 0, \quad \mathbf{mit}_{\Xi}(\bar{t}_i, \bar{t}_j) < 0
\end{aligned}$$

The Method 4 is the most commonly used in many areas, such as, information retrieval, natural language processing, document classification, sentiment analysis, and many related areas. More discussion on this method, including its properties and potential application problems, can also be found in [17].

V. Discussion

Some key points, which are helpful to understand the methods proposed under the general framework, are addressed in this section. These key points are also important to guide practical applications.

First, it should be possible, though it may not be easy, to construct a variety of estimation functions and then to define probability distributions and verify the corresponding constraints for formalising the MIT measures. For suitable choices of the estimation functions practically appropriate for and mathematically meaningful to a specific application problem, the term state distributions, when substituted into measures, $\mathbf{mit}_{\Xi}(t_i^{\delta_i}, t_j^{\delta_j})$ $(\delta_i, \delta_j = 0, 1)$ and/or $I(\delta_i; \delta_i)$, can be expected to capture the mutual information of terms. The information may be used to develop a variety of techniques in order to assign dependence values to individual term pairs and, then some decision is made on the values. A summary of the four example estimation methods proposed in this study is given in Table I. It is important to understand that the MIT measures formalised by different estimation methods may have entirely different properties. For instance, let us return to the four example estimations discussed in Section IV and consider an inequality,

$$\gamma_d(t_i, t_j) > p_d(t_i)p_d(t_j) = \psi_d(t_i)\psi_d(t_j)$$

Then some key points regarding the properties and relationships of the MIT measures of the four corresponding Methods 1–4 can be made below.

- Theorems/Corollaries 4.1–4.3 in respective Methods 1–3 tell us, when estimation functions (7), (9), (11) and (13) are used, that the above inequality always holds, and that terms co-occurring in document d must be more or less statistically dependent since it is always $\mathbf{mit}_d(t_i, t_j) > 0$ supporting a dependence assertion.
- Theorem/Corollary 4.4 in Method 4 tells us, when estimation functions (15) and (17) are used, that the above inequality does not always hold, and that terms may or may not be statistically dependent for their co-occurrence since the sign of $\mathbf{mit}_{\Xi}(t_i, t_j)$ might be different from term pair to term pair.

Therefore, we can learn from the Theorems/Corollaries: for two terms making the above inequality hold, some estimation functions ensure them to be more or less dependent for their co-occurrence, whereas other estimation functions cannot guarantee them to be dependent for their co-occurrence. This also clearly indicates, for the same term pairs, that different estimation methods may result in entirely different conclusions regarding the statistical dependence for their co-occurrence.

Second, as we all knew, the MIT measures may influence experimental performance significantly. However, as the probability distributions are normally obtained according to practical application, it seems that only the "form" of the mutual information measure has frequently been the main concern of research in literature, whereas the problem of verification of the probability distributions is often ignored as

TABLE I
A SUMMARY OF THE FOUR EXAMPLE ESTIMATIONS

Method	Function $\psi(t)$	Function $\gamma(t_i, t_j)$	Conditions
1	$\psi_{_d}(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_{_d}(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\varpi}$	$f_d^2(t_i) \leq \varpi_{_{t_i}}$, $f_d^2(t_j) \leq \varpi_{_{t_j}}$
2	$\psi_{_d}(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_{_d}(t_i, t_j) = \frac{2 f_d(t_i) f_d(t_j)}{\|\|d\|\| \cdot (\|\|d\|\| - 1)}$	$f_d(t_i) \leq \frac{\|\|d\|\| - 1}{2}$, $f_d(t_j) \leq \frac{\|\|d\|\| - 1}{2}$
3	$\psi_{_d}(t) = \frac{f_d(t)}{\|\|d\|\|}$	$\gamma_{_d}(t_i, t_j) = \frac{f_d(t_i)}{\|\|d\|\|} \frac{f_d(t_j)}{\|\|d\|\| - f_d(t_i)}$	none
4	$\psi_{_\Xi}(t) = \frac{n_{_\Xi}(t)}{\|\Xi\|}$	$\gamma_{_\Xi}(t_i, t_j) = \frac{n_{_\Xi}(t_i, t_j)}{\|\Xi\|}$	none

an unimportant matter. This implicitly means that a function with a form

$$i(x_1, x_2) = P(x_1, x_2) \log \frac{P(x_1, x_2)}{P(x_1) P(x_2)}$$

would be a "mutual information measure" of x_1 and x_2 for their co-occurrence, and that the discussion on the three criteria of $P(x)$ and $P(x_1, x_2)$ in the function are trivial. This is indeed not true. It is important to realise that it is not necessarily that the function, $i(x_1, x_2)$, is a mutual information measure. In fact, $i(x_1, x_2)$ is not a mutual information measure in the information-theoretic sense, if $P(x)$ and $P(x_1, x_2)$ are not probability distributions and/or, if $P(x_1)$ and $P(x_2)$ are not marginal distributions of the joint distribution $P(x_1, x_2)$ (even though they may be all probability distributions). It may not even converge if $P(x_1, x_2) \ll f_1(x_1)$ and $P(x_1, x_2) \ll P(x_2)$ do not hold. Therefore, in practical applications, it entirely makes no sense to use some function, looking like a mutual information measure, to compute the mutual information of terms when any one of the three criteria is not satisfied. We emphasize that the verification of $P(\delta)$ and $P(\delta_i, \delta_j)$ meeting the three criteria is the major premise when applying $I(\delta_i; \delta_j)$ to effectively capture the mutual information inherent among terms. A simple but interesting example given in our related study [15] may clarify our idea. We here give a brief explanation and details of computation can be found in [15]. Suppose we are given a document $d = \{t_1, t_2, t_2, t_2, t_3, t_4\} \in D$. This example considers the estimation functions given in Method 1 and illustrates a specific instance of failing to apply them for two terms $t_1, t_2 \in V_d$:

$$\varpi = \sum_{i' < j';\ t_{i'}, t_{j'} \in V_d} f_d(t_{i'}) f_d(t_{j'}) = 12$$

and, with expressions (7), (8), (9) and (10), we have $\gamma_{_d}(t_1, t_2) = \frac{1}{4}$ and

$$P_d(\delta_1 = 1, \delta_2 = 0) = \psi_{_d}(t_1) - \gamma_{_d}(t_1, t_2) = -\frac{1}{12}$$

It can be easily seen, for the term pair (t_1, t_2), that the corresponding $P_d(\delta_1, \delta_2)$ is not a probability distribution since the constraints given in (2) are not satisfied (i.e., $\psi_{_d}(t_1) < \gamma_{_d}(t_1, t_2)$). Consequently, $P_d(\delta_1 = 1, \delta_2 = 1)$ is not reliable for measuring dependence of t_1 and t_2 for their co-occurrence. The key points regarding the probability distributions are:

- There may be many term pairs, of which the corresponding $P_d(\delta_i, \delta_j)$ is indeed a probability distribution. However, it is possible that not all term pairs have the corresponding probability distribution.
- In order to compute MIT of terms, we must verify the constraints given in (2), that is, we have to check both $\psi_{_d}(t_i) \geq \gamma_{_d}(t_i, t_j)$ and $\psi_{_d}(t_j) \geq \gamma_{_d}(t_i, t_j)$ to be satisfied simultaneously, for each of the term pairs considered.

Thus, those term pairs (rather than two individual terms), of which the corresponding $P_d(\delta_i, \delta_j)$ does not satisfy the constraints, should be discarded immediately and omitted from the computation of MIT.

Third, the estimation functions given in Methods 1-3 can be applied to document representations not only for $\mathbf{m}_{_d} = \big[f_d(t)\big]_{1 \times n}$, but also for a more general case, where each document d can be represented by a $1 \times n$ (weight) matrix:

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

in which, each element is a real number, 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*, which indicates the importance of term t in representing document d. For instance, the weighting function in Methods 1-3 is $w_d(t) = f_d(t)$. The key points regarding the estimation functions are below.

- Methods 1-3 should be applicable to any quantitative document representation.
- $\psi_{_d}(t)$ and $\gamma_{_d}(t_i, t_j)$ should be used to capture the information of occurrence and co-occurrence of terms.
- $w_d(t)$ should be the main component of the estimation functions, it is construed by means of occurrence frequencies and co-occurrence frequencies of terms.

The extension of, for instance, Method 1 can be found in another of our studies [15]. It is beyond the scope of the current paper to discuss the issue of document representation in greater detail, and some formal discussion and technical treatment can be found in, for instance studies [18]–[20].

Fourth, it is certainly true that the MIT measures given in Definition 3.1 can be used to measure the extent of dependence of terms t_i and t_j. Also, it is certainly true that the larger quantities the measures offer, the higher the extent term t_i is statistically dependent on term t_j (and vice versa). However, the *implications* of the dependence obtained from the individual MIT measures are different. Remember that we always emphasize 'the dependence *under the state value* (δ_i, δ_j)'. This emphasis is necessary because it clearly indicates that it is the state value (δ_i, δ_j) that supports the dependence. For instance,

terms t_i and t_j may depend highly on one another, when t_i occurs but t_j does not occur in some document and, in this case, we are talking about the dependence under the state value $(\delta_i, \delta_j) = (1, 0)$. In a practical application, what we generally concentrate on is 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 constraints given in (2). For instance, for Method 1, we need only use the measure $\mathbf{mit}_d(t_i, t_j)$ and verify the condition:

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

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

Fifth, from a high expected mutual information value, we cannot state immediately that state value $(\delta_i, \delta_j) = (1, 1)$ makes a larger contribution to $I_{\Xi}(\delta_i; \delta_j)$ and, thus we cannot assert that terms t_i and t_j are highly dependent for their co-occurrence in Ξ. This is because, with the relations of the MIT measures learned from their signs, when $\gamma_{\Xi}(t_i, t_j) < \psi_{\Xi}(t_i)\psi_{\Xi}(t_j)$, the positive value $I_{\Xi}(\delta_i; \delta_j)$ will be dominated by the positive quantities $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$. In this case, the higher value the $I_{\Xi}(\delta_i; \delta_j)$ has, the larger quantities the $\mathbf{mit}_{\Xi}(t_i, \bar{t}_j)$ and $\mathbf{mit}_{\Xi}(\bar{t}_i, t_j)$ provide, the more it is indicated that t_i and t_j are highly dependent under state values $(1, 0)$ and $(0, 1)$ and that they should not co-occur in Ξ. We can clarify our viewpoint by an example given in [17]. Let us consider Method 4 and suppose $\Xi = \{d_1, d_2, d_3\}$, $V_{d_1} = \{t_1, t_2, t_3, t_4, t_5\}$, $V_{d_2} = \{t_1, t_4, t_5, t_7\}$ and $V_{d_3} = \{t_4, t_7, t_8\}$. Then, we have: $n_{\Xi}(t_1) = 2$, $n_{\Xi}(t_2) = 1$ and $n_{\Xi}(t_1, t_2) = 1$; $n_{\Xi}(t_5) = 2$, $n_{\Xi}(t_7) = 2$ and $n_{\Xi}(t_5, t_7) = 1$. Thus, we obtain (details of computation can be found in [17])

$$I_{\Xi}(\delta_1; \delta_2) \approx 0.1352 - 0.0959 - 0.0000 + 0.1352 = 0.1745,$$
$$I_{\Xi}(\delta_5; \delta_7) \approx -0.0959 + 0.1352 + 0.1352 - 0.0000 = 0.1745.$$

Clearly, the positive value of $I_{\Xi}(\delta_1; \delta_2)$ is dominated by both quantities $\mathbf{mit}_{\Xi}(t_1, t_2)$ and $\mathbf{mit}_{\Xi}(\bar{t}_1, \bar{t}_2)$, and t_1 and t_2 are highly dependent for their co-occurrence and co-not-occurrence in set Ξ; the positive value of $I_{\Xi}(\delta_5; \delta_7)$ is dominated by both $\mathbf{mit}_{\Xi}(t_5, \bar{t}_7)$ and $\mathbf{mit}_{\Xi}(\bar{t}_5, t_7)$, and t_5 and t_7 are highly dependent for their not co-occurrence in set Ξ. In addition, from this example, we can see that term pairs (t_1, t_2) and (t_5, t_7) have the same expected mutual information and, however, that the implications of for the individual state values are entirely different: Terms t_1 and t_2 provide the information highly supporting for both their co-occurrence and co-not-occurrence; whereas terms t_5 and t_7 provide the information highly supporting for occurrence of one but not occurrence of the other. It should be repeatedly pointed out that all the five different measures, the four MIT measures and the EMIM measure, may give us useful information, but each tells us different aspects about the dependences of terms and, in particular, it is likely that $I_{\Xi}(\delta_i; \delta_j)$ tells us nothing about the dependences of terms for their co-occurrence.

Sixth, it is worth mentioning that many studies use the following formula:

$$I(t_i; t_j) = P(t_i, t_j) \log \frac{P(t_i, t_j)}{P(t_i)P(t_j)}$$

to estimate the mutual information of terms t_i and t_j. It is 'equivalent' to the MIT measure for the state value $(\delta_i, \delta_j) = (1, 1)$ given in Definition 3.1,

$$\mathbf{mit}(t_i, t_j) = \mathbf{mit}(t_i^{\delta_i=1}, t_j^{\delta_j=1})$$
$$= P(\delta_i = 1, \delta_j = 1) \log \frac{P(\delta_i = 1, \delta_j = 1)}{P(\delta_i = 1)P(\delta_j = 1)}$$

as we denote $t^{\delta} = t$, $\bar{t}$ when $\delta = 1, 0$, respectively. The expression $I(t_i; t_j)$ seems simpler to that of $\mathbf{mit}(t_i, t_j)$. However, we point out, mathematically, that $\mathbf{mit}(t_i, t_j)$ is more appropriate and clearer than $I(t_i; t_j)$ from, for instance, a viewpoint of the probability space: It is obvious to see that $P(\delta_i, \delta_j)$ is over $\Omega \times \Omega$ as its each argument $\delta \in \Omega = \{0, 1\}$, whereas it is easy to cause confusion that $P(t_i, t_j)$ is over $V \times V$ as each of its arguments has a domain $t \in V = \{t_1, t_2, ..., t_n\}$ (rather than $t \in \{0, 1\}$). Also, $I(\cdot\,;\cdot)$, when used to expressed EMIM, is a traditional mathematical symbol, which is the summation of four items (rather than only one) corresponding to four state value pairs of each term pair.

Seventh, it is worth mentioning that there are five information measures widely used in the literature for computing term dependence (or, relatedness): directed divergence [1], divergence [1], information radius [21], Jensen difference [22] and the expected mutual information (i.e., EMIM, which is regarded as a special case of directed divergence) [1]. The five measures, which are what are generally called *information gain*, are by now familiar to many researchers. A detailed account of the concept of the measures is given in [1], and an axiomatic characterization can be found in [23]. The five measures are examined in our series of studies: Study [19] develops the measurement of term relatedness using the information radius measure, demonstrates how the relatedness measures may deal with some basic concepts of applications, and summarizes important features of, and differences between, the information radius measure and the first two information measures (directed divergence and divergence), from a practical perspective. Study [18] addresses the measurement of term relatedness based on the Jensen difference measure and points out, when Shannon entropy is used, that the Jensen difference measure is in fact the information radius measure, and that some formal methods proposed in many past studies in terms of these two measures are in principle the same matter. Study [15] proposes a method for estimating probability distributions required in EMIM, and provides examples to illustrate the possibility of failure of applying this method if the verification conditions are not satisfied. Study [17] reconsiders the $emim$ measure, which is widely used in applications, derived from simplifying EMIM

under a binary assumption, and discusses some potential but important problems of applying the $emim$ measure. Study [20] attempts to establish a unified theoretical framework for applying several information measures to the measurement of term discrimination information and to define relatedness measures according to the discrimination measures, and then discusses some potential problems arising from using the relatedness measures and suggests solutions.

Finally, we would like to point out that the current study is further work of study [15], [17]: it focuses on the establishment of a general framework for constructing estimation functions in order to define probability distributions required in EMIM for effectively distinguishing potentially dependent term pairs from many others. As this paper concentrates on a formal analysis and discussion, the reader interested in how the mutual information methods, as well as other information measures' methods, may be supported by empirical evidence drawn from a number of performance experiments is referred to those papers referenced.

VI. Conclusions

This study focused on the establishment of a general framework for defining probability distributions required in EMIM, which is crucial and remains an open issue, for effectively distinguishing potentially dependent term pairs from many others. Under the framework,

- the general forms of estimation functions with a set of constraints were introduced;
- the general forms of probability distributions under term state values were defined;
- the general form of MIT measures for computing the mutual information of terms was formalised;
- the general properties of the MIT measures were studied and the general relations between the MIT measures were revealed.

Four estimation methods were proposed to clarify and illustrate our ideas presented in this study by

- interpreting the mathematical meanings of the estimation functions within practical application contexts;
- discussing verification conditions for satisfying the constraints in order to ensure that probability distributions meet the three criteria;
- presenting the properties and relationships of the MIT measures given in the individual methods.

The key points of this study were pointed out and emphasised, some of them are:

- The different implications of the dependence obtained from the individual MIT measures and the EMIM measure should be carefully distinguished from one another.
- The estimation functions should be constructed using weighting functions capable of capturing the occurrence and co-occurrence information of terms.
- It is possible of failure of using the estimation functions to define probability distributions if the constraints are not satisfied.

Under the general framework, the probability distributions, when defined from the estimation functions satisfying the constraints, will meet the three criteria. Thus, the issue of defining the probability distributions becomes the issue of constructing the estimation functions and verifying the constraints, which is relatively simple for practical applications. Due to its generality, the general framework is applicable to many areas of science, involving statistical semantic analysis of features (concepts, terms, phrases, words, etc.) and quantitative representations of objects (documents, abstracts, sentences, queries, etc.).

References

[1] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
[2] C. E. Shannon, "A mathematical theory of communication," *Bell System and Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.
[3] A. Akadi, A. 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, no. 4, pp. 116–121, 2008.
[4] 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. 31, no. 10, pp. 1168–1174, 2010.
[5] 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.
[6] 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.
[7] G. Wang, F. Lochovsky, and Q. Yang, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2004, pp. 342–349.
[8] N. Vretos, V. Solachidis, and I. Pitas, "A mutual information based face clustering algorithm for movies," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo (ICME'06)*, 2006, pp. 1013–1016.
[9] 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.
[10] 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.
[11] H. Fang and C. X. Zhai, "Semantic term matching in axiomatic approaches to information retrieval," in *Proceedings of the 29th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 115–122.
[12] S. Gauch, J. Wang, and S. M. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 250–269, 1999.
[13] 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.
[14] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, 2000.
[15] D. Cai and T. McCluskey, "A simple method for computing term mutual information," *Journal of Computing*, vol. 4, no. 6, pp. 1–6, 2012.
[16] 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.
[17] D. Cai, "Reconsideration of potential problems of applying EMIM measure for text analysis," *International Journal of Advanced Computer Science and Applications (accepted)*.

[18] ——, "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.
[19] D. Cai and C. J. van Rijsbergen, "Learning semantic relatedness from term discrimination information," *Expert Systems with Applications*, vol. 40, no. 1, 2008.
[20] 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.
[21] R. Sibson, "Information radius," *Z. Wahrsch'theorie and verw. Geb*, vol. 14, pp. 149–160, 1969.
[22] C. R. Rao, "Diversity: Its measurement, decomposition, apportionment and analysis," *Sankhya: Indian Journal of Statistics*, vol. 44, pp. 1–22, 1982.
[23] A. Rényi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1961, pp. 547–561.