語と文書の共起に基づく特徴度の数量的表現について

相澤彰子†

本論文では語と文書の共起関係に注目し，与えられた文書集合中での語の特徴度の量的表現やその適用について，情報量的な観点から考察を加える．今日，情報検索の分野において広く用いられている $tf{\cdot}idf$（term frequency - inverse document frequency）は，語頻度と対数文書頻度の逆数を乗じた尺度である．ここで tf を語の総出現頻度で正規化した値は，語の出現確率の推定値に対応しており，さらに idf は一種の情報量として解釈できることから，$tf{\cdot}idf$ は確率と情報量をかけあわせた尺度であるといえる．本論文では，このような $tf{\cdot}idf$ の定義を拡張して，語の特徴度を,「語の出現確率」と「語の持つ情報量」の積の形で一般的に定義し，実際のテキストデータに適用した結果を示す．

On the Quantitative Representation of Term Specificity Based on Terms and Documents Co-occurrences

Akiko Aizawa†

This paper presents a mathematical definition of the *feature quantity*, a measure of specificity of terms in documents which is based on an information theoretic view of retrieval events. The proposed feature quantity is expressed as a product of the frequency of terms and their amouts of information, and has a good correspondence with $tf{\cdot}idf$-like measures commonly used in today's information retrieval systems. In the paper, the mathemtaical definition of the feature quantity is shown together with some illustrative examples.

1. まえがき

本論文では，与えられた文書集合中での語の特徴度の量的表現やその適用について，情報量的な観点から考察を加える．語の特徴度（term specificity）の数量的な尺度に関連する研究分野として，計算的語彙論における自動用語抽出およびテキスト分類における特徴語選択の 2 つをあげることができる．用語抽出の分野における特徴度は，n グラムや形態素解析などにより抽出された語単位の中から，与えられた文書集合を代表する基本語彙を自動識別するために用いられる．このための統計尺度として，頻度[1)]，idf[2)]，相互情報量[3)]，カイ 2 乗値[4)]，対数尤度比[4)]，$tf{\cdot}idf$[5)]，$representativeness$[6)]などが存在する[7),8),☆]．

テキスト分類の分野における特徴語選択は主に前処理として用いられており，あらかじめ語数を削減することで機械学習アルゴリズムを効率的に適用し，さらに過学習を回避することを目的としている．テキスト分類において一般的な尺度として，文書頻度[9)]，情報利得[10),11)]，カイ 2 乗値[11),12)]，Odds Ratio[13)]，平均クロスエントロピー（expected cross entropy）[14),15)]などが存在する[9),16)]．

一方，語の特徴の度合いを数量化した類似の尺度として，情報検索における語の重み（term weights）がある☆☆．特定の文書に注目して，その文書に含まれる語の重要度を数値で表現するもので，頻度[18)]，idf[19)]，信号雑音比[5)]，ベクトル空間モデルにおける $tf{\cdot}idf$[20)]

† 国立情報学研究所
National Institute of Informatics

☆ 用語抽出の分野では，近傍で共起する語と語の共起関係に注目して語の特徴度を定義する場合が多い．これに対して，テキスト分類の分野ではカテゴリと語の共起関係に，情報検索の分野では文書と語の共起関係に，主に注目して語の特徴度を定義する．本論文では，これらの共起関係が頻度行列という同一の形式で表されることを前提として，共起頻度が与えられた場合に，特徴度を数量化するための統計尺度に焦点をあてて議論を進める．

☆☆ 文献 17) では，term specificity を検索文から検索語を選別するための尺度，term weights を適合文書中に出現する語の重み付け尺度として用いている．これに対して用語抽出の分野における specificity は，たとえば文献 1) で用いられているように，語が意味階層の下位にあることを示唆している．本論文では，与えられた文書集合中の文書どうしを識別するための手がかりとしての語の有用性を term specificity，特定の文書を特徴づける語の重みを term weights と呼び区別するが，用語抽出の観点からは，前者は文献 6) の representativeness（分野代表性）に対応すると考えられる．

表 1 語の特徴を表す数量的尺度の分類と例
Table 1 Classification and examples of quantitative measures to express the specificity of terms.

	「網羅性」の尺度	「特定性」の尺度	「識別性」の尺度	「代表性」の尺度
語の特徴度（文書集合に注目）	例）文書集合中での総出現回数	例）idf, 信号雑音比	例）情報利得	例）総頻度 × idf
語の重み（特定の文書に注目）	例）語の文書内頻度	例）相互情報量	例）確率型検索モデルの重み付け法	例）文書内頻度 × idf

およびその多くのバリエーション，確率型検索モデルにおける P あるいは P/A 重み付け法[19),21)]などが代表的なものとしてあげられる[8),22),23),☆]．

本論文ではまず，理論的な観点に基づき，これらの尺度を以下の 4 つのタイプに分類する（**表 1**）．

(1)「網羅性」の尺度
特定の文書あるいは文書集合中での語の出現頻度に基づく尺度．多く出現する語ほど特徴的であることの数量的な評価となっている．出現頻度を総のべ語数で正規化し，語の出現確率（の推定値）を尺度とする場合もある．

(2)「特定性」の尺度☆☆
情報量あるいはエントロピーに基づく尺度．ただし，情報量は確率の対数に -1 を乗じたもの，エントロピーは情報量の期待値である．語の出現の偏りに関する数量的な評価となっている．相互情報量に加えて信号雑音比や idf なども特定性の尺度として解釈できる．

(3)「識別性」の尺度
一般的に確率や情報量を用いて定義される尺度で，適合文書と不適合文書，あるいは異なるカテゴリに属する文書どうしの識別における語の有用性の数量的な評価となっている．確率のロジット変換（確率を p として $\log(p/(1-p))$）を識別関数として導出される確率型検索モデルの各種重み付け法，語の有無によるエントロピーの差分を計算する情報利得などに代表される．

(4)「$tf{\cdot}idf$」あるいは「代表性」の尺度
語頻度と対数文書頻度の逆数である idf を乗じた尺度．特定の文書において語が特徴的に多く出現することの数量的な評価になっていると考えられる．

☆ P 重み付け法の定義式と文献 13), 16) で特徴語選択に用いられている Odds Ratio の定義式は等しい．
☆☆「網羅性」および「特定性」という語は文献 24) による．

一般に，(1) の「網羅性」を特徴的な語の選択基準として用いると，多くの文書に共通に出現する高頻度語が過剰に重み付けされてしまうことがよく知られている．一方で，(2) の「特定性」を選択基準として用いると，低頻度語に対する重みが強くなりすぎてしまう．このように特徴語選択においては，網羅性と特定性のバランスをうまくとることが重要なポイントとなる．(3) の「識別性」および (4) の「$tf{\cdot}idf$」は両者とも，このような効果を意図したものであると解釈できる．ただし，両者は異なる観点に基づいており，たとえば，ある文献に限りまれにしか出現しない語の文献内での重み付けについて考えると，識別性の尺度ではこのような語には高い重みが与えられるが，$tf{\cdot}idf$ による重みは低いという違いがある．

さて，(3) の識別性の尺度は一般に確率や情報量を用いて定義されるため，その理論的な背景は明確である．しかし，このことは必ずしも実用上の性能を保証するものではない．たとえば，確率型検索モデルにおける語の重み付けの計算には，「検索質問に適合する文書中での語の出現確率」などの数値が必要であり，その推定が容易でないという実際上の問題点がある．また，テキスト分類のための特徴語の選択基準の比較において，情報利得は他の尺度より性能が劣るとの報告もある[13),16)]．

これに対して (4) の $tf{\cdot}idf$ は語の出現頻度および文書頻度からただちに計算可能で，上記のような確率推定の問題がない．また $tf{\cdot}idf$ は，現在の検索システムで広く用いられている指標であり，その有用性は経験的に実証されている．ただし，$tf{\cdot}idf$ の理論的な裏づけは必ずしも明確ではないというのが一般的な見解であり[22),24),25)]，このため $tf{\cdot}idf$ は，近年注目されているテキスト分類などにおける機械学習的なアプローチとは一線を画しているのが現状である．また，$tf{\cdot}idf$ には多数の経験的なバリエーションが存在することが，新しい領域に適用する際の問題点となるという指摘もなされている[26)]．

ここで，idf については理論的考証が古くから行われており，70 年代には 2 値独立（binary independence）確率型モデルによる説明がなされていた[8),27)]．また文献 28) では idf と信号雑音比を比較し，両者がともにシャノンのエントロピーを用いて（2 値独立型モデルを想定することなく）説明できることを示している．文献 2) では，idf と語の出現頻度の関係を分析し，語の idf の値が，ポアソン分布による語の生起過程のモデル化と実際の観察結果との間のずれの指標になっていることを示している．ただし，これらの理論

的あるいは実証的な考察は idf を対象にして行われており，idf に（単なる 2 値ではない）語頻度をかけあわせた尺度である $tf{\cdot}idf$ について特に言及するものではない．

近年，文献 26) では $tf{\cdot}idf$ の確率的な解釈を試みて，テキスト分類のための新しい尺度 $PrTFIDF$ を提案している．具体的には，情報検索の確率型モデルである RPI（Retrieval with Probability Indexing）[29] の考え方を適用して，分類対象となる文書が与えられた場合の各分類カテゴリの事後確率を $PrTFIDF$ として定義するもので，その理論的な裏づけは明確である．しかし $PrTFIDF$ と $tf{\cdot}idf$ は，数式の形のうえでは類似するものの，両者の意味的な関連は明らかではなく，これより $PrTFIDF$ は必ずしも $tf{\cdot}idf$ の有用性に対して直接の根拠を与えるものではない．

このような問題意識のもとに本論文では，情報量的な観点から $tf{\cdot}idf$ の理論的な解釈を試みる．ここで文献 28) より idf は一種の情報量と見なせることから，$tf{\cdot}idf$ は語頻度と情報量をかけあわせた尺度であるといえる．確率とも情報量ともエントロピー（すなわち確率と情報量の積和）とも異なるこのような量が，情報理論の分野において明示的に用いられることは少ない．しかし，情報検索の分野においては $tf{\cdot}idf$ の有用性が経験的に広く認められていることから，本論文において特徴度の尺度として改めて有用性を検討するものである．本論文中では上記の考えに基づき $tf{\cdot}idf$ の定義を拡張し，「語の出現確率」と「語の持つ情報量」の積を「特徴量」（feature quantity）として新たに定義したうえで，一般的な語の代表性尺度としての有効性を調べる[30),31)]．

以下，まず 2 章で情報量的な観点に基づく $tf{\cdot}idf$ の理論的な解釈を示し，その一般化である「特徴量」の数学的定義を述べる．次に，3 章で語の特徴量の計算式を示し，実際の文書集合を用いて計算を行った数値をふまえて考察を加える．また 4 章で，用語抽出タスクへの適用を通して異なる特徴量の定義を比較し，5 章で今後の課題について述べる．

2. 特徴量の数学的定義

2.1 情報量に関する基本的な定義式[32),33)]

$\mathcal{X}$，$\mathcal{Y}$ を，事象集合 X，Y をそれぞれ値とする確率変数とし，$\mathcal{X}$，$\mathcal{Y}$ の任意の事象 x_i（$\in X$），y_j（$\in Y$）に関する同時生起確率 $P(x_i, y_j)$ が与えられているものとする．$P(x_i, y_j)$ が与えられるとき，ただちに

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

かつ

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

であり，相互情報量の一般的な定義式により，x_i と y_j の間の自己相互情報量は次式となる．

$$\mathcal{M}(x_i, y_j) = \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \tag{3}$$

また $\mathcal{X}$ と $\mathcal{Y}$ の間の平均相互情報量（以下では，平均を省略して単に「相互情報量」と呼ぶ）は，自己エントロピー $\mathcal{H}$ を用いて次式で与えられる．

$$\begin{aligned}\mathcal{I}(\mathcal{X};\mathcal{Y}) &= \mathcal{H}(\mathcal{X}) - \mathcal{H}(\mathcal{X}|\mathcal{Y}) \\ &= \mathcal{H}(\mathcal{X}) + \mathcal{H}(\mathcal{Y}) - \mathcal{H}(\mathcal{X}\mathcal{Y}) \\ &= \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \ \log \frac{P(x_i, y_j)}{P(y_j)P(x_i)} \\ &= \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \mathcal{M}(x_i, y_j) \end{aligned} \tag{4}$$

ここで相互情報量の定義より，式 (4) は $\mathcal{X}$ と $\mathcal{Y}$ に対して対称的であり，$\mathcal{I}(\mathcal{X};\mathcal{Y}) = \mathcal{I}(\mathcal{Y};\mathcal{X})$ となる．

事象 x_i が観察されることによって $\mathcal{Y}$ に関して得られる情報量を，2 つの確率分布 $P(\mathcal{Y}|x_i)$ と $P(\mathcal{Y})$ の間のカルバックライプラー情報量（2 つの確率分布の違いを表す尺度）で表すことにすると，カルバックライプラー情報量の定義式により，

$$\mathcal{K}(P(\mathcal{Y}|x_i)||P(\mathcal{Y})) = \sum_{y_j \in Y} P(y_j|x_i) \log \frac{P(y_j|x_i)}{P(y_j)} \tag{5}$$

となる．同様に，事象 y_j が観察されることによって $\mathcal{X}$ に関して得られる情報量を $P(\mathcal{X}|y_j)$ と $P(\mathcal{X})$ の間のカルバックライプラー情報量で求めると，

$$\mathcal{K}(P(\mathcal{X}|y_j)||P(\mathcal{X})) = \sum_{x_i \in X} P(x_i|y_j) \log \frac{P(x_i|y_j)}{P(x_i)} \tag{6}$$

となる．式 (4)，(5)，(6)，および条件付き確率の一般則 $P(x_i, y_j) = P(y_j|x_i)P(x_i) = P(x_i|y_j)P(y_j)$ により，相互情報量とカルバックライプラー情報量の間には次式の関係が成立することが分かる．

$$\begin{aligned}\mathcal{I}(\mathcal{X};\mathcal{Y}) &= \sum_{x_i \in X} P(x_i) \ \mathcal{K}(P(\mathcal{Y}|x_i)||P(\mathcal{Y})) \\ &= \sum_{y_j \in Y} P(y_j) \ \mathcal{K}(P(\mathcal{X}|y_j)||P(\mathcal{X})) \end{aligned} \tag{7}$$

2.2 $tf \cdot idf$ の情報量的解釈

次に，相互情報量の定義である式 (4) を用いて，情報量的な観点に基づく $tf{\cdot}idf$ の解釈を示す．まず，各

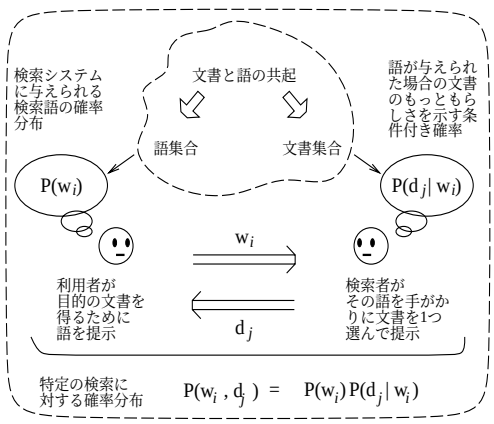


図 1 相互情報量の計算で想定している状況
Fig. 1 An illustrative situation assumed in the calculation of the expected mutual information.

文書は語の集合として与えられているものとし，文書集合を D，D に含まれるすべての語の集合を W とする．また，全文書数を N，語 w_i（$\in W$）を含む文書の数を N_i で表記する．D から 1 つ文書を選ぶという事象に対して定義される確率変数を $\mathcal{D}$，W から 1 つ語を選ぶという事象に対して定義される確率変数を $\mathcal{W}$ として，任意の語 w_i について，w_i を含む N_i 個の文書が既知である場合に，$\mathcal{D}$ と $\mathcal{W}$ の間の相互情報量の期待値を計算する（**図 1**）．

何も情報が与えられない状態において，すべての文書が同様に確からしいものとすると，D に含まれるすべての文書 d_j について $P(d_j) = 1/N$ であり，文書あたりの情報量は $-\log(1/N)$ となる．これより確率変数 $\mathcal{D}$ に関する情報量の期待値は次式で与えられる．

$$\begin{aligned}\mathcal{H}(\mathcal{D}) &= -\sum_{d_j \in D} P(d_j) \log P(d_j) \\ &= -N \cdot \frac{1}{N} \cdot \log \frac{1}{N} = -\log \frac{1}{N} \end{aligned} \tag{8}$$

さらに文書が語 w_i を含むという情報が与えられたものとする．w_i を含む N_i 個の文書はすべて同様に確からしいものとすると，文書あたりの情報量は $-\log(1/N_i)$ である．このとき，$\mathcal{D}$ に関する情報量の期待値は次式で与えられる．

$$\begin{aligned}\mathcal{H}(\mathcal{D}|w_i) &= -\sum_{d_j \in D} P(d_j|w_i) \log P(d_j|w_i) \\ &= -N_i \cdot \frac{1}{N_i} \cdot \log \frac{1}{N_i} = -\log \frac{1}{N_i} \end{aligned} \tag{9}$$

ここで，w_i を含まない文書に割り当てられる確率はゼロであることから，これら $(N - N_i)$ 個の文書に対応する項は上式の計算には現れていない．

次に，任意の語 w_i（$\in W$）が文書集合全体中での出現頻度に比例した確率で，ランダムに選ばれる場合について考える．全文書の総のべ語数を F，語 w_i の総出現頻度を f_{w_i} として，f_{w_i} が既知であるとき，w_i が選択される確率は f_{w_i}/F となる．このとき $\mathcal{D}$ と $\mathcal{W}$ の間の平均相互情報量は次式のように計算される．

$$\begin{aligned}\mathcal{I}(\mathcal{D};\mathcal{W}) &= H(\mathcal{D}) - H(\mathcal{D}|\mathcal{W}) \\ &= \sum_{w_i \in W} P(w_i) \left(H(\mathcal{D}) - H(\mathcal{D}|w_i)\right) \\ &= \sum_{w_i \in W} \frac{f_{w_i}}{F} \left(-\log \frac{1}{N} + \log \frac{1}{N_i}\right) \\ &= \sum_{w_i \in W} \frac{f_{w_i}}{F} \log \frac{N}{N_i} \\ &= \sum_{w_i \in W} \sum_{d_j \in D} \frac{f_{ij}}{F} \log \frac{N}{N_i} \end{aligned} \tag{10}$$

ただし f_{ij} は文書 d_j 中での語 w_i の出現頻度とする．

上式は，$\log N/N_i$ で定義される対数文書頻度（すなわち *idf*）と各語の出現頻度（すなわち tf）をかけあわせ，さらに定数項 $1/F$ を乗じて総和をとった値に等しい．これより tf·*idf* は，語と文書に関する相互情報量の計算に必要となる量で，特定の語と文書の共起による寄与分を表すものと解釈できる．

ただし上記の導出では整合性のため，文書 d_j に含まれる語集合を $W(d_j)$ として

$$P(d_j) = \sum_{W(d_j)} \frac{f_{w_i}}{F} \cdot \frac{1}{N_i} \approx \frac{1}{N} \tag{11}$$

の仮定が必要となる．すなわち，与えられた文書集合が比較的均一であることが暗黙の前提となっている．また逆に，このような仮定の適用自体が，tf·*idf* におけるヒューリスティックな戦略を表しているものと解釈できる．一方，近年の情報検索は従来の文献データベース検索から，Web 文書検索やテキスト分類など多様な範囲に拡大しており，このような状況では文書の均一性の仮定は必ずしも成立するものではない．そこで以下では，式 (10) を出発点として，tf·*idf* の定義をより一般的な定義に拡張することを試みる．

2.3 特徴量の一般的定義

図 1 の状況において，語 w_i と文書 d_j の同時生起確率 $P(w_i, d_j)$ が与えられているものとする．このとき，特定の語と文書の組合せが持つ「特徴」の量的表現を，式 (4) の相互情報量の計算における，指定された語と文書対の寄与分として，以下のように定義する．

$$\mathcal{F}(w_i, d_j) = P(w_i, d_j)\mathcal{M}(w_i, d_j) \tag{12}$$

$\mathcal{M}$ は式 (3) による自己相互情報量である．また，語 w_i の特徴量 $\mathcal{F}(w_i;\mathcal{D})$ を，式 (7) の相互情報量の計

算における，指定された語の寄与分として，次式で定義する．

$$\mathcal{F}(w_i;\mathcal{D}) = P(w_i)\ \mathcal{K}(P(\mathcal{D}|w_i)||P(\mathcal{D})) \quad (13)$$

同様に文書 d_j の特徴量 $\mathcal{F}(d_j;\mathcal{W})$ を，式 (7) の相互情報量の計算における，指定された文書の寄与分として，次式で定義する☆．

$$\mathcal{F}(d_j;\mathcal{W}) = P(d_j)\mathcal{K}(P(\mathcal{W}|d_j)||P(\mathcal{W})) \quad (14)$$

いずれの場合についても，特徴量は確率と情報量の積の形で表されており，情報量として式 (12) では自己相互情報量を，式 (13)，(14) ではカルバックライプラー情報量を用いている．さらに式 (13) は以下の形に書き改められる．

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

同様に式 (14) は以下の形に書き改められる．

$$\begin{aligned}\mathcal{F}(d_j;\mathcal{W}) &= \sum_{w_i\in W} P(d_j)P(w_i|d_j)\ \log\frac{P(w_i|d_j)}{P(w_i)} \\ &= \sum_{w_i\in W} P(w_i,d_j)\ \mathcal{M}(w_i,d_j) \\ &= \sum_{w_i\in W} \mathcal{F}(w_i,d_j) \qquad (16)\end{aligned}$$

最後に式 (4) の相互情報量は，式 (12)，(15)，(16) の特徴量を用いて以下のように計算される．

$$\begin{aligned}\mathcal{I}(\mathcal{D};\mathcal{W}) &= \sum_{d_j\in D}\sum_{w_i\in W} \mathcal{F}(w_i,d_j) \\ &= \sum_{w_i\in W} \mathcal{F}(w_i;\mathcal{D}) \\ &= \sum_{d_j\in D} \mathcal{F}(d_j;\mathcal{W}) \qquad (17)\end{aligned}$$

ここで，同時生起確率 $P(w_i,d_j)$ は，検索システムが適用する知識あるいは解釈を表しており，その値を決定する方法として，以下の 3 つが考えられる．

(1) $P(w_i)$ および $P(d_j|w_i)$ を推定する方法
図 1 において想定したように，検索システムに与えられる検索語の生起確率 $P(w_i)$ と検索語が与えられた場合の文書のもっともらしさ $P(d_j|w_i)$ が既知であるとする場合．

(2) $P(d_j)$ および $P(w_i|d_j)$ を推定する方法
テキスト分類における単純ベイズ法などで想定されるように，各文書のもっともらしさ $P(d_j)$ および各文書における語の生起確率 $P(w_i|d_j)$ が既知であるとする場合．

(3) $P(w_i,d_j)$ を直接推定する方法
確率的言語モデルの手法を適用して語–文書頻度行列から，$P(w_i,d_j)$ を直接推定する場合．単純に共起頻度に比例する確率を割り当てる方法，未知語の出現確率を差し引く方法，最大エントロピー法による推定などが考えられる[23),34)]．

ただし (1) では $P(d_j|w_i)P(w_i) = P(w_i,d_j)$，(2) では $P(w_i|d_j)P(d_j) = P(w_i,d_j)$ により $P(w_i,d_j)$ の値が決まることから，上記の 3 手法による推定は互いに書き換えが可能である．

なお文献 35) では，情報検索におけるベクトル空間モデルと確率分布モデルを双対的なものとしてとらえ（duality theory），前者が検索語を事象とする確率変数から文書の期待確率を計算する手法であるのに対して，後者は文書を事象とする確率変数から語の重みを計算する手法であると位置づけている．この定式化は上記における確率推定法の (1) および (2) によく対応している．

3. 語の特徴量に関する計算式

3.1 $tf\cdot idf$ と $tf\cdot kli$

前章と同様に，文書 d_j における語 w_i の出現頻度を f_{ij}，文書集合全体での w_i の出現頻度を f_{w_i}，全文書中でのすべての語の出現頻度の総和を F で表記する．また，文書 d_j に含まれるすべての語の出現頻度の総和を f_{d_j} とする．すなわち，$F = \sum_i\sum_j f_{ij} = \sum_j f_{d_j} = \sum_i f_{w_i}$ である．いま確率 $P(w_i,d_j)$ が語の出現頻度に比例して以下で与えられるものとする．

$$P(w_i,d_j) = \frac{f_{ij}}{F} \quad (18)$$

このとき式 (13) により定義される語の特徴量の計算式は次式のようになり，

$$\begin{aligned}\mathcal{F}(w_i;\mathcal{D}) &= P(w_i)\ \mathcal{K}(P(\mathcal{D}|w_i)||P(\mathcal{D})) \\ &= \frac{f_{w_i}}{F}\sum_{d_j\in D}\frac{f_{ij}}{f_{w_i}}\log\frac{\frac{f_{ij}}{f_{w_i}}}{\frac{f_{d_j}}{F}} \qquad (19)\end{aligned}$$

$\mathcal{F}(w_i;\mathcal{D})$ の値が大きいほど語は文書を特徴づける手がかりとして有用であると見なされることになる．語の選択基準が，語頻度とカルバックライプラー情報量

☆ 本論文では特に文書の特徴量については扱っていないが，ここでは特徴量の語と文書に関する対称性を示すため，双方についての定義式を並べる．

（Kullback-Leibler information, *kli*）の積で表されていることから，以下では式 (19) のような尺度を *tf·kli* で参照する．

さて，*tf·idf* を語の特徴量の尺度として用いる場合に，整合性のため定数 $1/F$ をかけあわせると，計算式は次式で与えられる．

$$tfidf(w_i;\mathcal{D}) = \frac{f_{w_i}}{F}\log\frac{N}{N_i} \tag{20}$$

式 (20) と式 (19) を比較すると，*idf* とカルバックライプラー情報量が類似の役割を果たしており，特に次の 2 つの条件が成立するとき，両者はよく一致することが分かる．

$$\text{(条件 1)} \qquad \frac{f_{d_j}}{F} \approx \frac{1}{N} \tag{21}$$

$$\text{(条件 2)} \qquad \frac{f_{ij}}{f_{w_i}} \approx \frac{1}{N_i} \qquad (f_{ij} > 0) \tag{22}$$

逆に (条件 1) と (条件 2) が成立すれば，式 (11) の条件 $P(d_j) = \sum_{W(d_j)} f_{w_i}/F \cdot 1/N_i = 1/N$ が成立し，各語の *tf·idf* 値の総和を相互情報量として矛盾なく解釈できる．ここで (条件 1) は，各文書ののべ語数がほぼ等しいことを意味し，(条件 2) は，ある語が共通に出現した文書の間で，その出現頻度に大きなばらつきがないことを意味している．具体的にこれらの条件が成立する例として，たとえば抄録など比較的短い文書の集合などがあげられる．また，文書に語が含まれるか否かにより，頻度を 1（含まれる）または 0（含まれない）に設定する場合には，(条件 2) は自動的に成立する．

また，上記では式 (13) の $\mathcal{F}(w_i;\mathcal{D})$ に注目して比較を行ったが，(条件 1)，(条件 2) が成立する場合には，*tf·idf* による文書 d_j 中の語 w_i の重みを $tfidf(w_i, d_j)$ で表記して，

$$\begin{aligned}\mathcal{F}(w_i, d_j) &= \frac{f_{ij}}{F}\log\frac{\dfrac{f_{ij}}{f_{w_i}}}{\dfrac{f_{d_j}}{F}} \\ &\approx \frac{f_{ij}}{F}\log\frac{N}{N_i} \\ &= tfidf(w_i, d_j)\end{aligned} \tag{23}$$

となり，特定の文書中での語の特徴量についても同様の結果が期待できる．

3.2 数値計算の例

tf·idf の情報量的な解釈の妥当性を確認するため，実際にデータベースなどから抽出した文書集合を用いて *tf·idf* と *tf·kli* の値を比較した結果を示す．実験では，学術情報センター学会発表データベースから，24 学会で発表された文献の抄録 327,904 件を抽出して，形態素解析により語の切り出しを行った．そして，(1) 各抄録を 1 つの文書に対応させる場合（GAKKAI-doc），(2) 同じ学会で発表された文献すべてを単一の「文書」であると見なす場合（GAKKAI-cls）の 2 通りについて，文書あたりの語の出現頻度を調べ，*idf* と *kli*，*tf·idf* と *tf·kli* の数値の相関を求めた．また，(3) Web 上からランダムにサンプリングした 72,386 個の HTML 文書に対しても同様の処理を適用し（WEB），*idf* と *kli*，*tf·idf* と *tf·kli* の数値の相関を，相関係数を用いて評価した．

これら 3 種類の文書集合のうち，GAKKAI-doc は，文書あたりのべ語数（f_{d_j}）が少なく，その標準偏差を平均値で正規化した値を見ると，f_{d_j} の文書による偏りも小さい．また，文書内での語の出現頻度（$f_{ij} > 0$）に注目して各語ごとの文書内出現頻度の偏差をすべての語について平均した値を見ると，f_{ij} の文書による偏りも小さいことが分かる．すなわち (条件 1) および (条件 2) が成立すると考えられる．一方 GAKKAI-cls では，文書あたりのべ語数が多く，f_{d_j}，f_{ij} ともに偏差が大きい．この場合には (条件 1) および (条件 2) は満足されないと考えられる．WEB に関しては，文書あたりののべ語数は比較的少いが，その文書間での格差は大きく，一方 f_{ij} の偏差は少ないことから，(条件 2) だけが成立していると考えられる．

各文書集合に対する相関係数の値を**表 2** にまとめる．*idf* と *kli* の相関係数の値を比較すると，GAKKAI-doc，WEB，GAKKAI-cls の順に相関が低くなっている．特に GAKKAI-doc の相関は 0.95 と高く，実際の文書集合においても，文書が比較的均一で (条件 1) および (条件 2) が満たされる場合には，*idf* がカルバックライプラー情報量の単純で頑強な推定値となっていることが確認できる．一方，*tf·idf* と *tf·kli* の相関係数の値に注目すると，GAKKAI-doc，WEB の両者において高い値を示しており，一般に (条件 2) が成立する場合，すなわち文書長が短く文書内頻度の偏りが少ない場合には，*tf·idf* と *tf·kli* はほぼ同じ値を与えることが推察される．

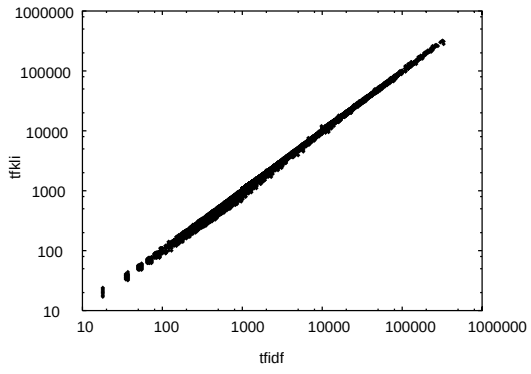
また，**図 2** は GAKKAI-doc と GAKKAI-cls について，*tf·idf* と *tf·kli* の値の相関を図示した結果である．各語について *tf·idf* と *tf·kli* の値を，それぞれ横軸と縦軸に示してある．GAKKAI-doc では，両者の値はほぼ一致しているのに対して，GAKKAI-cls ではばらつきが大きく，*tf·kli* の値が有意に高くなっていることが確認できる．

3.3 確率分布モデルに関する考察

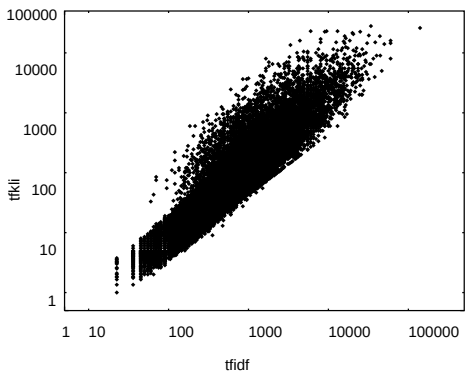
上記の実験では，式 (18) に基づく $P(w_i, d_j)$ の推

表 2 異なる文書集合における idf と kli, $tf{\cdot}idf$ と $tf{\cdot}kli$ の相関
Table 2 Correlation between idf and kli, and also $tf{\cdot}idf$and $tf{\cdot}kli$, for different document sets.

文書集合	文書数	文書あたり平均のべ語数 f_{d_j}	f_{d_j} の標準偏差（平均値で正規化）	f_{ij} 偏差の全語平均	文書の特徴	idf-kli 相関係数	$tf{\cdot}idf$-$tf{\cdot}kli$ 相関係数
GAKKAI-doc	327,904	8.39×10^1	0.38	0.16	語数少・均一	0.95	1.00
GAKKAI-cls	24	1.14×10^6	1.36	6.27	語数多・非均一	0.40	0.53
WEB	72,386	2.76×10^2	1.81	0.30	語数少・非均一	0.76	0.98



(a) GAKKAI-d における $tf{\cdot}idf$ 値と $tf{\cdot}kli$ 値の相関



(b) GAKKAI-c における $tf{\cdot}idf$ 値と $tf{\cdot}kli$ 値の相関

図 2 $tf{\cdot}idf$ と $tf{\cdot}kli$ の数値比較
Fig. 2 Numerical comparison of the $tf{\cdot}idf$ and $tf{\cdot}kli$ values.

定値を用いて比較を行ったが，検索タスクのモデルとして図 1 を想定する場合には，語の生起確率 $P(w_i)$ と条件付き確率 $P(d_j|w_i)$ を，それぞれ独立に定めることが可能である．この場合に，$P(w_i)$ は語 w_i が検索語としてシステムに与えられる確率，$P(d_j|w_i)$ は語 w_i が手がかりとして与えられた場合の文書 d_j のもっともらしさを表すことから，前者を利用者のモデル，後者を検索システムのモデルに対応づけて考えることができる．

これまでに考案されてきた $tf{\cdot}idf$ の数多くのバリエーションは，(1) tf に対する非線形重み付け（たとえば $\sqrt{tf}$ や $\log(tf)$ など），(2) idf の算出法の変形（たとえば $idf = \log(N/N_i) + 1$ など），および，(3) 両者の組合せ，のいずれかである．ここで，本論文における定式化に従って，(1) を $P(w_i)$ 推定の問題，(2) を $P(d_j|w_i)$ 推定の問題であるととらえると，$tf{\cdot}idf$ のバリエーションとは，検索タスクに対して想定する確率的なモデルの違いであると解釈できる．

さらに，$tf{\cdot}idf$ と $tf{\cdot}kli$ の違いもまた，確率モデル選択の問題としてとらえることが可能である．すなわち，idf は「語 w_i が出現した N_i 個の文書だけに非ゼロの確率を配分する」という制約条件のもとで，エントロピーを最大にする $P(d_j|w_i)$ の確率配分に基づく計算法，一方 kli は標本分布をそのまま真の確率の推定値として用いる計算法であり，ともに特徴量の定義式 (12)，(13) に従う尺度であると見なせる．したがって，いずれが優れているかという問題は，確率的な尺度の定義の問題ではなく，標本が与えられた場合の確率モデルの選択問題として定式化できる．

$P(w_i)$ や $P(d_j|w_i)$ の推定にあたっては，そのほか，ディスカウンティングや最大エントロピー法を含む確率的な言語モデル[23),34)]，テキスト分類における Laplace 推定[36)]などの適用が可能である．拡張性の観点からは，これらの確率モデルの妥当性を検証する方が，ヒューリスティックに考案された $tf{\cdot}idf$ の多様な計算法を取捨選択するよりも合理的であるといえる．

4. 用語抽出タスクへの適用例

4.1 実験の概要

2.3 節では語の特徴度を示す尺度として，文書 d_j における語の特徴度 $\mathcal{F}(w_i, d_j)$，および，文書集合 $\mathcal{D}$ における語の特徴度 $\mathcal{F}(w_i; \mathcal{D})$ の 2 種類の尺度を定義した．前者は，文書集合中の特定の文書に注目して，その文書に特徴的に多く出現する語を選別するための尺度であり，後者は，文書集合中の任意の文書どうしを互いに区別するうえで手がかりとなる語を選別するための尺度である．さらに 3.3 節の議論により，特徴量の計算値は，想定する確率モデルにも依存することが

表 3 特徴語抽出タスクにおいて設定した条件
Table 3 Conditions used in our term extraction experiments.

特徴量の計算で想定した条件	確率の推定に用いた文献集合	「文書」の単位	特徴量の計算に用いた文献集合	特徴量の計算式 ($P(d_j\|w_i) = f_{ij}/f_{w_i}$ は共通)
(A) AI 文書-tf	人工知能学会の文献	抄録	人工知能学会の 2,170 文献	$\mathcal{F}(w_i;\mathcal{D})$, $P(w_i) = f_{w_i}/F$
(B) AI 文書-log(tf)	人工知能学会の文献	抄録	人工知能学会の 2,170 文献	$\mathcal{F}(w_i;\mathcal{D})$, $P(w_i) \propto \log(f_{w_i}/F)$
(C) AI クラス	24 学会の文献	学会	人工知能学会	$\mathcal{F}(w_i,d_j)$, $P(w_i) = f_{w_i}/F$
(D) 非 AI クラス	24 学会の文献	学会	人工知能学会以外の 23 学会	$\mathcal{F}(w_i,d_j)$, $P(w_i) = f_{w_i}/F$
(E) AI クラス +	17 学会の文献	学会	人工知能学会	$\mathcal{F}(w_i,d_j)$, $P(w_i) = f_{w_i}/F$
(F) 非 AI クラス +	17 学会の文献	学会	人工知能学会以外の 16 学会	$\mathcal{F}(w_i,d_j)$, $P(w_i) = f_{w_i}/F$
(G) AI クラス +IG	17 学会の文献	学会	人工知能学会とそれ以外の 16 学会	$IG(w_i;\mathcal{D})$, $P(w_i) = f_{w_i}/F$

分かる．

以下では，特徴量の値が，このように想定する状況に応じて変わるものであることを例示するため，前章と同様に学会発表データベースに登録された文献を対象として，実際に特徴語のランキングを行った結果を示す．具体的には，学会発表データベースの登録文献から人工知能分野の特徴語を自動抽出して，TMREC[37] で人手により作成された人工知能分野の用語抽出タスク正解集合を用いて，上位にランキングされた語と正解集合の一致度を調べた．

実験ではまず，後述する条件に従って選んだ文献（タグなしテキスト）に形態素解析ツール「茶筅」バージョン 2.02[38]を適用して，品詞情報を手がかりに複合語の最短および最長単位を抽出した．ただし，本実験は複合語単位抽出における自然言語処理の評価を目的とするものではないことから，ここで用いたルールは多分に経験的で単純なものである．最短単位はたとえば「帰納」,「論理」,「プログラミング」，最長単位はたとえば「帰納論理プログラミング」などであり，ともに用語として抽出した．次に，著者キーワードとしてデータベースに登録された約 40 万語を用語候補辞書として，専門用語として意味をなさない不要語を機械的に除去した．最後に，得られたすべての語を対象として，特徴量の計算値に基づくランキングを行った．

計算にあたり想定した条件は，**表 3** に示す 7 通りである．表 3 の (A)「AI 文書-tf」および (B)「AI 文書-log(tf)」は人工知能学会で発表された 2,170 件の文献を対象とするもので，各文献を 1 つの文書に対応させ，文書どうしを区別するうえで有用な語を抽出した．$P(d_j|w_i)$ は頻度情報に基づき $P(d_j|w_i) = f_{ij}/f_{w_i}$ とした．一方，$P(w_i)$ について,「AI 文書-tf」では $P(w_i) = f_{w_i}/F$ による線形重み付けを,「AI 文書-log(tf)」では $P(w_i) \propto \log(f_{w_i}/F)$ による非線形重み付けを適用した．計算では論文中の定義式 (13) による語の特徴量 $\mathcal{F}(w_i;D)$ を用いたが，この場合に各文書は比較的均一であり，3.2 節の結果から，得られる特徴語のランキングは従来の $tf{\cdot}idf$ による結果とほぼ等価である．

表 3 の (C)「AI クラス」および (D)「非 AI クラス」では，全文献を「人工知能学会における発表文献」2,170 件と「人工知能学会以外の 23 学会における発表文献」335,734 件の 2 つのクラスに分割して，各クラスを 1 つの文書に対応させたうえで，それぞれ特徴的な語を抽出した．$P(w_i)$，$P(d_j|w_i)$ の値は「AI 文書-tf」の場合と同様に，$P(w_i) = f_{w_i}/F$，$P(d_j|w_i) = f_{ij}/f_{w_i}$ により定めた．特徴量の計算では論文中の定義式 (12) による語と文書の特徴量 $\mathcal{F}(w_i,d_j)$ を用いた．この場合の文書数は 2 であり，文書長の不均衡も著しいことから，従来の $tf{\cdot}idf$ の定義は直接適用できない．

さらに表 3 の (E)「AI クラス +」および (F)「非 AI クラス +」では,「人工知能学会における発表文献」2,170 件と「人工知能学会以外の 16 学会における発表文献」172,755 件の 2 つのクラスを設定して，上記と同様の処理を行った．ここで，(D) の「非 AI クラス」で対象とした文献の中には,「情報処理学会」など人工知能分野と関連が深い学会も含まれるが，(F) の「非 AI クラス +」では，これら人工知能学会以外の計算機・情報処理関連学会を対象外として，残りの 17 学会について計算を行った．

最後に表 3 の (G)「AI クラス +IG」では，識別性の尺度である情報利得（information gain）を用いて，(E) の「AI クラス +」および (F) の「非 AI クラス +」で設定した 2 つの文書クラスを互いに識別するために有用な語のランキングを行った．$P(w_i)$，$P(d_j|w_i)$ の推定値は「AI クラス +」の場合と同様に $P(w_i) = f_{w_i}/F$，$P(d_j|w_i) = f_{ij}/f_{w_i}$ により定めた．情報利得は機械学習の分野で多く用いられている尺度で，語 w_i が文書に「含まれる」/「含まれない」が既知となった場合のエントロピーの差分として定義される．すなわち，$\overline{w_i}$ を w_i 以外の語すべてに対応する事象として，情報利得 $IG(w_i;\mathcal{D})$ は以下で与えられる[9]．

表 4 異なる特徴量の定義によるランキング結果（上位 20 位）
Table 4 Top 20 Ranking of terms extracted using varied definitions of the feature quantity.

(A) AI 文書-tf	(B) AI 文書-log(tf)	(C) AI クラス	(D) 非 AI クラス	(E) AI クラス +	(F) 非 AI クラス +	(G) 非 AI クラス +IG
1618 提案	28 補間	1463 知識*	252631 結果	1463 知識*	153091 結果	2635 知識*
1245 モデル*	57 欠落	903 推論*	114875 特性	903 推論*	99097 検討	2412 学習*
1463 知識*	40 視野*	89 知的教育システム*	177192 検討	1166 学習*	62519 影響	18627 システム*
1351 問題*	39 ナビゲーション*	206 帰納*	70304 測定	437 エージェント*	48319 特性	1511 推論*
1166 学習*	62 振舞い*	122 オントロジー*	90398 影響	712 言語*	56615 量	1254 言語*
1325 研究	66 写像*	93 アブダクション*	82228 量	567 論理*	41306 測定	3316 支援
2185 システム*	49 意味論*	1166 学習*	46948 光	430 ユーザ*	35597 試験*	905 論理*
988 手法	65 対象モデル*	160 機械学習*	103618 報告	422 対話*	48613 変化	18790 提案
898 論文	53 対話システム*	61 帰納論理プログラ$*	45683 試験*	286 学習者*	29220 反応	467 エージェント*
1037 情報*	39 ASK	437 エージェント*	77451 変化	447 知的	27895 分子	6786 表現
993 表現	39 メイル	447 知的	34525 強度	315 プログラミング*	27890 強度	567 ユーザ*
931 方法	101 類推*	148 知的 CAI*	34168 温度	228 インタフェース*	29445 活性	597 対話*
986 利用	66 演繹*	286 学習者*	33094 波	206 帰納*	31891 調査	768 知的
903 推論*	73 インターネット*	32 アブダクティブ論$*	34027 体	224 知識獲得*	25554 酸	2004 ベース*
936 支援	53 故障診断*	47 行き詰まり	32482 反応	936 支援	24062 温度	9921 情報*
914 結果	73 失敗	125 ITS*	99037 解析*	240 CAI*	25474 体	2744 記述*
837 必要	93 アブダクション*	92 論理プログラム*	30408 分子	224 発話*	23625 水	3901 実現
712 言語*	53 談話*	97 帰納学習*	40269 分布	288 知能*	31506 合成	1272 獲得
749 処理	48 デザイン*	224 知識獲得*	29369 周波数	299 音声*	24033 質	20368 問題*
741 構築	35 メニュー*	40 AI システム*	35698 回路	205 知識表現*	21254 低下	8047 論文

$$IG(w_i;\mathcal{D}) = \sum_{d_j \in D} P(w_i, d_j)\mathcal{M}(w_i, d_j) + \sum_{d_j \in D} P(\overline{w_i}, d_j)\mathcal{M}(\overline{w_i}, d_j) \quad (24)$$

ここで，人工知能学会の文献集合から構成される文書クラスを d_1，人工知能学会以外の文献集合から構成される文書クラスを$\overline{d_1}$，$D = \{d_1, \overline{d_1}\}$ として式 (24) を書き改めると

$$\begin{aligned} IG(w_i;\mathcal{D}) &= P(w_i, d_1)\mathcal{M}(w_i, d_1) \\ &\quad + P(w_i, \overline{d_1})\mathcal{M}(w_i, \overline{d_1}) \\ &\quad + P(\overline{w_i}, d_1)\mathcal{M}(\overline{w_i}, d_1) \\ &\quad + P(\overline{w_i}, \overline{d_1})\mathcal{M}(\overline{w_i}, \overline{d_1}) \\ &= \mathcal{F}(w_i, d_1) + \mathcal{F}(w_i, \overline{d_1}) \\ &\quad + \mathcal{F}(\overline{w_i}, d_1) + \mathcal{F}(\overline{w_i}, \overline{d_1}) \quad (25) \end{aligned}$$

となり，第 1 項は前出の「AI クラス +」における特徴量の定義に等しい．また第 3 項および第 4 項は,「語 w_i が含まれない」という事象によるもので，計算結果への影響は無視できる程度と考えられる．これより，「AI クラス +」と「AI クラス +IG」の違いは主に，式 (25) の第 2 項，すなわち人工知能学会以外の学会で特に多く出現する語にかかわるものであることが予想される．

4.2 用語抽出タスクによる比較結果

表 4 に，条件 (A)〜(G) による特徴語のランキング上位 20 位を示す．表の左端の数値は対象とする文書集合内での語の出現頻度，右端の $*$ は，その語が人工知能分野の用語抽出タスクの正解集合[37]に含まれることを表している．ここで，人工知能学会の文献集合から得られる用語候補の数は 10,220 語，その中に含まれる正解語数は 3,963 語で条件 (A)，(B)，(C)，(E)，(G) に共通である．これより正解語数による評価を行うものとし，**図 3** に，ランキング上位 N 語（$N \leq 2{,}000$）について，上記の正解集合による正解語数を比較した結果を示す．

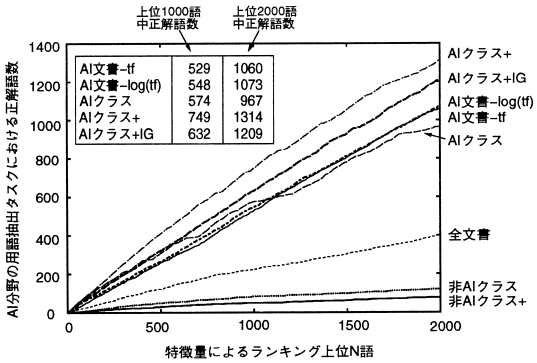


図 3 異なる特徴量の定義による用語抽出結果の比較
Fig. 3 Comparison of term extraction results using varied definitions of the feature quantity.

（1） 頻度の非線形重み付けによる影響

「AI 文書-tf」と「AI 文書-log(tf)」を比較することにより，tf に対する非線型重み付けの効果が確認できる．表 4 において，単純な線形重み付けでは，頻度を用いた場合と近い順位づけが行われており,「提案」や「研究」などの一般的な語が上位にあがっている．一方，対数による非線形重みを用いた場合では，低頻度語の重みが強まり,「類推」や「アブダクション」など人工知能分野に特徴的な語の順位が高くなっている．しかし，このように上位にランキングされる語はかなり異なるものであるにもかかわらず，図 3 の正解率

で比較すると，両者に大きな違いは見られない．非線形重み付けの方法をさらに工夫することによって，頻度の影響を調整することは容易であるが，多数のバリエーションのうちでいずれを選択するべきかという問題は多分に経験的なもので本論文の範囲を越えることから，ここでは比較の対象に含めなかった．

（2） 特徴量の定義式の違いによる影響

「AI 文書」と「AI クラス」を比較することにより，特徴量の定義式の違いによる影響を確認することができる．両者は同一の文献集合に対して特徴語を求めたものであるが，「AI 文書」では，人工知能学会の文献だけに注目して文献どうしを区別するうえで有用な語を抽出しているのに対して，「AI クラス」では，他学会と比較して人工知能学会に特徴的に多く出現する語を抽出している．結果として前者では，比較的高頻度で一般的な語が上位にランキングされるが，後者では，「帰納論理プログラミング」「アダプティブ論理プログラミング」（ただし表中では末尾を '$' により省略）など，高度に専門的な用語が上位にランキングされる．このような違いを図 3 の正解率で比較すると，上位約 1,000 語までは「AI クラス」の方が，それ以降は「AI 文書」の方が正解率が高くなっており，結論として両者のいずれかが優れているかの判定は困難である．

（3） 比較対象とする文書集合の違いによる影響

「AI クラス」と「AI クラス +」を比較することにより，参照する文書集合による違いを見ることができる．両者では，特徴量の計算に用いた文書集合は同じであるが，比較の対象として確率推定に用いた文書集合が異なる．「AI クラス」では，人工知能学会と対象分野が重複する学会を比較対象に含めているため，上位にランキングされる語は専門的になる傾向が見られるが，「AI クラス +」では，人工知能学会と重複が少ない学会だけを比較対象に設定していることから，「AI クラス」と比較すると一般的な語が得られている．図 3 の比較では，「AI クラス +」は「AI クラス」を含め，設定した 7 条件の中で一番正解率が高くなっている．

（4） 特徴量と情報利得の比較

「AI クラス +」と「AI クラス +IG」を比較することにより，代表性の尺度である特徴量と識別性の尺度である情報利得の違いを見ることができる．情報利得では，人工知能学会と他 16 学会を互いに識別するうえで有用な語を選択することから，「システム」や「提案」など，文書集合全体での頻出語が上位にランキングされている．図 3 の比較では，「AI クラス +」の方が「AI クラス +IG」よりも正解率が高い．ただし，かなり一般的と思われる高頻度語が上位に位置づけられているにもかかわらず，「AI クラス +IG」について，他の「AI 文書」や「AI クラス」よりも高い正解率が得られたことは，人工知能学会との対比のために設定した「非 AI クラス +」の妥当性を裏づけるものとして注目に価する．

以上の結果から，語の特徴度が想定する状況に応じて変化するものであること，および，実験で確認した範囲の中では，十分広範囲でかつ分野的な重複が少ない文献集合と対比することにより，比較的人間の直感と適合した特徴語のランキングが行えることが分かる．ここで TMREC の正解集合は，文献のキーワードではなく分野全体の専門用語の抽出を目的として作成されたものであることから，文献を単位とするランキングと比較して，学会全体を単位とするランキングの方が高い正解率が得られることは自然である．また TMREC では，用語抽出の分野からのアプローチとして，複合語の構成要素としての有用性に注目する C-value[39] や *Pre & Post* 重要度[40] などの手法の有効性が報告されている．これらの手法は，「AI 文書」の場合と同様に関連する分野の文献集合だけに注目するものであるが，語を文書ではなく，その語と前後して共起する他の語によって特徴づける点が異なる．本論文において優れた性能を観察した「AI クラス +」との比較や併用の効果については，今後の検討課題となっている．

5. むすび

本論文では，語の出現確率と情報量の積を「特徴量」として新たに定義することにより，$tf{\cdot}idf$ の考え方を一般化し，用語抽出タスクの例題を通して，その妥当性を検証した．本論文における検討は，主に $tf{\cdot}idf$ をはじめとする従来尺度との整合性を重視しており，現在のところ，個別の適用分野における性能改善に直接結び付くものではない．しかしながらこのような考察によって，$tf{\cdot}idf$ を情報量的な裏づけを持つ尺度として無理なく解釈するための条件が明らかになり，他の尺度との比較においても同一の確率モデルを用いることができると考えられる．

本論文で用いた特徴量の定義は，語に限らず，文書や，共起関係にある他の要素どうし，たとえば語と語，語とカテゴリ，文書と中間概念表現などの組合せにも適用可能である．これに基づき現在，テキスト分類，複合語や対訳抽出といった異なる領域について，特徴量の適用や確率モデルに関する検討を進めている．

謝辞 本研究は学術振興会の未来開拓学術研究推進事業による「高度分散情報資源活用のためのユービキ

タス情報システムに関する研究」のもとで行われた．本研究を行うにあたり，活発な議論やデータの提供をいただいた国立情報学研究所の影浦峡氏，高須淳宏氏，相原健朗氏に感謝の意を表する．

参 考 文 献

1) Caraballo, S.A. and Charniak, E.: Determining the Specificity of Nouns from Text, *Proc. 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (*EMNLP'99*), pp.63–70 (1999).
2) Church, K. and Gale, W.: *Inverse Document Frequency (IDF): A Measure of Deviations from Poisson*, pp. 283–295, Kluwer Academic Pub. (1999). (in "Natural Language Processing Using Very Large Corpora").
3) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information and Lexicography, *Proc. 27th Annual Meeting of the Association for Computational Linguistics* (*ACL'98*), pp.76–83 (1989).
4) Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).
5) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
6) Hisamitsu, T., Niwa, Y. and Tusjii, J.I.: A Method of Measuring Term Representativeness – Baseline Methods Using Co-occurrence Distribution, *Proc. 18th International Conference on Computational Linguistics* (*COLING2000*) (2000). (to appear)
7) Kageura, K. and Umino, B.: Methods of Automatic Term Recognition: A Review, *Terminology*, Vol.3, No.2, pp.259–289 (1998).
8) Manning, C.D. and Schtüze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press (1999).
9) Yang, Y. and Pedersen, O.: A Comparative Study on Feature Selection in Text Categorization, *Proc. 14th International Conference on Machine Learning* (*ICML'97*), pp.412–420 (1997).
10) Lewis, D.D. and Ringuette, M.: Comparison of Two Learning Algorithms for Text Categorization, *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval* (*SDAIR'94*), pp.81–93 (1994).
11) Yang, Y. and Liu, X.: A Re-examination of Text Categorization Methods, *Proc. 22nd International Conference on Research and Development in Information Retrieval* (*SIGIR'99*), pp.42–49 (1999).
12) Wiener, E., Pedersen, J.O. and Weighend, A.S.: A Neural Network Approach to Topic Spotting, *Proc. DAIR'95*, pp.317–332 (1995).
13) Mladenić, D.: Feature Subset Selection in Text-Learning, *Proc. 10th European Conference on Machine Learning* (*ECML'98*), pp.95–100 (1998).
14) Koller, D. and Sahami, M.: Toward Optimal Feature Selection, *ICML'96*, pp.284–292 (1996).
15) Koller, D. and Sahami, M.: Hierarchically Classifying Documents using Very Few Words, *ICML'97*, pp.170–178 (1997).
16) Mladenić, D. and Grobelnik, M.: Feature Selection for Classification based on Text Hierarchy, *Working notes of Learning from Text and the Web, CONALD'98* (1998).
17) Robertson, S.E.: Documentation Note on Term Selection for Query Expansion, *Journal of Documentation*, Vol.46, No.4, pp.359–364 (1990).
18) Luhn, H.P.: A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, Vol.1, No.4, pp.309–317 (1957).
19) Spark-Jones, K.: A Statistical Interpretation of Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–21 (1972).
20) Salton, G. and Buckley, C.: Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol.24, No.5, pp.513–523 (1988).
21) Robertson, S.E. and Spark-Jones, K.: Relevance Weighting of Search Terms, *Journal of the American Society of Information Science*, Vol.27, pp.129–146 (1976).
22) 岸田和明：情報検索の理論と技術，勁草書房 (1998).
23) 北　研二：確率的言語モデル，東京大学出版会 (1999).
24) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
25) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, ACM press and Addison Wesley (1999).
26) Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *ICML'97*, pp.143–151 (1997).
27) Croft, W. and Harper, D.J.: Using Probabilistic Models of Document Retrieval without

Relevance Information, *Journal of Documentation*, Vol.35, pp.285–279 (1979).
28) Wong, S. and Yao, Y.: An Information Theoretic Measure of Term Specificity, *Journal of the American Society for Information Science*, Vol.43, No.1, pp.54–61 (1992).
29) Fuhr, N.: Models for Retrieval with Probabilistic Indexing, *Information Processing and Management*, Vol.25, No.1, pp.55–72 (1989).
30) 相澤彰子：語と文書の共起に基づく「特徴量」の定義と適用，情報処理学会自然言語処理研究会，NL 136-4, pp.25–32 (2000).
31) Aizawa, A.: The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, *Proc. ACM SIGIR2000*, pp.104–111 (2000).
32) 宮川 洋：情報理論，コロナ社 (1954).
33) Cover, T.M. and Thomas, J.A.: *Elements of Information Theory*, John Wiley and Sons, Inc. (1991).
34) 山本幹雄：統計的言語モデル—理論と実験，第5回言語処理学会チュートリアル資料，pp.9–24 (1999).
35) Amati, G. and van Rijsbergen, K.: *Semantic Information Retrieval*, Kluwer Academic Pub., pp.189–219 (1998). (in "Information Retrieval: Uncertainty and Logics").
36) McCallum, A. and Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification, *AAAI-98 Workshop on learning for text categorization*, pp.42–49 (1998).
37) Kageura, K., Yoshioka, M., Tsujii, K., Yoshikane, F., Takeuchi, K. and Koyama, T.: Evaluation of the Term Recognition Task, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition* (*NTCIR Workshop 1*), pp.417–434 (1999).
38) 松本裕治，北内 啓，山下達雄，平野善隆，松田寛，浅原正幸：日本語形態素解析システム「茶筌」Version 2.0 使用説明書第2版，NAIST Technical Report NAIST-IS-TR99012，奈良先端科学技術大学院大学 (1999).
39) Frantzi, K.T. and Ananiadou, S.: Extracting Nested Collocations, *Proc. COLING'96*, pp.41–46 (1996).
40) Nakagawa, H. and Mori, T.: Nested Collocation and Compound Noun for Term Extraction, *Proc. 1st Workshop on Computational Terminology* (*COMPTERM'98*), pp.64–70 (1998).

(平成12年 7 月21日受付)
(平成12年10月 6 日採録)

相澤 彰子（正会員）

1985年東京大学工学部電子工学科卒業．1990年同大学大学院電気工学専攻博士課程修了．工学博士．1990〜1992年，イリノイ大学アーバナ・シャンペイン校客員研究員．現在，国立情報学研究所助教授．遺伝的アルゴリズム，統計的情報処理，自動用語抽出等の研究に従事．