

# 書誌情報における著者名の曖昧性 解消のためのクラスタリング

## Clustering for Name Disambiguation in Author Citations

正田 備也<sup>\*</sup> 高須 淳宏<sup>\*</sup> 安達 淳<sup>\*</sup>

Tomonari MASADA Atsuhiko TAKASU  
Jun ADACHI

本論文では、書誌情報に現れる省略著者名を、フルネームに正しく対応付けるためのクラスタリング手法を提案する。クラスタリングには、ナイーブ・ベイズ混合モデルと、新たに提案する2変数混合モデルを用いた。実験ではDBLPデータ・セットを用い、50以上のフルネームに対応する47の省略名で評価した。その結果、2変数混合モデルは、適合率と再現率の良いバランスを実現することが分かった。

In this paper, we propose a clustering method for disambiguating abbreviated author names appearing in citation data by finding the correct full name for each instance of an abbreviated name. We use the standard naive Bayes mixture model and the two-variable mixture model, which is a newly proposed model having two hidden variables. In the experiment, we have used the DBLP data set and have selected 47 abbreviated author names corresponding to more than or equal to 50 full names for evaluation. The results show that our model can achieve a good balancing of precisions and recalls.

### 1. はじめに

実世界のデータを活用するときよく直面する問題に名前  
の曖昧性解消がある。これは、(1)同じものが違う名で指されるとき同じものを指す名だけを正しく束ねる問題であるし、(2)異なるものが同じ名で指されるとき同じ名の個々の出現にそれが指示するものを正しく対応付ける問題でもある。本論文では、書誌情報において、複数の著者が同じ名で指されるとき、同じ名の複数の出現のうちどれが同じ著者を指すかを正しく言い当てるといったタイプ(2)の問題を扱う。実際、学術論文の参考文献一覧では、著者名の姓名の名がほぼイニシャルであり、個々の省略名が誰に対応するかが問題となる。

本論文では、名がイニシャルにされた省略名にフルネームを正しく対応付けるという問題に取り組む。同じフルネームが異なる著者を指すこともあるが、正解データを作る困難さもあり、本論文では扱わなかった。また、引用データに含まれないこともあるデータ（ページ数、年号、雑誌の巻号、会議の開催場所など）は用いず、共著者名、タイトル、雑誌名または国際会議名という3つのフィールドだけを用いる。この3フィールドからなるデータを、以下、引用データと呼ぶ。

もちろん、著者のフルネームが登録されたデータベースが既にあれば、タイトルや雑誌・会議名のマッチングにより、

与えられた引用データに対応する書誌情報を探すことで、フルネームは分かるだろう。しかし、本論文では、このような整備されたデータベースが利用できない状況を想定する。

省略著者名の曖昧性解消問題を解く手順は、次の通りである。まず、特定の1つの省略名を共著者名フィールドに含む引用データを収集する。そして、集めた引用データをクラスタリングし、同じクラスタに属する引用データに現れる省略名は同じフルネームに対応すると解釈する。評価に際しては、このクラスタリング結果を正解データとつきあわせる。

著者名曖昧性解消問題を解くために、本論文では2つの確率モデルを用いる。ひとつはナイーブ・ベイズ混合モデル(NBM)である[9]。同じ省略名を含む引用データ全てを生成する混合多項分布を想定し、パラメータを最尤推定する。そして、各引用データが、混合されている多項分布のどれによって生成されたとするのが妥当かを判定する。混合される多項分布の個数がクラスタ数であり、これは事前に指定する。著者名曖昧性解消では、真のクラスタ数は、与えられた省略著者名に対応するフルネームの数である。本論文で用いるも2つめのモデルは、新たに提案するモデルであり、依存関係にある隠れ変数を2つ備え、その値の組み合わせがクラスタのIDとなる。また、このモデルでは、タイトルと、雑誌・会議名とは、別々の1つの隠れ変数の値に依存して生成され、共著者名だけが2つの隠れ変数の値に同時に依存して生成される。このモデルを2変数混合モデル(TVM)と呼ぶ。NBMでは3フィールドが同じ1つの隠れ変数の値に依存して生成されるため、これらが同じ粒度で区別され、その結果、大きな多様性を示すタイトル・フィールドが、クラスタリング結果を大きく左右する。その一方、TVMでは、共著者名だけが2つの隠れ変数に同時に依存して生成されるので、共著者名を区別する粒度よりも、タイトルや雑誌・会議名を区別する粒度が粗くなる。評価実験では、DBLP[1]が公開しているデータ・セットからdblp20040213.xml.gzという長期間同内容で公開されているものを用い、実験の再現性を確保した。評価実験は、真のクラスタ数が未知と想定する場合と既知と想定する場合の両方でおこなう。

本論文の構成は次のとおりである。2節で先行研究を紹介する。3節で2つの確率モデルを定式化し、パラメータ推定のためのEMアルゴリズムを説明する。4節では、実験の内容を詳述し、5節でその結果を提示した後、6節で全体をまとめる。

### 2. 先行研究

姓名の名がイニシャルに略された著者名に、正しくフルネームを対応させる問題には、Hanら[4]が教師あり学習を採用して解法を提示している。しかし、省略著者名のすべてに訓練データを準備することは現実的でない。よってDongらの研究[3]やKalashnikovらの研究[7]は、教師なし学習でこの問題を解いている。しかも、1節で示したタイプ(1)とタイプ(2)の曖昧性を同時に解消する手法を提案している。しかし、いずれもメール・アドレスや所属機関名など、引用データから得られない情報の存在を想定している。本論文では、共著者名、タイトル、雑誌・会議名という3フィールドだけを使うため、より難しい問題を解くことになるが、タイプ(2)の曖昧性の解消だけに問題を限定する。さらにHanら[6]がスペクトラル・クラスタリングによって、またHanら[5]が確率モデルによって、やはり教師なし学習による解法を提案しており、ともにDBLPのデータを評価に用いている。だが、いずれも、真のクラスタ数、つまり、与えられた省略著者名に対応する

<sup>\*</sup> 正会員 長崎大学工学部 [masada@cis.nagasaki-u.ac.jp](mailto:masada@cis.nagasaki-u.ac.jp)

<sup>\*</sup> 正会員 国立情報学研究所 [takasu.adachi@nii.ac.jp](mailto:takasu.adachi@nii.ac.jp)

フルネームの数が既知と想定している。本論文では、真のクラスタ数を既知と想定する場合だけでなく、真のクラスタ数よりかなり大きい数をすべての省略名について共通してクラスタ数として用いる場合の実験もおこなう。また[5][6]では評価にmicroaveraged precisionしか使っていないが、本論文ではこれを含む4種の尺度を使う。

### 3. クラスタリングのための確率モデル

曖昧性解消問題の入力は、特定の省略著者名を含む引用データの集合  $D = \{d_1, \dots, d_j\}$  である。各引用データは、共著者名、タイトル、雑誌・会議名の3フィールドからなる。 $D$ に現れる省略名の集合を  $A = \{a_1, \dots, a_u\}$ 、雑誌・会議名の集合を  $B = \{b_1, \dots, b_v\}$  とする。各引用データは、ちょうど1つの雑誌・会議名を含む。また、 $D$ に属する引用データのタイトルに現れる語彙の集合を  $W = \{w_1, \dots, w_j\}$  とする。共著者の順序や、タイトルでの単語の順序は問わない。目標は、 $D$ をクラスタに分け、 $D$ を得るために使った省略名の曖昧性を解消することである。理想的なクラスタリングでは、同じフルネームに対応する省略名すべてが、ちょうど1つのクラスタに属する引用データのすべてに現われる。

#### 3.1 ナイーヴ・ベイズ混合モデル(NBM)

ナイーブ・ベイズ混合モデル(NBM)は1つの隠れ変数を持つ。この隠れ変数を取る値の集合を  $C = \{c_1, \dots, c_k\}$  とする。これらの値はクラスタのIDとみなされる。NBMでは1つの引用データ  $d_i$  が次のように生成される。まず、隠れ変数の値が多項分布  $P(c_k)$  にしたがって  $C$  からひとつ選ばれる。この値を  $c_k$  とする。次に  $c_k$  に対応する多項分布  $P(a_u | c_k)$  にしたがって、 $d_i$  の共著者数だけ省略著者名が  $A$  から選ばれる。タイトルを構成する単語も、 $c_k$  に対応する多項分布  $P(w_j | c_k)$  にしたがって  $d_i$  のタイトルの長さだけ  $W$  から選ばれる。雑誌名・会議名も、 $c_k$  に対応する多項分布  $P(b_v | c_k)$  にしたがって  $B$  からひとつ選ばれる。なお共著者数やタイトルの長さは明示的にモデル化しない[9]。こうして1つの引用データ  $d_i$  が生成される。 $d_i$  に省略名  $a_u \in A$  が現れる回数を  $\alpha(I, u)$ 、 $d_i$  のタイトルに単語  $w_j \in W$  が現れる回数を  $n(i, j)$  とする。さらに  $d_i$  の投稿された雑誌・会議名が  $b_v \in B$  のとき1となり、それ以外るとき0となる値を  $\delta(i, v)$  とする。このとき NBM によって引用データ  $d_i$  が生成される確率は  $P(d_i) = \sum_k P(c_k) P(d_i | c_k)$  と書ける。ただし  $P(d_i | c_k)$  は

$$\prod_u P(a_u | c_k)^{\alpha(i, u)} \prod_j P(w_j | c_k)^{n(i, j)} \prod_v P(b_v | c_k)^{\delta(i, v)} \quad (1)$$

に等しい。データ集合全体の尤度は  $P(D) = \prod_i P(d_i)$  である。詳細は割愛するが、NBMの場合、最尤推定によるパラメータ推定のためのEMアルゴリズムのEステップは

$$P^{\theta}(c_k | d_i) = P^{\theta-1}(d_i, c_k) / \sum_k P^{\theta-1}(d_i, c_k) \quad (2)$$

となる[9][10]。 $P^{\theta-1}(d_i, c_k)$  は  $P^{\theta-1}(c_k) P^{\theta-1}(d_i | c_k)$  に等しく、 $P^{\theta-1}(c_k)$  は、ひとつ前のMステップで得られており、 $P^{\theta-1}(d_i | c_k)$  もひとつ前のMステップで得られたパラメータ値によって式(1)から計算できる。Mステップでのパラメータ値の更新式は[11]を参照されたい。

今回の実験では30回の反復計算で十分な収束が得られた。計算が収束した後、各  $d_i$  について  $P(c_k | d_i)$  を最大とする  $c_k$  を  $d_i$  が属するクラスタのIDとみなす。よって  $c_1, \dots, c_k$  のうち、どの引用データについても  $P(c_k | d_i)$  を最大にしなかったものは、空のクラスタに対応すると言える。

#### 3.2 2変数混合モデル(TVM)

本論文が提案する2変数混合モデル(TVM)は、2つの隠れ変数をもつ。これらの変数を取る値の集合を各々  $Y = \{y_1, \dots, y_s\}$ 、 $Z = \{z_1, \dots, z_t\}$  とする。そして、これら2種類の値のペアをクラスタのIDとみなす。TVMでは、1つの引用データ  $d_i$  が次のように生成される。まず、一方の隠れ変数の値が多項分布  $P(y_s)$  にしたがって  $Y$  から1つ選ばれる。これを  $y_s$  とする。次に、 $y_s$  に対応する多項分布  $P(b_v | y_s)$  にしたがって、雑誌・会議名が1つ選ばれる。もう一方の隠れ変数の値も、 $y_s$  に対応する多項分布  $P(z_t | y_s)$  にしたがって  $Z$  から1つ選ばれる。この値を  $z_t$  とする。そして、この  $z_t$  に対応する多項分布  $P(w_j | z_t)$  にしたがって、 $d_i$  のタイトルの長さだけ単語が  $W$  から選ばれる。最後に、隠れ変数の値の組み合わせ  $(y_s, z_t)$  に対応する多項分布  $P(a_u | y_s, z_t)$  にしたがって、 $d_i$  の共著者数だけ省略名が  $A$  から選ばれる。ここでも、共著者数やタイトルの長さはモデル化しない。TVMでは、雑誌・会議名とタイトルとは、1つの隠れ変数のみに依存して生成され、共著者名だけが2つの隠れ変数に依存して生成される。これにより、共著者名の生成に寄与する多項分布のパリエーションは、引用データのクラスタの粒度と一致するようにし、雑誌・会議名とタイトルとは、よりパリエーションの乏しい多項分布群から生成されるようにした。なぜなら、先行研究[5][6]の指摘によると、書誌情報の著者名曖昧性解消では、共著者名が最も有効な情報を与えるためである。TVMによって引用データ  $d_i$  が生成される確率は

$$P(d_i) = \sum_s \sum_t P(y_s) P(z_t | y_s) P(d_i | z_t, y_s) \quad (3)$$

となる。ただし  $P(d_i | z_t, y_s) = \prod_u P(a_u | z_t, y_s)^{\alpha(i, u)} \prod_j P(w_j | z_t)^{n(i, j)} \prod_v P(b_v | y_s)^{\delta(i, v)}$  と計算される。詳細は割愛するが、このTVMについて、EMアルゴリズムのEステップは

$$P^{\theta}(y_s, z_t | d_i) = P^{\theta-1}(d_i, y_s, z_t) / \sum_s \sum_t P^{\theta-1}(d_i, y_s, z_t)$$

となる。 $P^{\theta-1}(d_i, y_s, z_t)$  は  $P^{\theta-1}(y_s) P^{\theta-1}(z_t | y_s) P^{\theta-1}(d_i | y_s, z_t)$  に等しく、 $P^{\theta-1}(y_s) P^{\theta-1}(z_t | y_s)$  はEMアルゴリズムの1つ前のMステップで得られたパラメータ値であり、 $P^{\theta-1}(d_i | y_s, z_t)$  は、同じく1つ前のMステップで得られたパラメータ値から式(3)により求められる。Mステップでのパラメータ値の更新式は[11]を参照されたい。TVMも30回の反復計算で十分な収束が得られた。計算の収束後、各  $d_i$  について  $P(y_s, z_t | d_i)$  を最大とする隠れ変数値のペア  $(y_s, z_t)$  を、 $d_i$  が属するクラスタのIDとみなす。隠れ変数値のペアは  $ST$  通りあるが、どの  $d_i$  についても  $P(y_s, z_t | d_i)$  を最大にしなかった  $(y_s, z_t)$  は、空のクラスタに対応すると言える。なお実験では  $S=T$  となるように設定した。なぜなら、予備実験によると  $S \neq T$  以外の場合は興味深いふるまいを示さなかったからである。また、TVMの2つの隠れ変数の役割を入れ替えたモデルも考えられるが、やはり予備実験で興味深い違いを示さなかったため、上に提示したTVMだけを扱う。なお、NBMとTVMとで、EMアルゴリズムにスムージングやアニーリングを併用したが、その詳細は[11]を参照されたい。

## 4. 実験と評価

### 4.1 実験方法

実験では、DBLP書誌情報データベース[1]が公開しているデータ・セットのうち、長期間同じ内容で公開されている `dblp20040213.xml.gz` というデータ・セットを用いた。共著者名、タイトル、雑誌・会議名という3つのフィールドを備

えていないデータや、著者名の姓名の名が元々イニシャルになっているデータは除去した。残ったデータで著者名の名をイニシャルにし、こうして得られた省略著者名のうち、対応するフルネームが 50 以上ある 47 の省略名[11]を実験に使った。タイトルからは stop word を除去し porter stemmer[2] を適用した。47 の省略名の各々について、次のような手順を踏んだ。例えば、`S. Lee` について実験する場合、まずこの省略名を含む引用データを集め、引用データ集合  $D$  を作成する。そして、 $D$  を以下の 3 通りの方法でクラスタリングする。1) NBM をそのまま用いる。このクラスタリング方法を NBM と書く。2) 共著者名フィールドだけを残し NBM を用いる。この方法を NBMa と書く。3) TVM を用いる。この方法を TVM と書く。なお、3 つのクラスタリング手法すべてで、ランダムに決めた 10 通りの異なる初期値から EM アルゴリズムを開始する。クラスタ数は、真のクラスタ数が未知と想定する場合は、NBM, NBMa では  $K=256$ , TVM では  $S=T=16$  と設定し、真のクラスタ数が既知と想定する場合は、TVM で  $S, T$  を真のクラスタ数の正の平方根を切り上げた自然数、NBM, NBMa では  $K=ST$  と決めた。計算時間は、クラスタ数が未知の場合、省略名 `S. Lee` について、30 回の反復計算で NBM が約 19 秒、TVM が約 16 秒、NBMa が約 6 秒 (Xeon 3.20GHz, 全データがメモリ上) だった。

### 4.2 評価方法

クラスタリング結果の評価方法は、以下のとおりである。例えば `S. Lee` を含む引用データの集合  $D$  に、NBM, NBMa, TVM いずれかの方法で得たクラスタリング結果を  $\mathcal{G}$  とする。 $D$  に含まれる引用データの各々について、`S. Lee` と略される前のフルネームを、元のデータに戻って確認する。そして、各クラスタ  $G \in \mathcal{G}$  に属する引用データを元のデータで確認したとき、最も多くのデータに現われるフルネーム、例えば `Sunghyun Lee` を  $G$  のラベルと呼ぶ。 $G$  に属するデータの数を  $N_{size}(G)$ 、そのうち  $G$  のラベルが現われるデータ数を  $N_{pos}(G)$  とする。また、 $D$  の全引用データのうち  $N_{cor}(G)$  個に  $G$  のラベルが現われているとする。このとき  $G$  の precision を  $N_{pos}(G) / N_{size}(G)$ , recall を  $N_{pos}(G) / N_{cor}(G)$  と定義する。 $\mathcal{G}$  自体の評価には、下記の 4 種類の値を使う。

$$P_{mac}(\mathcal{G}) = \frac{\sum_{G \in \mathcal{G}} \frac{N_{pos}(G)}{N_{size}(G)}}{|\mathcal{G}|}, R_{mac}(\mathcal{G}) = \frac{\sum_{G \in \mathcal{G}} \frac{N_{pos}(G)}{N_{cor}(G)}}{|\mathcal{G}|}$$

$$P_{mic}(\mathcal{G}) = \frac{\sum_{G \in \mathcal{G}} N_{pos}(G)}{\sum_{G \in \mathcal{G}} N_{size}(G)}, R_{mic}(\mathcal{G}) = \frac{\sum_{G \in \mathcal{G}} N_{pos}(G)}{\sum_{G \in \mathcal{G}} N_{cor}(G)}$$

順に、macroaveraged precision/recall, microraveraged precision/recall である。以上 4 つの評価値を、NBM, NBMa, TVM それぞれで、10 通りのランダムな初期値から出発したクラスタリング結果について計算する。そして、10 通りの結果の評価値の平均と標準偏差を求め、NBM, NBMa, TVM それぞれの、特定の省略名に関する曖昧性解消の性能とする。

## 5. 実験結果

### 5.1 真のクラスタ数を未知と想定した場合

図 1 は、真のクラスタ数を未知と想定した場合に得られた空でないクラスタの数を、各省略名ごとに示している。値は、10 通りのパラメータ初期値に対応する結果の平均である。

マーカは標準偏差の  $\pm 1$  倍の幅を示す。×印が各省略名に対応するフルネーム数、つまり真のクラスタ数である。NBMa では、多くの省略名で非空のクラスタ数が真のクラスタ数に迫っている。全体として、NBM, TVM, NBMa の順でクラスタの数が多く、NBM で細分化が甚だしい。これは、NBMa では共著者名しか用いなかった結果、引用データの多様性が減った一方、NBM ではタイトルを用いることで引用データの多様性が増し、引用データが異なるクラスタに分散しやすくなったためだろう。TVM では、タイトルの生成が 1 つの隠れ変数にのみ依存するため、中間的にふるまったと考えられる。だが、NBMa には次のような問題があった。図 2 は、少なくともひとつのクラスタのラベルとなりえたフルネームの数を、各省略名について、10 通りの結果の平均と標準偏差とで示しているが、NBMa ではこの値が低く、よって NBMa では、多くのフルネームを発見し損なってしまった。

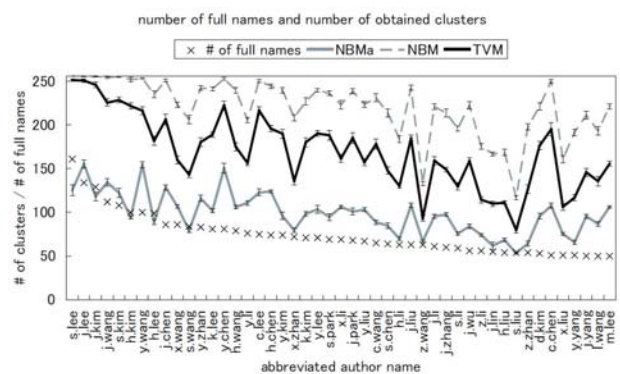


図 1 真のクラスタ数が未知の場合の非空のクラスタ数  
Fig. 1 Number of non-empty clusters under the assumption that we do not know the true cluster number

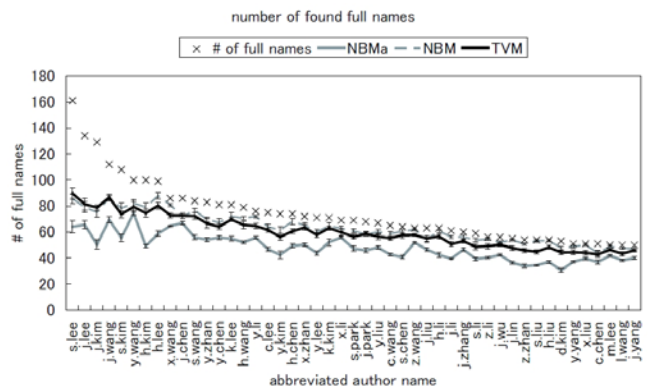


図 2 真のクラスタ数が未知の場合に見つかったクラスタ数  
Fig. 2 Number of found clusters under the assumption that we do not know the true cluster number

次に、各省略名についてのクラスタリング結果を  $P_{mic}$ ,  $R_{mic}$ ,  $P_{mac}$ ,  $R_{mac}$  の 4 つの評価値で評価し、10 通りの値の平均と標準偏差を求め、さらに 47 の省略名全体で平均をとった結果を表 1 にまとめた。 $F_{mic}$ ,  $F_{mac}$  は、それぞれ  $P_{mic}$  と  $R_{mic}$ ,  $P_{mac}$  と  $R_{mac}$  の調和平均である。 $P_{mic}$  は、クラスタが細分化されるほど高くなりやすく、サイズの大きなクラスタの precision に強く影響される。 $R_{mic}$  は、クラスタが細分化されるほど低くなりやすく、 $D$  内で頻繁に出現するフルネームをラベルとするクラスタの recall に強く影響される。 $P_{mac}$  は、 $P_{mic}$  と同様に細分化が甚だしいほど高くなるが、

すべてのクラスタの precision が平等に寄与する。  $R_{mac}$  は、  $R_{mic}$  と同様に細分化が甚だしいと低くなるが、すべてのクラスタの recall が平等に寄与する。 precision は NBM より NBMa が良いが、 recall は NBMa より NBM が良い。これは、NBM が NBMa より大きな多様性を示すデータを使っており、細分化を起こしやすいためであろう。 TVM は、 precision と recall の両方で、NBM と NBMa の中間の値を示している。つまり、2変数混合モデルは、 precision と recall の良いバランスを与えていると言える。  $F_{mic}$ 、  $F_{mac}$  で見れば NBMa が最も良いが、今回のように recall を上げることが難しい状況では、理想的なクラスタリングへ近づくために、元々高い precision を下げてでも recall を上げるよりは、同じフルネームに対応する引用データが複数のクラスタに分かれることで recall は下がってしまっても、個々のクラスタはうまくできていて precision がある程度高い、という方向を目指すのが望ましいと考える。これは、特定のフルネームを探すには複数のクラスタを調べる必要があるものの、個々のクラスタの質は高い、という状況を目指すことに対応する。この意味では TVM はバランスがとれていると言える。

表 1 真のクラスタ数を未知とした場合の評価結果

Table 1 Evaluation results under the assumption that we do not know the true number of clusters

方法	$P_{mic}$	$P_{mac}$	$R_{mic}$	$R_{mac}$	$F_{mic}$	$F_{mac}$
NBMa	0.6295	0.8653	0.1477	0.3845	0.2312	0.5274
NBM	0.8595	0.9013	0.0784	0.2610	0.1415	0.4019
TVM	0.7866	0.8720	0.1034	0.3109	0.1784	0.4539

## 5.2 真のクラスタ数を既知と想定した場合

表 2 は真のクラスタ数が既知と想定する場合に、4通りの評価値について、すべての省略名にわたって平均を求めた結果である。 Precision と Recall の大小の傾向は表 1 と同様だが、真のクラスタ数が既知としているため、無駄なクラスタができにくく、表 1 より recall が高い。まとめると、NBMa は、真のクラスタ数が未知でも妥当な数のクラスタを与え、また  $F_{mic}$ 、  $F_{mac}$  で最も優れた結果を示すが、多くのフルネームを見つけ損ねるという欠点をもつ。 TVM は、NBM と同程度のフルネームを見つけると同時に、高い precision をあまり落とさず recall を上げるという方向で NBM より優れている。

表 2 真のクラスタ数を既知とした場合の評価結果

Table 2 Evaluation results under the assumption that we know the true number of clusters

方法	$P_{mic}$	$P_{mac}$	$R_{mic}$	$R_{mac}$	$F_{mic}$	$F_{mac}$
NBMa	0.5506	0.7687	0.1813	0.4189	0.2638	0.5378
NBM	0.5566	0.5627	0.1283	0.3060	0.2037	0.3935
TVM	0.5680	0.6620	0.1453	0.3478	0.2252	0.4527

## 6. おわりに

本論文では、姓名の名がイニシャルにされたかたちで引用データに現われる著者名に、正しくフルネームを対応付けるという意味での、著者名の曖昧性解消問題に取り組んだ。実験の結果、真のクラスタ数が未知とした場合、クラスタの過細分化が起こり、3つのどの方法でも recall が低くなった。 NBMa では、非空クラスタの数が真のクラスタ数に近かったが、見つけ損なったフルネームの数も多かった。どの省略著者名についても、TVM は NBMa と NBM 間の中間的な結果

を示しており、2変数混合モデルのねらいを反映していた。つまり、できるだけ多くのフルネームを見つけつつ、 precision と recall のバランスをとりたい場合は、TVM を使うとよい。だが、やはり全体として性能は高くない。実用に耐える著者名曖昧性解消システムをつくるには、例えば、引用データが元々そこから取ってこられた論文の情報を保存しておき、それらの論文に現われる様々な情報との依存関係を、積極的にモデルに組み込む必要があると思われる。

## 【文献】

- [1] <http://www.informatik.uni-trier.de/~ley/db/>
- [2] <http://www.tartarus.org/~martin/PorterStemmer/>
- [3] Dong, X., Halevy, A., and Madhavan, J.: "Reference Reconciliation in Complex Information Spaces", Proc. of SIGMOD, pp. 85-96 (2005).
- [4] Han, H., Giles, C. L., Zha, H., Li, C., and Tsioutsoulklis, K.: "Two Supervised Learning Approaches for Name Disambiguation in Author Citations", Proc. of JCDL, pp. 296-305 (2004).
- [5] Han, H., Xu, W., Zha, H., and Giles, C. L.: "A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations", Proc. of SAC, pp. 1065-1069 (2005).
- [6] Han, H., Zha, H., and Giles, C. L.: "Name disambiguation in author citations using a  $k$ -way spectral clustering method", Proc. of JCDL, pp. 334-343 (2005).
- [7] Kalashnikov, D. V., Mehrotra, S., and Chen, Z.: "Exploiting Relationships for Domain-Independent Data Cleaning", Proc. of SDM (2005).
- [8] Rose, K., Gurewitz, E., and Fox, G.: "A Deterministic Annealing Approach to Clustering", Pattern Recognition Letters, Vol. 11, pp. 589-594 (1990).
- [9] Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. M.: "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning, Vol. 39, No. 2/3, pp. 103-134 (2000).
- [10] 上田修功: "ベイズ学習 [I] ---統計的学習の基礎---", 電子情報通信学会誌, Vol. 85, No. 4, pp. 265-271 (2002).
- [11] 正田備也, 高須淳宏, 安達淳: 書誌情報における著者名の曖昧性解消のためのクラスタリング手法の提案, 第 18 回データ工学ワークショップ (DEWS2007), L1-4 (2007).

## 正田 備也 Tomonari MASADA

長崎大学工学部情報システム工学科助教。2004 東京大学大学院情報理工学系研究科博士課程修了。博士 (情報理工学)。テキストマイニング、情報検索の研究に従事。情報処理学会正会員。日本データベース学会正会員。

## 高須 淳宏 Atsuhiro TAKASU

国立情報学研究所教授。1989 東京大学大学院工学系研究科博士課程修了。工学博士。データ工学、特にデータ解析と解析モデルの学習の研究に従事。電子情報通信学会、情報処理学会、人工知能学会、ACM、IEEE 各会員。

## 安達 淳 Jun ADACHI

国立情報学研究所教授。東京大学大学院情報理工学系研究科教授を併任。1981 東京大学大学院工学系研究科博士課程修了。工学博士。データベースシステム、テキストマイニング、情報検索、電子図書館システム等の研究開発に従事。電子情報通信学会、情報処理学会、IEEE、ACM 各会員。