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

正田　備也† 　高須　淳宏†† 　安達　淳††

† 長崎大学工学部　〒 852-8521 長崎県長崎市文教町 1-14
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †masada@cis.nagasaki-u.ac.jp, ††{takasu,adachi}@nii.ac.jp

あらまし　本論文では，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を提案する．論文の多くの引用では，著者の姓名の名がイニシャルに省略される．そのため，論文情報データベースの構築時に，省略された著者名をフルネームに正しく対応付ける必要が生じる．本論文では，省略された共著者名，タイトル，雑誌名（または会議名）の 3 フィールドからなる引用データを，2 つの確率論的モデルでクラスタリングする．一方のモデルは，通常のナイーヴ・ベイズ混合モデルである．各引用データについて，最も高い確率を与える，隠れ変数の値によって，引用データをクラスタリングする．そして，同じクラスタに属するに引用データに現われる省略名は，すべて同じフルネームに対応すると考える．もう一方は，本論文が提案するモデルであり，隠れ変数をふたつ含む．これらの 2 変数の値の組み合わせで，引用データをクラスタリングする．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 の省略著者名を選んだ．各省略名について，それを含む全ての引用データを集め，クラスタリングをおこなった．実験の結果，提案手法は，ナイーヴ・ベイズ混合モデルに比べ，適合率と再現率のよりよいバランスを実現することが分かった．

キーワード　クラスタリング，教師なし学習，名前曖昧性解消

Proposal of a New Clustering Method for Name Disambiguation in Author Citations

Tomonari MASADA†, Atsuhiro TAKASU††, and Jun ADACHI††

† Faculty of Engineering, Nagasaki University　Bunkyo-machi 1–14, Nagasaki, 852–8521 Japan
†† National Institute of Informatics　Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †masada@cis.nagasaki-u.ac.jp, ††{takasu,adachi}@nii.ac.jp

Abstract　In this paper, we propose a clustering method for author name disambiguation in citation data. Most article citations include first names of authors with their initials. Therefore, we need to disambiguate the abbreviated author names and to find the correct full name for each of them when constructing a bibliographic database. In this paper, we obtain a clustering of citation data, which consist of the three fields, i.e., co-author names, title words, journal or proceeding title words, with the two probabilistic models. The one model is a standard naive Bayes mixture model. For each citation data, we regard the most probable value of the hidden variable as the ID of the cluster to which the data belongs. Then all abbreviated name instances appearing in the same citation data cluster are taken as the abbreviation of the same full name. The other is a newly proposed model, which has two hidden variables. We partition citation data into clusters according to the most probable combination of the two hidden variable values. In the experiment, we used the DBLP bibliographic data set and selected 47 abbreviated author name to which more than or equal to 50 full names correspond. For each abbreviated name, we collected all citation data including the name and partition the citation data into clusters. The result of experiments shows that our new model can achieve a better balancing of precisions and recalls than the naive Bayes mixture model.

Key words　clustering, unsupervised learning, name disambiguation

1. はじめに

実世界から得られる様々なデータを活用するにあたって，頻繁に直面する問題のひとつに，名前の曖昧性解消 (name disambiguation) が挙げられる．これは，(1) 同じものが異なる名前によって指されている場合に，同じものを指す名前だけを，正しく束ねる問題であったり，(2) 異なるものが同じ名前によって指されている場合に，同じ名前の個々の出現について，それが指示するものを正しく対応付ける問題であったりする．

本論文では，論文の引用データにおいて，複数の著者が同じ名前で指されているとき，同じ名前の個々の出現について各々が指す著者を正しく言い当てる，という問題を扱う．つまり，(2) のタイプの曖昧性解消問題を引用データに現われる著者名について考える．実際，多くの学術論文末尾の参考文献リストでは，著者名の姓名の名がイニシャルにされている．このような引用データを書誌情報データベースに追加する際は，個々の省略著者名が誰を指すのか，正しく判定する必要がある．もちろん，各々の引用データが表す論文について，予め詳細な書誌情報を持っているならば，文字列マッチング等により，省略名が指示する著者を判定できるだろう．本論文では，このように詳細な書誌情報を予め保有していないとき，姓名の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，曖昧性解消問題に取り組む．現実には同じフルネームが異なる著者を指しうるが，正解データを作ることも困難であり，本論文では扱わなかった．しかし，フルネームの同姓同名問題にも，類似の手法が適用可能と考えている．さらに，現実の引用データに現われる省略名からフルネームを再現するという状況を想定しているため，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 つのフィールドの情報だけを用いる．以下，これら 3 つのフィールドからなると見なされたデータを，単に引用データと呼ぶ．

省略著者名の曖昧性解消問題を解くために，本論文では 2 つの確率論的なモデルを用いる．問題を解く手順は，次の通りである．まず，例えば “S. Lee” など，1 つの省略著者名を固定する．次に，この “S. Lee” を共著者名に含む引用データをすべて集める．これは，書誌情報データベースでの，省略名を用いた検索の場面を想定している．そして，この引用データ群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われる “S. Lee” は，すべて同じフルネームに対応すると解釈する．本論文で用いる 2 つのモデルのうちの一方は，ナイーヴ・ベイズ混合モデルである [9]．各引用データを bag of words と見なし，同じ省略名を含む引用データ群全てを生成する混合多項分布を想定して，そのパラメータを推定する．そして，各引用データが，混合されている多項分布のうちどの多項分布によって生成されたとするのが妥当かを判定することで，クラスタリングをおこなう．混合される多項分布の個数，すなわちクラスタ数は，あらかじめ指定する必要がある．著者名の曖昧性解消の場合，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となる．実験は，真のクラスタ数が未知の場合と既知の場合との両方で行う．

本論文で用いるもう一方のモデルは，新たに提案するモデルである．ナイーヴ・ベイズ混合モデルでは，混合されたどの多項分布によってデータが生成されるかを表すものとして，1 つの隠れ変数が用意されていた．新たに提案するモデルでは，依存関係にある隠れ変数を 2 つ用意し，それらの値の組み合わせによって，引用データをクラスタリングする．また，このモデルでは，タイトルと，雑誌名・会議名との出現確率は，それぞれ別の 1 つだけの隠れ変数に依存し，共著者名の出現確率だけが 2 つの隠れ変数に同時に依存する．以下，このモデルを **2 変数混合モデル**と呼ぶ．2 変数混合モデルを使うと，共著者名を区別する粒度よりも，タイトルや雑誌名・会議名を区別する粒度のほうが粗くなる．ナイーヴ・ベイズ混合モデルの場合は，共著者名，タイトル，雑誌名・会議名の 3 つのフィールドの情報が，すべて同じ粒度によって区別されるため，3 つのうち最も多様性を示しやすいタイトル・フィールドが，クラスタリングの結果を大きく左右する．そこで，曖昧性解消に共著者名の違いがより大きく寄与するように，新しいモデルを提案した．

評価実験で用いるデータとしては，実験の再現性を確保するため，DBLP [1] が一般に公開しているデータ・セットを用いた．しかも，頻繁に内容が変わる最新のデータ・セットではなく，長期間同じ内容のまま公開されているものを用いた．

本論文の構成は次のとおりである．2. 節では，著者名の曖昧性解消問題に取り組んだ先行研究を紹介する．3. 節では，ナイーヴ・ベイズ混合モデルと 2 変数混合モデルを定式化し，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，クラスタリング結果の評価を提示する．最後に 5. 節では，全体のまとめと今後の課題を述べる．

2. 先行研究

姓名の名がイニシャルに略された著者名に，正しく元のフルネームを割り当てるという問題は，すでに多く研究されている．Han ら [4] は，本論文と同じく DBLP のデータを使い，教師あり学習によってこの問題を解いている．しかし，省略著者名のすべてについて学習のための訓練データを準備することは，非現実的である．よって，教師なし学習を採る研究が近年増えている．Dong らの研究 [3] や Kalashnikov らの研究 [7] は，教師なし学習によって名前の曖昧性解消問題を解いている．しかも，1. 節で示した (2) のタイプの曖昧性だけでなく，(1) のタイプの曖昧性も同時に解消する手法を提案している．しかし，いずれの研究も，メール・アドレスや所属機関名など，引用データからは得られない情報も利用できると想定している．本論文では，共著者名，タイトル，雑誌名・会議名という 3 種類の情報のみを使う．そのため，より難しい問題を解くことになるが，かわりに，解くべき問題のタイプを (2) のタイプに限定している．

教師なし学習によって省略著者名にフルネームを正しく対応付けるという問題には，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] が確率論的なモデルによって，それぞれ解法を提案している．ともに DBLP の書誌データで評価をおこなっている．しかし，いずれの研究も，真のク

ラスタ数，つまり，曖昧性を解消すべき省略著者名に対応するフルネームの数が，既知と想定している．本論文では，真のクラスタ数が既知の場合の実験もおこなうが，真のクラスタ数よりも多い一定の値を，様々な省略著者名の曖昧性解消で共通して用いるという実験もおこなう．また，これらの研究は評価尺度に microaveraged precision しか使っていない．本論文では microaveraged precision/recall，macroaveraged precision /recall という 4 種類の評価尺度を使うことにする．

3. クラスタリングのためのデータ生成モデル

3.1 引用データ

曖昧性解消問題の入力として与えられるのは，特定の 1 つの省略著者名を含む全引用データの集合 $D = \{d_1, \ldots, d_I\}$ である．各引用データは，共著者名，タイトル，雑誌名または国際会議名の 3 つのフィールドからなるとする．D に属する引用データに現われる省略著者名の集合を $A = \{a_1, \ldots, a_U\}$，雑誌名・会議名の集合を $B = \{b_1, \ldots, b_V\}$ とする．各引用データは，ちょうど 1 つの雑誌名・会議名だけを含む．また，D に属する引用データのタイトルに含まれる語彙の集合を $W = \{w_1, \ldots, w_J\}$ とする．共著者名に含まれる省略著者名の順序や，タイトルに含まれる単語の順序は問わないことにする．

D をクラスタに分けることで，D を得るために使った省略著者名の曖昧性を解消することが目標である．理想的なクラスタリングとは，各クラスタに属する引用データにおいて，そこに現われる省略名がすべて同じフルネームに対応しており，かつ，同じフルネームに対応する省略名のすべてが，1 つのクラスタに属する引用データに現われるようなクラスタリングである．

3.2 ナイーヴ・ベイズ混合モデル

ナイーヴ・ベイズ混合モデルは，1 つの隠れ変数を持つ．この隠れ変数が取りうる値の集合を $C = \{c_1, \ldots, c_K\}$ とする．これらの値は，クラスタの ID とみなすことができる．ナイーヴ・ベイズ混合モデルでは，1 つの引用データ d_i が次のように生成されると考える．まず，隠れ変数の値が，多項分布 $P(c_k)$ s.t. $\sum_{k=1}^{K} P(c_k) = 1$ にしたがって C からひとつ選ばれる．次に，選ばれた隠れ変数の値に対応する多項分布 $P(a_u|c_k)$ s.t. $\sum_{u=1}^{U} P(a_u|c_k) = 1$ にしたがって，d_i の共著者数だけ共著者名が A から選ばれる．タイトルを構成する単語も，選ばれた隠れ変数の値に対応する多項分布 $P(w_j|c_k)$ s.t. $\sum_{j=1}^{J} P(w_j|c_k) = 1$ 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．さらに，雑誌名・会議名も，選ばれた隠れ変数の値に対応する多項分布 $P(b_v|c_k)$ s.t. $\sum_{v=1}^{V} P(b_v|c_k) = 1$ にしたがって，1 つ B から選ばれる．共著者数やタイトルの長さは明示的にモデル化しないことにする [9]．こうして 1 つの引用データ d_i が生成される．

引用データ d_i に現われる著者名 $a_u \in A$ の個数を o_{iu}，d_i のタイトルに含まれる単語 $w_j \in W$ の個数を c_{ij} とする．また，d_i が投稿された雑誌名・会議名が $b_v \in B$ のとき 1 となり，それ以外のとき 0 となる値を δ_{iv} とする．このとき，ナイーヴ・ベイズ混合モデルによって，引用データ d_i が生成される確率は

$$P(d_i) = \sum_{k=1}^{K} P(d_i, c_k) = \sum_{k=1}^{K} P(c_k) P(d_i|c_k) \tag{1}$$

と書ける．ただし $P(d_i|c_k)$ は

$$P(d_i|c_k) = \prod_{u=1}^{U} P(a_u|c_k)^{o_{iu}} \prod_{j=1}^{J} P(w_j|c_k)^{c_{ij}} \prod_{v=1}^{V} P(b_v|c_k)^{\delta_{iv}} \tag{2}$$

という式によって求められる．引用データ集合全体 D の尤度は $P(D) = \prod_{i=1}^{I} P(d_i)$ となる．詳細は割愛するが，このナイーヴ・ベイズ混合モデルの場合の，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

$$P(c_k|d_i) = \frac{\bar{P}(d_i, c_k)}{\sum_{k=1}^{K} \bar{P}(d_i, c_k)} \tag{3}$$

となる [9] [10]．なお，$\bar{P}(d_i, c_k)$ は，$\bar{P}(c_k)\bar{P}(d_i|c_k)$ に等しい．$\bar{P}(c_k)$ 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，$\bar{P}(d_i|c_k)$ は，同じく 1 つ前の M ステップで得られたパラメータ値を使って式 (2) により計算する．そして，M ステップでのパラメータ値の更新のための式は

$$\begin{aligned}
P(c_k) &= \frac{\sum_{i=1}^{I} P(c_k|d_i)}{\sum_{k=1}^{K} \sum_{i=1}^{I} P(c_k|d_i)} \\
P(a_u|c_k) &= \frac{\sum_{i=1}^{I} P(c_k|d_i) o_{iu}}{\sum_{u=1}^{U} \sum_{i=1}^{I} P(c_k|d_i) o_{iu}} \\
P(b_v|c_k) &= \frac{\sum_{i=1}^{I} P(c_k|d_i) \delta_{iv}}{\sum_{v=1}^{V} \sum_{i=1}^{I} P(c_k|d_i) \delta_{iv}} \\
P(w_j|c_k) &= \frac{\sum_{i=1}^{I} P(c_k|d_i) c_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{I} P(c_k|d_i) c_{ij}}
\end{aligned} \tag{4}$$

となる．今回の実験では，30 回の繰り返し計算で十分な収束が得られた．EM アルゴリズムの計算が収束した後，各引用データについて得られた $P(c_k|d_i)$ の値を見て，これを最大とする c_k を，その引用データ d_i が属するクラスタの ID とみなす．よって，$c_1, \ldots, c_K$ 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．

3.3 2 変数混合モデル

本論文が提案する 2 変数混合モデルは，2 つの隠れ変数をもつ．一方の隠れ変数が取りうる値の集合を $Y = \{y_1, \ldots, y_S\}$ とし，もう一方の隠れ変数が取りうる値の集合を $Z = \{z_1, \ldots, z_T\}$ とする．そして，これら 2 種類の値の組み合わせによって，引用データのクラスタを表現することにする．

2 変数混合モデルでは，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 $P(y_s)$ s.t. $\sum_{s=1}^{S} P(y_s) = 1$ にしたがって，Y から 1 つ選ばれる．次に，選ばれた隠れ変数の値に対応する多項分布 $P(b_v|y_s)$ s.t. $\sum_{v=1}^{V} P(b_v|y_s) = 1$ にしたがって，雑誌名・会議名が 1 つ選ばれる．また，もう一方の隠れ変数の値が，先に選ばれた隠れ変数の値に対応する多項分布 $P(z_t|y_s)$ s.t. $\sum_{t=1}^{T} P(z_t|y_s) = 1$ にしたがって，Z から 1 つ

選ばれる．そして，選ばれた第 2 の隠れ変数の値に対応する多項分布 $P(w_j|z_t)$ s.t. $\sum_{j=1}^{J} P(w_j|z_t) = 1$ 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，選ばれた 2 つの隠れ変数の値の組み合わせに対応する多項分布 $P(a_u|y_s, z_t)$ s.t. $\sum_{u=1}^{U} P(a_u|y_s, z_t) = 1$ にしたがって，d_i の共著者数だけ共著者名が A から選ばれる．ここでも，共著者数やタイトルの長さは，明示的にモデル化しないことにする．

2 変数混合モデルでは，雑誌名・会議名とタイトルは 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．これによって，共著者名の生成に寄与する多項分布のバリエーションだけが，引用データのクラスタの粒度と一致するようにし，雑誌名・会議名，および，タイトルは，よりバリエーションの乏しい多項分布によって生成されるようにしている．なぜなら，先行研究が指摘するように [4] [5]，引用データの著者名の曖昧性解消では，共著者名が最も有効な情報を与えるからである．

2 変数混合モデルによって引用データ d_i が生成される確率は

$$\begin{aligned} P(d_i) &= \sum_{s=1}^{S}\sum_{t=1}^{T} P(d_i, z_t, y_s) \\ &= \sum_{s=1}^{S}\sum_{t=1}^{T} P(y_s)P(z_t|y_s)P(d_i|z_t, y_s) \end{aligned} \tag{5}$$

となる．ただし $P(d_i|z_t, y_s)$ は

$$\begin{aligned} &P(d_i|z_t, y_s) \\ &= \prod_{u=1}^{U} P(a_u|z_t, y_s)^{o_{iu}} \prod_{j=1}^{J} P(w_j|z_t)^{c_{ij}} \prod_{v=1}^{V} P(b_v|y_s)^{\delta_{iv}} \end{aligned} \tag{6}$$

という式で計算される．やはり詳細は割愛するが，2 変数混合モデルの場合の，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

$$P(y_s, z_t|d_i) = \frac{\bar{P}(d_i, y_s, z_t)}{\sum_{s=1}^{S}\sum_{t=1}^{T} \bar{P}(d_i, y_s, z_t)} \tag{7}$$

となる．$\bar{P}(d_i, y_s, z_t)$ は，$\bar{P}(y_s)\bar{P}(z_t|y_s)\bar{P}(d_i|z_t, y_s)$ に等しい．$\bar{P}(y_s)$ と $\bar{P}(z_t|y_s)$ には，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，$\bar{P}(d_i|z_t, y_s)$ は，同じく 1 つ前の M ステップで得られたパラメータ値を使って式 (6) により求める．M ステップでのパラメータ値の更新のための式は

$$\begin{aligned} P(y_s) &= \frac{\sum_{t=1}^{T}\sum_{i=1}^{I} P(y_s, z_t|d_i)}{\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{i=1}^{I} P(y_s, z_t|d_i)} \\ P(z_t|y_s) &= \frac{\sum_{i=1}^{I} P(y_s, z_t|d_i)}{\sum_{t=1}^{T}\sum_{i=1}^{I} P(y_s, z_t|d_i)} \\ P(b_v|y_s) &= \frac{\sum_{t=1}^{T}\sum_{i=1}^{I} P(y_s, z_t|d_i)\delta_{iv}}{\sum_{v=1}^{V}\sum_{t=1}^{T}\sum_{i=1}^{I} P(y_s, z_t|d_i)\delta_{iv}} \\ P(w_j|z_t) &= \frac{\sum_{s=1}^{S}\sum_{i=1}^{I} P(y_s, z_t|d_i)c_{ij}}{\sum_{j=1}^{J}\sum_{s=1}^{S}\sum_{i=1}^{I} P(y_s, z_t|d_i)c_{ij}} \\ P(a_u|y_s, z_t) &= \frac{\sum_{i=1}^{I} P(y_s, z_t|d_i)o_{iu}}{\sum_{u=1}^{U}\sum_{i=1}^{I} P(y_s, z_t|d_i)o_{iu}} \end{aligned} \tag{8}$$

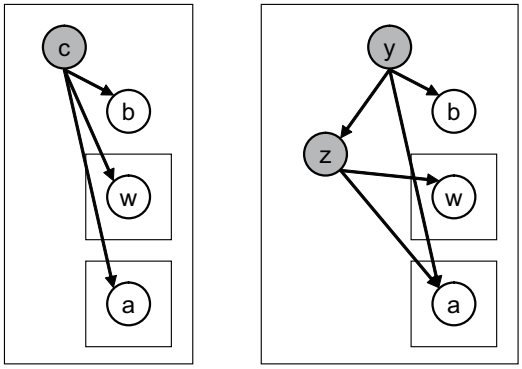

図 1　ナイーヴ・ベイズ混合モデル（左）と 2 変数混合モデル（右）のグラフィカルモデル

Fig. 1　Graphical models of naive Bayes mixture model (right) and two variable mixture model (left).

となる．2 変数混合モデルのほうも，今回の実験では 30 回の繰り返し計算で十分な収束が得られた．EM アルゴリズムの計算が収束した後，各引用データについて得られた $P(y_s, z_t|d_i)$ の値を見て，これを最大とする隠れ変数の値のペア (y_s, z_t) を，その引用データ d_i が属するクラスタの ID とみなす．隠れ変数が取りうる値のペアは ST 通りあるが，このうちどの引用データについても $P(y_s, z_t|d_i)$ を最大にしなかったものは，空のクラスタに対応する．実験では，$S = T$ となるように設定した．なぜなら，予備実験の結果，$S = T$ の場合以外は，ナイーヴ・ベイズ混合モデルと，$S = T$ に設定した 2 変数混合モデルとの，中間的なふるまいを示しただけだったからである．

2 変数混合モデルについては，2 つの隠れ変数の役割を入れ替えたモデルを考えることができる．このモデルでは，式 (5) が $P(d_i) = \sum_{s=1}^{S}\sum_{t=1}^{T} P(z_t)P(y_s|z_t)P(d_i|z_t, y_s)$ という式に置き換えられる．しかし，予備実験の結果，名前の曖昧性解消において興味深い違いを示さなかったため，上に提示した 2 変数混合モデルだけを扱う．ナイーヴ・ベイズ混合モデルと 2 変数混合モデルのグラフィカル・モデルを，図 1 に示す．

3.4　EM アルゴリズムの詳細

3.4.1　スムージング

ナイーヴ・ベイズ混合モデルの式 (4) で得られたパラメータ $P(a_u|c_k), P(b_v|c_k), P(w_j|c_k)$ の値を式 (3) で使う前に，

$$\begin{aligned} P(a_u|c_k) &= (1-\gamma)\frac{\sum_{i=1}^{I} P(c_k|d_i)o_{iu}}{\sum_{u=1}^{U}\sum_{i=1}^{I} P(c_k|d_i)o_{iu}} \\ &\quad + \gamma\frac{\sum_{i=1}^{I} o_{iu}}{\sum_{u=1}^{U}\sum_{i=1}^{I} o_{iu}} \\ P(b_v|c_k) &= (1-\gamma)\frac{\sum_{i=1}^{I} P(c_k|d_i)\delta_{iv}}{\sum_{v=1}^{V}\sum_{i=1}^{I} P(c_k|d_i)\delta_{iv}} \\ &\quad + \gamma\frac{\sum_{i=1}^{I} \delta_{iv}}{\sum_{v=1}^{V}\sum_{i=1}^{I} \delta_{iv}} \\ P(w_j|c_k) &= (1-\gamma)\frac{\sum_{i=1}^{I} P(c_k|d_i)c_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I} P(c_k|d_i)c_{ij}} \\ &\quad + \gamma\frac{\sum_{i=1}^{I} c_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I} c_{ij}} \end{aligned} \tag{9}$$

表 1　実験で使われた省略名
Table 1　Abbreviated names used in our experiment.

省略名	フルネーム数	データ数	省略名	フルネーム数	データ数
s.lee	161	1067	j.park	68	397
j.lee	134	934	y.liu	67	313
j.kim	129	822	c.wang	65	366
j.wang	112	583	s.chen	64	309
s.kim	108	625	h.li	63	220
h.kim	100	550	j.liu	63	409
y.wang	100	557	z.wang	63	142
h.lee	99	349	j.li	61	314
j.chen	86	493	j.zhang	60	308
x.wang	86	322	s.li	59	245
s.wang	84	274	j.wu	56	329
y.zhang	83	412	z.li	56	210
k.lee	81	386	j.lin	55	196
y.chen	81	531	h.liu	54	197
h.wang	79	389	s.liu	54	122
y.li	76	261	z.zhang	54	255
c.lee	75	480	d.kim	53	334
h.chen	74	419	c.chen	51	483
y.kim	74	406	x.liu	51	187
x.zhang	72	287	y.yang	51	250
k.kim	71	333	j.yang	50	310
y.lee	71	385	l.wang	50	253
s.park	69	397	m.lee	50	315
x.li	69	321			

と式で表されるスムージングを適用する．2 変数混合モデルの式 (8) にも同様のスムージングを使う．引用データにおいてはデータのスパースネス (sparseness) が甚だしく，スムージングが有効である．γ の値は，予備実験の結果，0.5 と設定した．

3.4.2　アニーリング

本論文では，ナイーヴ・ベイズ混合モデルでの式 (3)，2 変数混合モデルでの式 (7) を計算する際，EM アルゴリズムにおいて局所最適解につかまりにくくする工夫として，Rose らによるアニーリング法 [8] を用いる．式 (3)，式 (7) の代わりに各々

$$P(c_k|d_i) = \frac{\{\bar{P}(d_i, c_k)\}^\beta}{\sum_{k=1}^{K}\{\bar{P}(d_i, c_k)\}^\beta} \tag{10}$$

$$P(y_s, z_t|d_i) = \frac{\{\bar{P}(d_i, y_s, z_t)\}^\beta}{\sum_{s=1}^{S}\{\sum_{t=1}^{T}\bar{P}(d_i, y_s, z_t)\}^\beta} \tag{11}$$

を用い，$\beta = 0.5$ を初期値とし，EM アルゴリズムの反復計算が 1 ステップ進むたびに β を 0.8 乗するという方法で，アニーリング法を実装する．このように実装すると，反復計算の初期では，式 (10) では，隠れ変数の値の各々の確率 $P(c_k|d_i)$ の違いが，また，式 (11) では，隠れ変数の値の組み合わせの各々の確率 $P(y_s, z_t|d_i)$ の違いが，あまり目立たない．反復が進むにつれ β が 1 に近づき，確率の違いが目立つようになってくる．

4. 実　　験

4.1 実験方法

本論文では，実験の再現性を確保するため，DBLP 書誌情報データベース [1] が一般に公開しているデータ・セットを評価実験に用いた．しかも，頻繁に内容が変わる最新のデータ・セットではなく，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．まず，共著者名，タイトル，雑誌名ないし国際会議名という 3 つのフィールドを備えていないデータは除去し，著者名の姓名の名が元々イニシャルにされているデータも除去した．残ったデータですべての著者名の姓名の名をイニシャルにした．こうして省略された著者名のうち，表 1 に示した，対応するフルネームが 50 以上あるものを実験に使った．1，4 列目が省略著者名，2，5 列目が対応するフルネームの個数，3，6 列目が各省略名を含む引用データの個数である．省略名はすべて小文字にし，名のイニシャルを姓とピリオドでつないで表記することにする．なお，タイトルからは stop word を除去し，porter stemmer [2] を適用した．

表 1 の省略名の各々について，次のような実験をおこなった．例えば，省略名 “s.lee” について実験する場合を考える．

（1） “s.lee” という省略名を含む引用データをすべて集め，実験用の引用データ集合 D をつくる．

（2） D を，以下の 3 通りの方法でクラスタリングする．なお，モデルのパラメータの初期値はランダムに設定し，3 通りのどのクラスタリングを行う際にも，10 通りの異なる初期値から EM アルゴリズムを開始する．つまり，同じ引用データについて，各クラスタリング方法で，10 通りずつの結果を得る．

（a） ナイーヴ・ベイズ混合モデルを用いる．このクラスタリング方法を NBM と書くことにする．

（b） すべての引用データについて，共著者名フィールドだけを残し，これにナイーヴ・ベイズ混合モデルを用いる．このクラスタリング方法を NBMa と書くことにする．

（c） 2 変数混合モデルを用いる．このクラスタリング方法を TVM と書くことにする．

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 $S = T = \lceil\sqrt{\text{真のクラスタ数}}\ \rceil$，NBM，NBMa では $K = (\lceil\sqrt{\text{真のクラスタ数}}\ \rceil)^2$ と設定する．真のクラスタ数が未知の場合については $K = 200$，$S = 10, T = 20$ などの値も用いたが，本質的に新しい結果は得られなかったため割愛する．なお計算時間は，クラスタ数未知の場合の省略名 “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．

（1） 例えば “s.lee” という省略名を含む引用データの集合 D について，4.1 節に示した NBM,NBMa,TVM の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，そのクラスタのラベルと呼び，$Label(G)$ と書く．

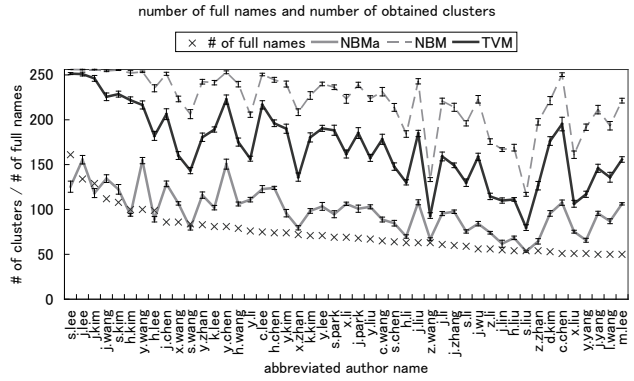

図 2　各省略名に対応するフルネーム数と得られたクラスタ数
Fig. 2　Number of full names and number of obtained clusters.

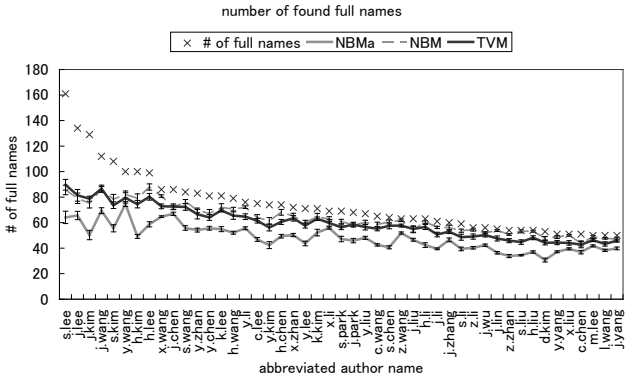

図 3　各省略名に対応するフルネーム数と見つかったフルネーム数
Fig. 3　Number of full names and number of found full names.

（2） 各クラスタ $G \in \mathcal{G}$ に含まれる引用データを元のデータで確認したとき，$Label(G)$ が現われるデータの数を $N_{pos}(G)$ とする．G のサイズを $N_{size}(G)$ とする．また，D に含まれる全引用データを元のデータで確認したとき，$Label(G)$ が現われるデータの総数を $N_{cor}(G)$ とする．このとき G の precision は $N_{pos}(G)/N_{size}(G)$，recall は $N_{pos}(G)/N_{cor}(G)$ と定義される．

（3） クラスタリング結果全体での precision と recall を，macroaveraged precision/recall および microaveraged precision/recall として算出する．macroaveraged precision/recall は

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

と，microaveraged precision/recall は

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

と，それぞれ定義される．以上 4 種類の評価値を，NBM,NBMa,TVM という 3 手法それぞれで，10 通りのランダムな初期値から出発した 10 通りの結果すべてについて計算する．そして，これら 10 通りの評価値の平均と標準偏差を求め，3 種類のクラスタリング手法それぞれの，特定の省略著者名に関する曖昧性解消の性能評価とする．以下，macroaveraged precision/recall をそれぞれ P_{mac}，R_{mac} と，microaveraged precision/recall をそれぞれ P_{mic}，R_{mic} と略記する．

4.3　実験の結果

真のクラスタ数が未知とした場合に得られた空でないクラスタの数を，各省略名について図 2 に示した．値は，10 通りの初期値から始めた EM アルゴリズムによって得られた，10 通りのクラスタリング結果での平均である．標準偏差は，マーカのプラスとマイナスの方向の幅によって示した．× 印は，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，空でないクラスタ数が真のクラスタ数に迫っている．全体としては，NBM，TVM，NBMa の順でクラスタの数が多く，特に NBM でクラスタの過細分化 (oversegmentation) が甚だしい．これは，NBMa では共著者名しか用いておらず，引用データの示す多様性が減ぜられた一方，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であろう．TVM では，タイトルの生成が 1 つの隠れ変数にしか依存しないため，中間的なふるまいを示したと考えられる．図 3 は，各省略名について，少なくとも 1 つのクラスタのラベルとなったフルネームの数，つまりクラスタリングによって見つかったフルネーム数を示している．標準偏差は，やはりマーカで示した．NBM と TVM は，対応するフルネーム数が少ない省略名で，ほとんどのフルネームを見つけることができたが，NBMa では，見つけそこねたフルネームが全体的に多い．

次に，各省略名について，10 通りの初期値から始めて得られた 10 通り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，図 4 から図 7 にまとめた．標準偏差は，プラス方向とマイナス方向にマーカの幅で示した．図 4 は P_{mic} を示しており，この値は，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．図 5 は R_{mic} を示しており，この値は，クラスタが細分化されるほど低くなりやすく，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．図 6 は P_{mac} を示しており，P_{mic} と同様，クラスタが細分化されるほど高くなりやすいが，すべてのクラスタの precision が平等に寄与している．図 7 は R_{mac} を示しており，R_{mic} と同様，クラスタが細分化されるほど低くなりやすいが，すべてのクラスタの recall が平等に寄与している．図 4 から図 7 を見ると，precision では，NBM よりも NBMa のほうが良い結果を示しているが，recall では，NBMa よりも NBM のほうが良い結果を示している．これは，NBM が NBMa よりも，大きな多様性を示すデータを使っており，過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 の中間のふるまいを示している．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．その理由はやはり，最も大きな多様性を示すタイトル情報が，2 変数混合モデルでは 1 つの隠れ変数にのみ依存して生成されているからであろう．

recall の値が全体的に低いのは，NBM，NBMa，TVM に共通の問題点である．しかし，いくつかの省略名，例えば “z.wang”

表 2　真のクラスタ数が未知の場合の評価結果
Table 2　Evaluation results under the assumption that the correct number of clusters is unknown.

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

や “s.liu” については，NBM，NBMa，TVM すべてで比較的 recall が良い．これらの省略名については，図 2 を見ると，パラメータ推定の結果として，適度な数の空でないクラスタが自然に残ったことが分かる．つまり，これらの省略名については，ナイーヴ・ベイズ混合モデルや，2 変数混合モデルによってとらえられた引用データ間の類似性が，正確なクラスタリングに近い引用データのグルーピングを表現していたのであろう．

表 2 は，以上の 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．第 6，7 列は，それぞれ，P_{mic} と R_{mic} の調和平均，P_{mac} と R_{mac} の調和平均である．この結果から，多くのフルネームを見つけ損ねてよいなら NBMa，できるだけ多くのフルネームを見つけたいなら TVM を選ぶとよいことが分かる．また，タイトルや雑誌名・会議名が，共著者名に比べて曖昧性解消に寄与しないことも分かる．

真のクラスタ数が既知と仮定した場合の評価結果を，表 3 にまとめた．紙面の都合上，省略著者名ごとのデータは割愛する．真のクラスタ数が既知とした場合は，TVM では $S = T = \lceil\sqrt{\text{真のクラスタ数}}\ \rceil$，NBM，NBMa では $K = (\lceil\sqrt{\text{真のクラスタ数}}\ \rceil)^2$ と設定した．表 2 と表 3 を比べると，precision が大きく下がり，recall が大きく上がっている．これは，真のクラスタ数に合わせてクラスタ数を設定することで，過細分化が回避されたためであろう．TVM が，NBM と NBMa の中間的なふるまいを示している点では，表 2 と同様である．ただし，表 8 に示したように，クラスタリングによって見つかったフルネーム数は，各省略名に対応するフルネーム数よりもかなり少なくなってしまった．この点では，真のクラスタ数に合わせてクラスタ数を設定したことが，不利にはたらいている．

最後に，参考として，Han らの研究 [5] で使われた 14 の省略名について，本論文の NBM，NBMa，TVM によって改めて実験をおこない，得られた P_{mic} の値を，表 4 にまとめた．真のクラスタ数は既知と仮定している．[5] に示されたデータも併記したが，Han らがどのような DBLP データを使ったか分からないため，正確な比較ではない．また，4 種類の評価値のうち P_{mic} の値しか示さなかったのは，Han らがこの値でしか評価していないためである．表 4 の最下行は 14 の省略名での平均値だが，本論文の手法の値がより大きいからといって，本論文の手法が良いとは言えない．なぜなら，すでに見たように，クラスタリングにおいて過細分化さえ起こっていれば，P_{mic} は自ずと大きくなるからである．しかし，少なくとも，共著者名，タイトル，雑誌名・会議名の 3 つのフィールドだけを手がかりに名前の曖昧性問題を解くことが，どの程度難しい問題かは，この表から見てとれるだろう．今後，[5] の提案手法を実装し，正確な比較をおこなう予定である．

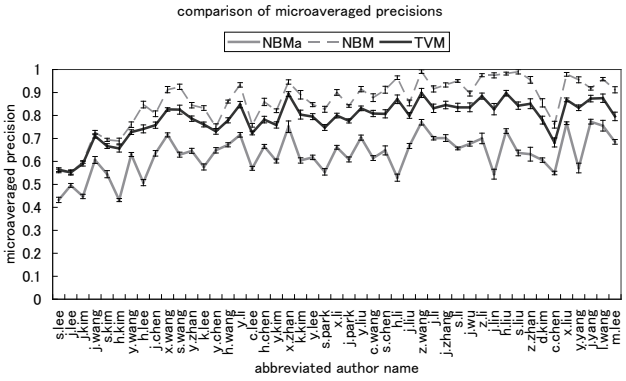

図 4　microaveraged precision の比較
Fig. 4　Comparision of microaveraged precisions.

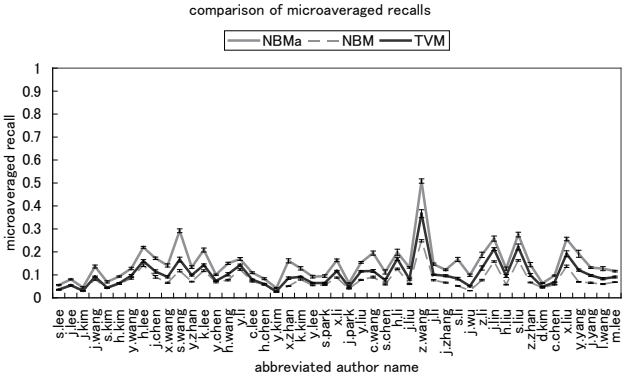

図 5　microaveraged recall の比較
Fig. 5　Comparision of microaveraged recalls.

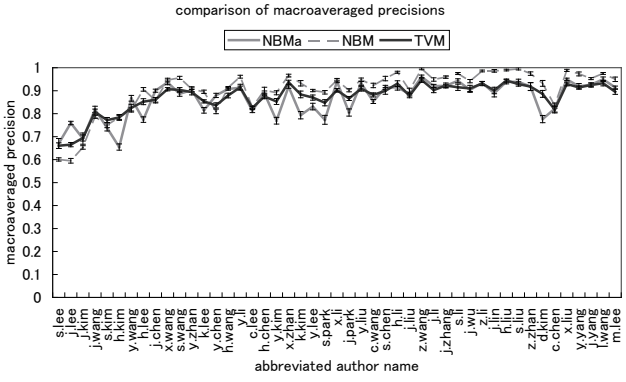

図 6　macroaveraged precision の比較
Fig. 6　Comparision of macroaveraged precisions.

5. 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．まず，特定の省略著者名を含む引用データをすべて集め，既存のナイーヴ・ベイズ混合モデル，および，本論文で提案した 2 変数混合モデル (TVM) によって，この引用データ集合をクラスタ

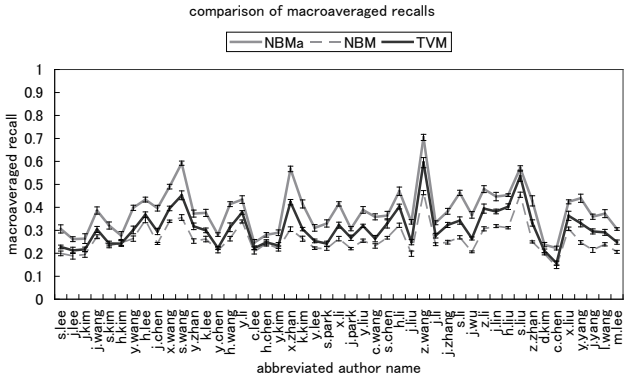

図 7　macroaveraged recall の比較
Fig. 7　Comparision of macroaveraged recalls.

表 3　真のクラスタ数が既知の場合の評価結果
Table 3　Evaluation results under the assumption that the correct number of clusters is known.

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

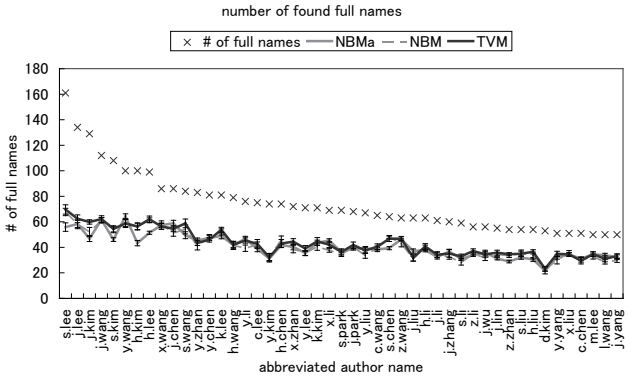

図 8　真のクラスタ数が既知としたときの，各省略名に対応するフルネーム数と見つかったフルネーム数
Fig. 8　Number of full names and number of found full names under the assumption that the correct number of clusters is known.

リングした．なお，ナイーヴ・ベイズ混合モデルを適用する際には，元の引用データをそのまま用いる場合 (NBM) と，共著者名だけを用いる場合 (NBMa) とを実験した．実験の結果，真のクラスタ数が未知と仮定した場合には，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 を使うと，空でないクラスタの数が真のクラスタ数に近づいてはいたが，同時に見つけ損ね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た結果となった．つまり，できるだけ多くのフルネームを見つけつつ，クラスタリングの性能を上げたい場合は，TVM を使うとよい．真のクラスタ数が既知と仮定した場合，microaveraged precision で比較する限りは，Han らの先行研究 [5] と同等の結果が得られた．Han らの先行研究が recall による評価をおこなっていないため，これは正確な比較ではないが，共著者名，タイトル，雑誌名・会議名という 3 つのフィールドだけを使った著者名の曖昧性解消がどの程度難しいか，共通の認識を持つことができた．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などする必要があると思われる．

表 4　Han ら [5] の評価結果との比較
Table 4　Comparison of our evaluation results with those reported in [5].

省略名	[5]	NBMa	NBM	TVM
s.lee	0.464	0.426	0.502	0.495
j.lee	0.495	0.462	0.464	0.453
y.chen	0.490	0.521	0.456	0.462
c.chen	0.382	0.446	0.422	0.418
a.gupta	0.476	0.661	0.586	0.606
k.tanaka	0.595	0.774	0.714	0.742
j.smith	0.617	0.556	0.569	0.560
j.martin	0.656	0.608	0.597	0.571
a.kumar	0.463	0.590	0.618	0.604
m.brown	0.660	0.606	0.686	0.669
m.miller	0.598	0.722	0.624	0.642
j.robinson	0.571	0.818	0.789	0.761
d.johnson	0.445	0.658	0.686	0.671
m.jones	0.657	0.639	0.557	0.569
Avg	0.541	0.606	0.591	0.587

文　献

[1] http://www.informatik.uni-trier.de/~ley/db/
[2] http://www.tartarus.org/~martin/PorterStemmer/
[3] X. Dong, A. Halevy, and J. Madhavan, Reference Reconciliation in Complex Information Spaces, in *Proc. of SIGMOD2005*, pp. 85-96, 2005.
[4] H. Han, C. L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, Two Supervised Learning Approaches for Name Disambiguation in Author Citations, in *Proc. of JCDL2004*, pp. 296-305, 2004.
[5] H. Han, W. Xu, H. Zha, and C. Lee Files, A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations, in *Proc. of SAC'05*, pp. 1065-1069, 2005.
[6] H. Han, H. Zha, and L. Giles, Name disambiguation in author citations using a k-way spectral clusering method, in *Proc. of JCDL2005*, pp. 334-343, 2005.
[7] D. V. Kalashnikov, S. Mehrotra, and Z. Chen, Exploiting Relationships for Domain-Independent Data Cleaning, in *Proc. of the SIAM International Conference on Data Mining*, 2005.
[8] K. Rose, E. Gurewitz, and G. Fox, A Deterministic Annealing Approach to Clustering, *Pattern Recognition Letters*, Vol. 11, pp. 589-594, 1990.
[9] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell, 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.