DBSJ Journal, Vol.9, No.2
November 2010

二段階クラスタリングを単語重み付与に応用した人名曖昧性解消

Person Name Disambiguation by Term Weighting via Two-Stage Clustering

吉田 稔 ♥　池田 雅紀 ♦
小野 真吾 ♠　佐藤 一誠 ♣
中川 裕志 †

Minoru Yoshida　Masaki Ikeda
Shingo Ono　Issei Sato
Hiroshi Nakagawa

本稿では、Web 検索結果における人名曖昧性解消問題を対象とした新たなクラスタリング手法を提案する。提案手法では、固有名詞等の「強い素性」を利用してクラスタリングを行ったあと、そのクラスタリング結果を利用して単語等の「弱い素性」に重み付けを行ない、クラスタの再構成に役立てるという二段階クラスタリングを行う。公式データセットを使った実験の結果、提案手法は、比較対象の手法と比べ、同等か高い性能を示すことを確認した。

In this paper, we report our system that disambiguates person names in Web search results. The system uses named entities, compound key words, and URLs as features for document similarity calculation, which typically show high precision but low recall clustering results. We propose to use a *two-stage clustering algorithm* to improve the low recall values, in which clustering results of the first stage are used to extract features used in the second stage clustering. Experimental results revealed that our algorithm shows the score better than (in B-Cubed measures) or competitive to (in P-IP measures) the best systems at the WePS-1 and WePS-2 workshops.

1. はじめに

Web 検索において、**人名曖昧性解消**とは、人名による検索の際に、同名であるが違う実体を指すページを自動的にまとめあげる（クラスタリングする）タスクであり、近年、人名曖昧性解消に関する国際ワークショップ WePS(Web People Search Workshop)[2, 3] が開催されるなど活発に研究が行われている。この問題には、特に人名クエリと共起する固有名詞に着目した手法が効果を発揮することが明らかになっている [9]。固有名詞の利用によって、高い精度で同一人物を判定できる半面、こうした固有名詞は、必ずしも文書に出現するとは限らず、網羅性という面で限界がある。

網羅性を上げるためには、文書類似度に対する（まとめるか否かを判定する）閾値を下げる、という方法があるが、この場合、通常は、精度が犠牲となる。また、その他の方策として、固有名詞以外の一般の単語（名詞）を用いる方法が挙げられるが、一般の単語は、人物との関連性が高い単語ばかりではないため、これらを手掛かりとすることによって、関連のない人物どうしを結びつけてしまうミスが発生しやすくなる。

これに対し本研究では、一般の単語のうち、**有用な単語**に高い重みを与え、手掛かりとして用いる手法を提案する。そのために、まず、固有名詞等の有用な特徴量を用いた、**第一段階クラスタ**（クラスタ＝文書のまとまり）を生成し、これら第一段階クラスタの結果を利用し、クラスタの重みを単語に伝播させることで、一般の単語に重みを付与する。この重みづけされた単語を用い**第二段階クラスタ**を生成する。このため、提案手法を、**二段階クラスタリング法**と呼ぶ。これにより、例えば、野球選手とコンピュータ科学者が同姓同名だった場合に、「ホームラン」や「ボール」、あるいは「メモリ」や「アルゴリズム」といった単語を重みの高い単語として人物の判定に利用することが可能になる。

この二段階クラスタリングは、「入力として初期集合のみが与えられたとき、その集合に特徴的な手がかり表現を獲得し、獲得された手がかりを用いてもとの集合を拡張する」手順として捉えることができる。このような手順を実現する手法として、自然言語処理の分野で**ブートストラップ法**と呼ばれる手法が提案されている。典型的には、ブートストラップ法は、入力として、**シード**と呼ばれる、元となる**インスタンス**の集合（例えば、「映画の題名」の集合）をとる。アルゴリズムは、インスタンスを抽出するための**パターン**（この場合、映画の題名と共起しやすい言語パターン）の集合を、「多くのインスタンスと共起するパターンは良いパターンである」という考えに基づき、共起するインスタンスのスコアをもとに獲得する。それに続き、今度は逆に新たなインスタンスの集合を、「多くのパターンと共起するインスタンスは良いインスタンスである」という考え方に基づき、（上記で獲得された）パターンのスコアをもとに、獲得する。これは、インスタンスからパターンへのスコア伝播と、それに続くパターンからインスタンスへのスコア逆伝播の手順として捉えることができ、提案手法では、このスコア伝播手法を人物クラスタリングに応用する。すなわち、この場合、インスタンスは**文書**に、パターンは**単語素性**に、シードは**第一段階で得られたクラスタ**にそれぞれ対応する。アルゴリズムを通じて、有用な単語（素性）に高い重み（信頼度）が与えられ、そうでない単語の重みは低くなる。その後、得られた単語の重みから、新たに文書の重みが求まる。

♥ 正会員　東京大学　mino@r.dl.itc.u-tokyo.ac.jp
♦　東京大学　ikeda@r.dl.itc.u-tokyo.ac.jp
♠　東京大学　ono@r.dl.itc.u-tokyo.ac.jp
♣ 学生会員　東京大学　sato@r.dl.itc.u-tokyo.ac.jp
†　東京大学　nakagawa@dl.itc.u-tokyo.ac.jp

この信頼度計算をクラスタ毎に行う（各クラスタに対する単語・文書の重みを求める）ことにより、文書の各クラスタへの帰属度（重み）が求まり、集合が拡張され、新たなクラスタリング結果を得ることができる。ただし、一般的にブートストラップ法においては、上記手順を複数回行いインスタンス集合を順次拡張していくが、本提案手法では、シード集合（初期集合）がインスタンス抽出の問題設定に比べて大きいこともあり、後述の通り、実際にはこの反復を一回とする。このため、提案手法は、ブートストラップ法の反復一回分、すなわち、「文書クラスタから素性クラスタへの重みの伝播」および、「素性クラスタから文書クラスタへの重みの逆伝播」を一回ずつ行う、重み伝播アルゴリズムと言える。我々は、ブートストラップ法の一つとして近年提案された *Espresso* アルゴリズム [19] を用い、これを人名曖昧性解消に用いる手法を提案する。

2. 関連研究

Bagga ら [4] は、単語出現頻度に基づくベクトル空間モデルによる手法を提案した。Niu ら [17] は、単語に加えて、情報抽出ツールにより抽出された情報も素性として用いた。また、Mann[15] らの手法では、人物に関するプロフィール情報を抽出し利用している。これらの研究では、人工的な小規模データセットにのみ実験を行っており、実際の Web ページに適用可能かどうかは不明である。これに対し、Wan ら [22] の提案した WebHawk は、サーチエンジンの検索結果を対象としているが、そのアルゴリズムは、英語の人名に特化しており、特にミドルネームを手掛かりとして用いているため、我々の問題設定とは若干の違いがある。

少し異なるアプローチとしては、Bekkerman ら [6] による、Web のリンク構造を利用したアルゴリズムがあるが、既知のソーシャルネットワーク上の人物を対象としており、これも我々の問題設定（「サーチエンジンを経由した一般の人名曖昧性解消」）とは異なっている。

Bollegala ら [7] の手法では、文書から抽出されたキーワードを文書類似度の測定に用いている。彼らはクラスタからのキーワード再抽出も提案しているが、再抽出されたキーワードを文書クラスタリングに利用することは行っていない。

Bunescu ら [8] は、Wikipedia から教師付き学習のための素性を抽出し、曖昧性解消に用いることを提案している。Wikipedia のような外部リソースを用いることは興味深い方向性であるが、特に有名でない人物までターゲットとするわれわれの問題設定には、外部リソースの利用は向いていないと考えられる。

人名曖昧性解消に関して、近年、WePS[2][3] と呼ばれる評価型ワークショップが開催された。参加チームの手法は、概ね上記の手法のいずれかに当てはまるものであり、形態素解析および固有名詞抽出の結果を用いて文書間類似度を計算し、一般的なクラスタリング手法を用いて文書をまとめるというものであった。

また、参加チーム以外で WePS データセットを用いた研究もいくつか報告されている。Kala ら [11] は、人名や組織名の共起関係を検索エンジンのヒット数をもとに計算し、文書類似度を測定する手法を提案し、極めて高い精度を達成しているが、検索エンジンへの大量のアクセス（毎回約四万回）を必要とするため、リアルタイムに検索結果をクラスタリングする用途には不向きである（論文では「検索エンジンサーバ上での利用に使うべき」と述べられている。）Balog ら [5] は、1 パスクラスタリング手法とベクトル空間モデルという単純な手法を用い、WePS のトップシステムと同等の性能を実現している。彼らのアイデアは、検索結果中の各文書の順位まで利用することと、HTML の文書構造を用いて有用な部分ページを抜き出すことである。我々の手法は、一般的な文書クラスタリングの設定のものであるため、彼らの手法を併用することで、より高い精度を得られると考えられ、今後の課題として考えている。

また、二段階クラスタリング手法として過去に提案されたものとしては、以下のものが挙げられる。Slonim ら [20] は、情報ボトルネック法に基づく、ダブルクラスタリングを提案している。具体的には、「文書クラスタの情報を最大限保持する」ように単語をクラスタリングし、得られた単語クラスタを新たに素性として用いることで、再び文書をクラスタリングするというものである。この手法は文書クラスタリング一般を扱っており、人名曖昧性解消には（「はじめに」で述べた通り、）単純な単語頻度や TF-IDF を用いることが難しいため、そのまま本手法を用いることは難しい。二段階クラスタリング手法は、我々のものも含めて WePS-2 において二つ提案されているが、[21][10], どちらも一般的な単語素性は用いていない。

一方、Liu ら [14] は、第一段階クラスタから、各クラスタに特徴的な素性（単語、単語ペアおよび固有名詞）を抽出し、抽出された素性を用いて（各文書で、どのクラスタの素性が多いかの頻度比較を行って）第二段階においてクラスタを再構築するという、我々の手法と近いクラスタリング手法を提案している。我々の手法は、この Liu らの二段階クラスタリングの考え方を、ブートストラップ法の枠組みで実現したものと捉えることもできる。また、この手法では、全ての素性を同等に扱っているが、WePS の多くのシステムにおいては、固有名詞と単語素性をそのまま同時に用いても、固有名詞だけ用いた場合と比べてそれほど精度が向上しないと報告されている。これに対し、提案手法のモデルは、第一段階と第二段階の素性を別々にすることで、単語素性とその他の素性の間の階層を導入している。日本語を対象とした人名曖昧性解消では、片岡ら [12] の二段階クラスタリング手法があるが、単語等の「弱い素性」については扱っていない。

3. 第一段階クラスタリング

本節と次節において、提案する**二段階クラスタリング法**を説明する。本稿では、文書から、その文書の特徴量として抽出される文字列を、**素性**と呼ぶ。素性としては、固有名詞、複合語、URL、そして単語が用いられる。提案手法では、素性を**強い素性**（固有名詞、複合語、URL）と**弱い素性**（単語）に区別する。前者は後者に比べ強い分類能力がある（すなわち、人物を特定す

る能力が強い）と仮定する。提案手法では、まず強い素性のみを用いてクラスタリングを行い（「第一段階クラスタリング」）、その結果を利用して弱い素性を重みづけし、クラスタに属さなかった文書のうち、各クラスタに併合できる文書を見つけ、新たに各クラスタに加える。（「第二段階クラスタリング」）。

本節では第一段階クラスタリングについて詳説し、次節において第二段階クラスタリングについて述べる。

文書の前処理として、 HTML からテキストへの変換ツール lxml [1] および文の切り分けツール[2]を用い、各 HTML 文書を、テキストファイル（文のリスト）に変換する。その後、クエリ周辺の文字列を取得[3] し、各文字列を Tree Tagger [4] による形態素解析および、Stanford NER[5] による固有名詞抽出にかける。その結果を利用し、最終的に人名、地名、組織名、URL、単語を各文書の素性として得る。

前処理された文書に対し、強い素性を利用して、精度の高い（間違いの少ない）クラスタを生成する。この際、先行研究で用いられていた「固有名詞」のほかに、「複合名詞」と「URL」も利用する。複合名詞は、ある専門分野の特徴的な概念を表していることが多い。また、URL は、その人物に関連する Web ページを表現しているということで、人物の特定能力が高いと考えられる。

固有名詞としては、人名のほか、組織名と地名も用いる。ただし、これらのうち、一般的な（一般の文書に高い頻度で現れる）ものは使用しない。これは、例えば「中央区」のような、どこにでも現れる地名を使用することによるミスを防ぐためである。また、複合名詞の利用の際には、専門用語抽出アルゴリズム「言選 Web」[16] を利用し、複合名詞のスコア付けを行っている。このスコアが閾値 θ_{CKW} を超えた複合名詞のみを使用する。

URL としては、HTML の<a>タグで囲まれた部分のほか、対象の Web 文書の URL そのものも使用する。また、組織名・地名と同様、あまりにも Web 上での頻度が高い URL は使用しない。

これらの特徴量を利用し、特徴量毎に、文書と文書の類似度を、オーバーラップ係数 $Overlap(d_x, d_y) = \frac{|\mathbf{f}_x \cap \mathbf{f}_y|}{\max(\min(|\mathbf{f}_x|,|\mathbf{f}_y|),\theta_{overlap})}$ で定義する。[6]ただし、$\mathbf{f}_x$ と $\mathbf{f}_y$ は、それぞれ文書 d_x と d_y における特徴（固有名詞など）の集合である。基本的にはオーバーラップ係数の値をそのまま類似度（$sim(d_x, d_y)$）として採用するが、d_x と d_y の間に直接リンクがあるときの「URL による類似度」は 1 とする。また、$\theta_{overlap}$ は、分母が小さくなりすぎることを防ぐためのパラメータであり、現在は $\theta_{overlap} = 4$ としている。

以上で定義される複数の類似度から、最終的な文書同士の類似度を定義するが、まず、固有名詞に関しては、「人名による類似度」「地名による類似度」「組織名による類似度」を、それぞれ係数 α_P, α_L, α_O で重み付けした線形補間により組み合わせることにより固有名詞全体での類似度を得る。（ただし $\alpha_P + \alpha_L + \alpha_O = 1$。また、これらの値は、WePS-1 訓練データ[7]による調整を経て、$\alpha_P = 0.78$, $\alpha_O = 0.16$, and $\alpha_L = 0.06$ と設定されている。）さらに、こうして計算された「固有名詞による類似度」「複合名詞による類似度」「URL による類似度」に関して、最大値をとることで文書の類似度と定義する[8]。

これらの文書間類似度をもとに、文書のクラスタリング（纏め上げ）を行う。クラスタリングには、一般的な手法である「群間平均法による階層クラスタリング」を用いた。

4. 第二段階クラスタリング

第二段階クラスタリングでは、第一段階で得られたクラスタを改良することを目指し、クラスタの拡張を行う。そのための手法として、近年自然言語処理（情報抽出）の研究で提案されている Espresso アルゴリズム [19] を採用する。Espresso は、最初にある名詞集合（あるいは名詞ペアの集合）が与えられると、それが出現しやすい言語的パターンを発見し、そのパターンをもとに類似する名詞（名詞ペア）を新しく獲得するという、**ブートストラップ法**と呼ばれる手法の一種である。Komachi ら [13] は、Espresso アルゴリズムの理論的解析を行い、アルゴリズムを行列計算で表現した。本稿でもその表現に倣い、アルゴリズムを行列により $\mathbf{i}^{(t+1)} = \frac{1}{|I||T|} \cdot \mathrm{M}^T\mathrm{M}\mathbf{i}^{(t)}$ と表現する[9]。ここで、M はインスタンスとパターンの関係の強さを表す行列、$\mathbf{i}$ は各インスタンスの重みを表すベクトルである。アルゴリズムは、与えられた名詞集合（クラスタ）を表すベクトル $\mathbf{i}^{(0)}$ （各インスタンスが、クラスタに含まれていれば 1、そうでなければ 0）を用意し、これに M, M^T、さらに正規化のための値 $\frac{1}{|I||T|}$ を乗算することにより $\mathbf{i}$ を更新する。

我々は、この Espresso における名詞を**文書**、パターンを**素性**（単語）と考え、人名曖昧性解消の問題に適用する。初期ベクトルは、（第一段階クラスタリングの結果得られる）クラスタ毎に定義される。すなわち、クラスタ数と同じ数の初期ベクトルが用意され、それぞれに上記の更新を行っていく。複数ベクトルの更新を同時に行うため、ベクトル $\mathbf{i}$ のかわりに行列 R を用いる。R の各列が各クラスタに対応するベクトルとなる。

具体的には、$\mathrm{R}_D^{(t)} = \{r_{d,C}^{(t)}\}$（ここで、$r_{d,C}$ は、文書 d とクラ

[1] http://codespeak.net/lxml/
[2] http://www.answerbus.com/sentence/
[3] 前後 50 語, 100 語, 200 語, 文書すべての 4 種類を試し、予備実験で成績の良かったパラメータ（複合語は 100 語以内、固有名詞は文書すべて）を利用。
[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[5] http://nlp.stanford.edu/software/CRF-NER.shtml
[6] オーバーラップ係数は、Web 文書間にサイズのばらつきがあることを考慮し、「大きい文書が小さい文書の内容を包含している」という場合にも高いスコアが与えられるよう採用した。
[7] テストデータとは別に提供されており、49 の人名を含む。
[8] 最大値をとるのは、強い素性が単体でも人物を特定できる能力があり、半面、文書中に存在しない場合がある（その場合、類似度が 0 になってしまう）ため、各類似度の OR を取るような計算が必要であることによる。
[9] これは、彼らが *simplified Espresso* と呼ぶ、パターンやインスタンスのフィルタリングを省略したバージョンであり、本稿で用いるブートストラップ法はこの simplified Espresso である。

スタ C の関連の強さ。t は後述の繰り返しの際の添字。）を定義する。これに加え、i 番目の素性と j 番目の文書の関係の強さを表す値を (i,j) 要素に持つ行列 P、および素性-クラスタ行列 $\mathrm{R}_F^{(t)} = \{r_{f,C}^{(t)}\}$ を定義する。P は、Espresso の定義式と同様に、自己相互情報量に従って以下のように定義されている。

$$\mathrm{P}[f,d] = \begin{cases} \frac{1}{\max pmi} \log \frac{p(f,d)}{p(f)p(d)} & \text{if } \frac{p(f,d)}{p(f)p(d)} > 1 \\ 0 & \text{otherwise} \end{cases}, (f \in F, d \in D)$$

（ただし、$\max pmi = \max(\mathrm{P}[f',d'])$ $(f' \in F, d' \in D)$）アルゴリズムは、$\mathrm{R}_D^{(t)}$ と $\mathrm{R}_F^{(t)}$ を、P および P^T を掛けることにより交互に更新していく。

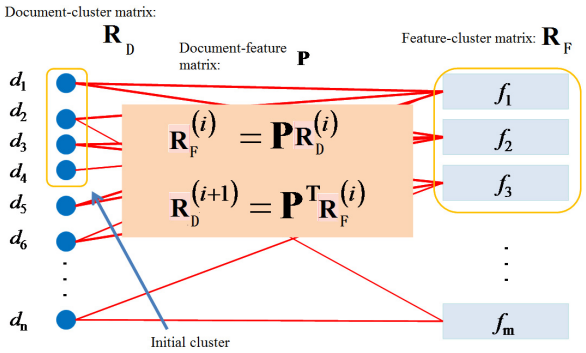

図 1　提案手法の概念図

図 1 に、この行列積計算の意味を説明する。ここで、文書 d_j が素性 f_i を持てば、f_i と d_j がエッジで結ばれている。（行列では、素性-文書行列 P の要素 $p_{i,j}$ がゼロでない状態を示す。）図中では、文書 d_1, d_2, d_3, d_4 が同一のクラスタに入っており、このことは、行列表現では、$r_{1,k}^{(0)} = r_{2,k}^{(0)} = r_{3,k}^{(0)} = r_{4,k}^{(0)} = 1$ （クラスタ番号を k とする）として表現される。このようなクラスタ情報は、文書-素性行列 P を通して、素性-クラスタ行列 $\mathrm{R}_F^{(0)}$ に伝播する。すなわち、文書-クラスタ行列 $\mathrm{R}_D^{(0)}$ に行列 P を掛けることで、素性-クラスタ行列 $\mathrm{R}_F^{(0)}$ が得られる。この素性-クラスタ行列では、k 番目のクラスタに強く関連する素性は、k 列目に高い値（重み）を持つことになる。その後、今度は素性-クラスタ行列 $\mathrm{R}_F^{(0)}$ に文書-素性行列（の転置）P^T 掛けて、新たな文書-クラスタ行列 $\mathrm{R}_D^{(1)}$ を得る。これは、素性-クラスタの関連度情報を逆に各文書に伝播させることに相当する。

アルゴリズムの流れは、Algorithm 1 に示す通りになる。入力として与えられるのは、第一段階クラスタリングによって得られたクラスタを初期クラスタ集合 $\mathcal{C}^{(0)}$ である。これに基づき文書-クラスタ関連度行列の初期値 $\mathrm{R}_D^{(0)}$ を定義し（$d \in C$ なら $r_{d,C}^{(0)} = 1$、それ以外なら 0）、どのクラスタにも属していない文書（より正確には、サイズ 1 のクラスタを形成する文書）の所属をこのアルゴリズムで決定する。Step 2 で文書から素性へ、Step 3 で素性から文書への重み伝播計算を行い、最終的に得られた文書-クラスタ行列から、Step 4 において、各文書 d を、関連度 $r_{d,C'}$ を最大化させるクラスタ $C' \in \mathcal{C}^{(0)}$（のうち、$|C'| \geq 2$ または、$C' = \{d\}$ を満たすもの）へ所属させることで、最終的なクラスタリング結果 $\mathcal{C}^{(T)}$ を得て、出力とする。なお、Algorithm 1 に示されている通り、Step-2 と 3 は、複数回繰り返すことが可能であるが、後述の実験では、二回以上の反復は効果が無いという結果となった。

Algorithm 1 第二段階クラスタリング

Step-1: 素性-文書行列 P を本文中の式に従って定義する。

Step-2: $\mathrm{R}_F^{(t)} = \frac{1}{|D|}\mathrm{P}\mathrm{R}_D^{(t)}$
Step-3: $\mathrm{R}_D^{(t+1)} = \frac{1}{|F|}\mathrm{P}^{\mathrm{T}}\mathrm{R}_F^{(t)}$
$\Big\}\ t \in 0, \cdots, T-1$ で繰り返し

Step-4: 新たなクラスタ $\mathcal{C}^{(T)}$ を定義

5. 実験

5.1 実験設定

実験は、WePS データセットを用いて行った。WePS は、人名検索に関する評価型ワークショップであり、人名曖昧性解消を主要なタスクとしている。2007 年に開催された WePS-1[2] では 16 チーム、2009 年に開催された WePS-2[3] では 17 チームが参加した。WePS-1 テストセット[10]は 30 個の名前で構成され、名前毎に 100 個の Web 文書と、その人手による正解クラスタリング結果が提供される。WePS-2 テストセット[11] は、同じく 30 個の名前で構成され、名前毎に 150 個の Web 文書および正解クラスタリング結果が提供されている。

二段階クラスタリングの比較評価のため、異なる二種類のベースラインを用いた。一つはトピック推定法 [18] であり、弱い素性（単語素性）を利用する。各文書に対応する**潜在トピック**を求め、二つの文書の潜在トピック（その文書に対して確率最大となるトピック）が一致した場合、その文書どうしをまとめる。もう一つはキーワード抽出法 [10] であり、第一段階で出力された各クラスタから、（クラスタ内文書を一つの大きな文書と見なすことによって）複合語を抽出し、クラスタ再構成に用いる手法である。

第二段階クラスタリングの素性としては、単語素性（ユニグラム）のほか、二単語の連接（バイグラム）についても実験を行った。このとき、事前に設定したストップワードリストに含まれる単語は除いた。各素性の重みは、TF-IDF により設定する。IDF 値の計算には、Web 1T 5-gram[12] を利用した。クラスタリングのための閾値 θ_f は、WePS-1 と WePS-2 を交互に用いて決定した。（すなわち、例えば WePS-2 での実験では、パラメータは WePS-1 のデータを訓練データとして用い決定した。）

WePS の公式評価尺度としては、WePS-1 で提案された Purity (P), Inverse Purity (IP) と、その欠点を克服するために WePS-2 で提案された extended B-Cubed Precision (BEP),

[10] http://nlp.uned.es/weps/weps-1-data/
[11] http://nlp.uned.es/weps/weps-2-data/
[12] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

表 1　WePS-1 データセットでの結果

Method	BEP	BER	F_B	P	IP	F_P
WePS-1 データセットでの結果						
ORIGINAL	0.84	0.73	**0.77**	0.82	0.73	0.76
CKW	0.82	0.76	**0.77**	0.84	0.73	0.77
TOPIC	0.81	0.73	0.76	0.72	0.82	0.75
PROPOSED	0.82	0.76	**0.77**	0.83	0.72	0.76
WePS 1st	0.67	0.81	0.71	0.72	0.88	**0.79**
WePS 2nd	0.68	0.73	0.68	0.75	0.80	0.77
WePS 3rd	0.68	0.71	0.67	0.73	0.82	0.77
WePS 4th	0.79	0.50	0.58	0.81	0.60	0.69
WePS-2 データセットでの結果						
ORIGINAL	0.92	0.70	0.78	0.94	0.79	0.86
CKW	0.87	0.77	0.81	0.91	0.84	0.87
TOPIC	0.94	0.70	0.79	0.72	0.79	0.75
PROPOSED	0.89	0.82	**0.85**	0.93	0.87	**0.89**
WePS 1st	0.87	0.79	0.82	0.91	0.86	0.88

extended B-Cubed Recall (BER) があり、現在では後者が公式評価尺度として採用されているが、ここでは両者ともに用いた結果を報告する。それぞれの詳しい定義については、誌面の都合上文献 [2, 1] に譲るが、P, BEP が一般的な定義による精度に、IP, BER が一般的な再現率に相当する。F 値は、精度と再現率の調和平均であり、システムの総合的な良さの指標として用いられる。(F_B が BEP と BER による F 値、F_P が P と IP による F 値。)

5.2　結果

表 1 に、両データセットによる結果を示す。“ORIGINAL” が第一段階クラスタリングのみを用いた場合、“CKW” がキーワード抽出法による第二段階クラスタリング、“TOPIC” がトピック推定法による第二段階クラスタリング、そして “PROPOSED” が提案手法となる。(反復回数は一回、素性は 1-gram を使用。その他の条件については後述。) WePS-1 参加システムの BEP,BER 値は [1] より引用した。WePS-1 におけるトップシステムの手法は、固有名詞素性を主に用い、HAC(階層併合型クラスタリング) によりクラスタリングを行っている。

WePS-2 データセットにおいて、CKW (キーワード抽出法) は、ORIGINAL を上回る F 値を達成したが、WePS-1 データセットにおいては改善しなかった。トピック法 (TOPIC) を用いた結果も、WePS-2 では改善したが、WePS-1 では逆に F 値は悪化した。提案手法も WePS-1 では改善が見られなかったが、WePS-2 での結果は、B-Cubed の F 値で 0.85 という高い値を達成した。トピック法と提案手法の F 値で大きな差が出たが、これは、トピック法では、提案手法のような「トピック特有の単語」に高い重みを与えることが難しいためと思われる。提案手法が WePS-2 データセットにおいて高い F 値を達成した理由としては、WePS-2 データセットが、第一段階クラスタリングが高い精度を達成できるような性質を持っていたことが考えられる。WePS-1 データセットにおいては、そのような性質が弱かったと思われるが、その場合でも性能 (F 値) は悪化しないという結果となった。

提案手法は、BEP-BER の F 値において WePS-1 のトップシステムを上回り、P-IP の F 値においてもトップ 3 に次ぐ値を達成している。(ただし、[11] による大量の Web 検索を用いる手法には劣っている。) P-IP の F 値で若干トップ 3 に劣った結果となったが、これは、WePS-1 のシステムが P-IP の尺度に対し最適化されているのに対し、提案手法は WePS-2 に準拠した BEP-BER 尺度に対しパラメータ学習等を行っていることが理由ではないかと推察される。“CKW” は WePS-2 に参加して 2 位となったアルゴリズムであるが、WePS-2 のデータに対して、提案手法 “PROPOSED” は CKW を 0.04 ポイント上回った。これは、WePS-2 のトップシステムを上回る結果である。

単語連接 (2-gram) を素性として用いた場合についても実験したが、第一段階のクラスタから変化せず、F 値の改善は見られなかった。一般的に、2-gram 素性は疎 (同一素性が 2 回以上登場しづらい) であり、提案手法のような「弱い素性の登場頻度を統計的手法で利用する」手法や、本研究のような比較的小規模のデータセットに対するクラスタリングタスクにおいては有効に働かなかったものと思われる。また、単語素性 (1-gram) を用いた場合に、反復回数を増やした実験も行ったが、精度が著しく悪化し、全体の F 値も悪化した。2 回以降の反復は、弱い素性による文書どうしの結びつきが強化され、本来弱い結びつきしか無かった文書同士が同一のクラスタに入ってしまうためであると思われる。このほか、[13] において提案されている、2 回以上の反復を減衰係数と共に行うための、Von Neumann カーネルや Graph Laplacian を用いた手法も試したが、改善は見られなかった。この結果と、上記の $T = 1$ (反復回数 1 回) が $T = 2, 3$ よりも高い F 値を示したという結果は、ブートストラップ法による重みの伝播は一回 (文書クラスタから素性クラスタへの伝播および素性クラスタから文書クラスタへの逆伝播) で十分であり、それ以上の反復は過剰な重みの伝播につながり、精度を悪化させてしまうということが示唆されていると考えられる。

6.　まとめと今後の課題

人名曖昧性解消問題において、固有名詞等の「強い素性」と単語等の「弱い素性」を効率的に併用するための二段階クラスタリングを提案した。実験の結果、提案手法は、WePS-1 データセットにおいては F 値の改善につながらなかったが、WePS-2 データセットにおいて大きく F 値を改善させることがわかった。今後の課題としては、より多種多様なデータによる実験や、ノイズを含む文書の除去手法の導入等が挙げられる。また、反復回数 1 回と 2 回の間で大きくクラスタが変化してしまうため、これらの中間のクラスタを得る手法を開発することも考えている。

[文献]

[1] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4), 2009.

[2] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *SemEval-2007*, pages 64–69, 2007.

[3] J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009)*, 2009.

[4] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL 1998*, pages 79–85, 1998.

[5] K. Balog, L. Azzopardi, and M. Rijke. Personal name resolution of web people search. In *NLPIX 2008*, 2008.

[6] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *WWW2005*, pages 463–470, 2005.

[7] D. Bollegala, Y. Matsuo, and M. Ishizuka. Extracting key phrases to disambiguate personal name queries in web search. In *Proc. of the Workshop: How can Computational Linguistics improve Information Retreival? at COLING-ACL 2006*, pages 17–24, 2006.

[8] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL-06*, 2006.

[9] E. Elmacioglu, Y. F. Tan, S. Yan, M. Kan, and D. Lee. PSNUS: Web people name disambiguation by simple clustering with rich features. In *SemEval-2007*, pages 268–271, 2007.

[10] M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. Person Name Disambiguation on the Web by Two-Stage Clustering. In *WePS 2009*, 2009.

[11] D. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse.: a web-querying approach to web people search. In *SIGIR '08*, pages 27–34, 2008.

[12] S. Kataoka, H. Ueda, H. Murakami, and S. Tatsumi. Two-step clustering based on person names to identify diiferent people with identical names on the web. In *Proc. of the 22nd Annual Conf. of the Japanese Society for Artificial Intelligence (In Japanese)*, 2008.

[13] M. Komachi, T. Kudo, M. Shimbo, and Y. Matsumoto. Graph-based analysis of semantic drift in espresso-like bootstrapping algorithms. In *EMNLP 2008*, pages 1010–1019, 2008.

[14] X. Liu, Y. Gong, W. Xu, and S. Zhu. Document clustering with cluster refinement and model selection capabilities. In *SIGIR 2002*, pages 191–198. ACM Press, 2002.

[15] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *CoNLL2003*, pages 33–40, 2003.

[16] H. Nakagawa and T. Mori. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219, 2003.

[17] C. Niu, W. Li, and R. K. Srihari. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. In *ACL-2004*, pages 598–605, 2004.

[18] S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. Person name disambiguation in web pages using social network, compound words and latent topics. In *PAKDD2008*, pages 260–271, 2008.

[19] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *COLING-ACL 2006*, pages 113–120, 2006.

[20] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *SIGIR 2000*, pages 208–215. ACM New York, NY, USA, 2000.

[21] R. N. Turan, Z. Chen, D. Kalashnikov, and S. Mehrotra. Exploiting web querying for web people search in WePS2. In *WePS 2009*, 2009.

[22] X. Wan, M. Li J. Gao, and B. Ding. Person resolution in person search results: WebHawk. In *CIKM2005*, pages 163–170, 2005.

吉田 稔 Minoru Yoshida

1998 年東京大学卒。2003 年東京大学大学院博士課程修了。現在、東京大学情報基盤センター助教。Web からの情報抽出に興味を持つ。

池田 雅紀 Masaki Ikeda

2008 年東京大学卒。2010 年東京大学大学院修士課程修了。

小野 真吾 Shingo Ono

2004 年東京大学卒。2009 年東京大学大学院博士課程修了。現在、ヤフー株式会社勤務。

佐藤 一誠 Issei Sato

2006 年早稲田大学卒。2008 年東京大学大学院修士課程修了。現在、同大学院博士課程在学中。

中川 裕志 Hiroshi Nakagawa

1975 年東京大学卒。1980 年東京大学大学院博士課程修了。1999 年まで横浜国立大学。現在、東京大学情報基盤センター教授。機械学習、自然言語処理に興味を持つ。