2007－NL－181 (3)
2007／9／25

Web検索結果における人名の曖昧性解消への半教師有りクラスタリングの適用

杉山 一成† 奥村 学†

† 東京工業大学 精密工学研究所
〒226–8503 神奈川県横浜市緑区長津田町 4259
E-mail: †sugiyama@lr.pi.titech.ac.jp, †oku@pi.titech.ac.jp

あらまし 人名は検索語として，しばしば検索エンジンに入力される．しかし，この入力された人名に対して，検索エンジンは，いくつかの同姓同名人物についての Web ページを含む長い検索結果のリストを返すだけである．Web の検索結果における人名の曖昧性を解消するほとんどの従来研究は，凝集型クラスタリングを適用している．一方，本研究ではある種文書に類似した文書をマージする半教師有りクラスタリングを用いる．我々の提案する半教師有りクラスタリングは，クラスタの重心の変動を抑えるという点において，新規性がある．実験の結果，最適な場合において，purity で 0.72，inverse purity で 0.81，これらの調和平均である F 値で 0.76 の評価値が得られた．

キーワード 情報検索，半教師有りクラスタリング，人名の曖昧性解消

Applying Semi-Supervised Clustering to Personal Name Disambiguation in Web Search Results

Kazunari SUGIYAMA† and Manabu OKUMURA†

† Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, Kanagawa, 226–8503 Japan
E-mail: †sugiyama@lr.pi.titech.ac.jp, †oku@pi.titech.ac.jp

Abstract Personal names are often submitted to search engines as query keywords. However, in response to a personal name query, search engines return a long list of search results that contains Web pages about several namesakes. Most of the previous works that disambiguate personal names in Web search results often employ agglomerative clustering approaches. In contrast, we have adopted a semi-supervised clustering approach to integrate similar documents into a seed document. Our proposed semi-supervised clustering approach is novel in that it controls the fluctuation of the centroid of a cluster. Experimental results show that in the best case, our proposed approach achieved a purity of 0.72 and inverse purity of 0.81, and their harmonic mean F was 0.76.

Key words Information retrieval, Semi-supervised clustering, Personal name disambiguation

1. はじめに

World Wide Web (WWW) 上の情報は増加し続けているため，ユーザの要求に適合する情報を見つけることはますます難しくなっている．こうした状況の中で，Web 検索エンジンは，WWW 上の情報を検索するための有用な手段である．検索エンジン *ALLTheWeb*(注1)の英語の検索語の約 1 割が人名を含むという報告(注2)があるように，特に人名は検索語として検索エンジンにしばしば入力される．しかし，人名が検索語として入力されると，検索エンジンは，その人名に対する同姓同名人物についての Web ページを含む長い検索結果のリストを返すだけである．例えば，ユーザが検索エンジン Google(注3)に “William Cohen” という人名を入力すると，その検索結果には “William Cohen” という名前を有する複数の人物が含まれる．この検索結果には，情報科学の教授，アメリカ合衆国の政治家，外科医，歴史家などが含まれ，これらの人物は各実体ごとに分類されておらず，混在している．

Web 検索結果における人名の曖昧性を解消する関連研究のほとんどは，いくつかの種類の凝集型クラスタリング

(注1)：http://www.alltheweb.com/
(注2)：http://tap.stanford.edu/PeopleSearch.pdf
(注3)：http://www.google.com/

を利用している．しかし，人物の実体について述べているいくつかのWebページを半教師有りの方法で導入すれば，人名の曖昧性解消に対するクラスタリングは，より正確に行なうことができると推測される．以下，本論文では，このようなWebページを「seedページ」と呼ぶことにする．また，Web検索結果における人名の曖昧性を解消するため，クラスタリングの精度を改善できるように，このseedページを用いた半教師有りクラスタリングを適用する．我々の半教師有りクラスタリングは，種文書を含むクラスタの重心の変動を抑える点において新規性がある．

2. 関連研究

本章では，凝集型クラスタリングに基づいた人名の曖昧性解消，ならびに従来の半教師有りクラスタリングの手法について振り返る．

2.1 凝集型クラスタリングに基づいた人名の曖昧性解消

まず，Web検索結果における人名の曖昧性解消に関して，次のような研究を挙げることができる．

Mannら[1]の研究では，まず，誕生日，出生地，職業などの人物情報を抽出する．次に，各文書に対して，抽出した人物情報や固有名と，検索結果中の文書から計算されたTF-IDF値[2]から構成される特徴ベクトルを生成する．最後に，この特徴ベクトルを用い，凝集型クラスタリングアルゴリズムに基づいて人物のクラスタを生成することで，人名の曖昧性を解消する．また，Mannらの研究と同様の手法を用いて，Wanら[3]は，*WebHawk*というシステムを開発した．

Pedersenら[4]は，与えられた名前の実体をグループに分類することによって，人名の曖昧性を解消している．まず，曖昧性を有する名前の各実体に関連する文脈を抽出し，その中から重要なbi-gramをとらえることで，二次的な文脈ベクトルを生成する．次に，互いに類似した実体が同じクラスタに属するように，クラスタリングを行なう．

Bekkermanら[5]は，(1) Webページのハイパーリンク構造に基づいた手法，(2) 凝集／寄せ集め型クラスタリング[6]に基づいた手法，(3) (1)と(2)を組み合わせた手法，の3つの教師無し手法を提案している．これらの手法では，曖昧性を解消すべき人物間のつながりや人物のリストが，既知であることを仮定している．したがって，こうした情報が事前にあるならば，この手法は効果的であると推測される．しかし，ほとんどの場合において，曖昧性を解消したい人物間のつながりについて事前に十分な情報があるわけではない．したがって，この手法は実際の人物検索には適切ではないと考えられる．

Bollegalaら[7]は，まず，対象となる文書集合に対して凝集型クラスタリングを行ない，得られたクラスタから異なる同姓同名人物を識別するために，重要語句を選択する．次に，はじめの対象文書集合から重要語句を抽出し，クラスタと文書集合のそれぞれから抽出された語句の間の類似度に応じて，クラスタどうしをマージする．

2.2 半教師有りクラスタリング

一般に，教師無しクラスタリングは，文書の組織化，閲覧，大規模文書の要約といったデータ解析を行なうために，重要な技術である．また，クラスタリングは，データへのラベル付けが実際的でなかったり，不可能であったりするような大規模なデータ集合を解析する処理において，有用である．このような教師無しクラスタリングは，何らかの教師を用いることによって，その精度が改善され得る．最近では，こうした何らかの教師を用いる，すなわち，半教師有りの手法でクラスタリングの精度を向上させることを目的とした研究が注目されている．

これまでの半教師有りクラスタリングの手法は，(1) 制約に基づいた手法，(2) 距離に基づいた手法，の2つに分類することができる．制約に基づいた手法は，ユーザが付与したラベルや制約を利用し，より適切にデータをクラスタリングできるようにする手法である．例えば，Wagstaffら[8], [9]の半教師有りK−meansアルゴリズムは，“must-link”（2つの事例が同じクラスタに属さなければならない）と，“cannot-link”（2つの事例が異なるクラスタに属さなければならない）という2種類の制約を導入し，これらの制約が侵されないことを保証して，データのクラスタリングを行なう．Basuら[10]もまた，初期の種クラスタを生成し，クラスタリングを正確に行なうために，ラベル付きデータを利用する半教師有りK−meansアルゴリズムを開発している．距離に基づいた手法では，特定のクラスタリング尺度を用いる既存のクラスタリングアルゴリズムが利用されている．しかし，クラスタリング尺度は教師付きデータにおけるラベルや制約を満たすために学習を必要とする．これらの研究では，最短経路アルゴリズム[11]によって修正されたユークリッド距離や最適化を行なったマハラノビス距離[12]などのいくつかの適応的な距離尺度が使われている．

3. 提案手法

2.1で述べた凝集型クラスタリングに基づいた人名の曖昧性解消は，類似したクラスタをマージする基準によっては，単一の要素を含むクラスタを生成しやすい．一方，**2.2**で述べた半教師有りクラスタリングは，クラスタ数Kをあらかじめ設定する必要があるK−means アルゴリズム[13]を改良することを目的としている．しかし，我々の研究においては，Web検索結果における同姓同名人物の数は，事前にわかっているわけではない．また，従来の半教師有りクラスタリングアルゴリズムは，制約を導入したり，距離を学習したりすることにのみ着目している．さらに，これらのアルゴリズムは，クラスタの重心の変動を抑えることを考慮していない．しかし，半教師有りクラスタリングにおいては，より正確なクラスタリング結果を得るためには，制約を導入するとともに，seedページを含むクラスタの重心の変動を抑えることが重要であると考えられる．

本章では，まず**3.1**で，従来研究のほとんどが利用している凝集型クラスタリングの手法について振り返り，次に**3.2**で，我々の提案する半教師有りクラスタリングについて述べる．本研究で提案する半教師有りクラスタリングの手法は，seedページを含むクラスタの重心の変動を抑える点において，新規性がある．

以下の議論において，検索結果集合におけるWebページpの特徴ベクトル$\boldsymbol{w}^p$を式(1)のように表す．

$$\boldsymbol{w}^p = (w^p_{t_1}, w^p_{t_2}, \cdots, w^p_{t_m}) \tag{1}$$

ここで，mはWebページpにおける単語の異なり数であり，t_k $(k = 1, 2, \cdots, m)$は，各単語を表す．

ここで，予備実験として，TF, IDF, residual IDF, TF-IDF, x^I-measure, gain[14]の6つの単語重み付け法を比較した．その結果，gainが我々のタスクにおけるクラスタリングのための特徴ベクトルを生成するのに，最も効果的な単語の重み付け法であることがわかった．このgainを用いた場合，$\boldsymbol{w}^p$の各要素$w^p_{t_k}$は式(2)のように定義される．

$$w^p_{t_k} = \frac{df(t_k)}{N}\left(\frac{df(t_k)}{N} - 1 - \log\frac{df(t_k)}{N}\right) \tag{2}$$

ここで，$df(t_k)$は単語t_kが出現するWebページ数を，Nは検索結果のWebページの総数を表す．さらに，クラスタの重心ベクトル$\boldsymbol{G}$を式(3)のように定義する．

$$\boldsymbol{G} = (g_{t_1}, g_{t_2}, \cdots, g_{t_m}) \tag{3}$$

ここで，g_{t_k}はクラスタの重心ベクトルにおける各単語の重みであり，t_k $(k = 1, 2, \cdots, m)$は，各単語を表す．

3.1 凝集型クラスタリング

凝集型クラスタリングにおいては，はじめに各Webページは，個々のクラスタとして扱われる．次に，二つのクラスタの類似度が，あらかじめ設定された閾値より小さくなるまで，類似度が最も大きくなる二つのクラスタをマージして新たなクラスタを生成する．図1に凝集型クラスタリングアルゴリズムの詳細を示す．このアルゴリ

Algorithm: Agglomerative clustering
Input: Set of search-result Web page $p_i (i = 1, 2, \cdots, n)$,
$Wp = \{p_1, p_2, \cdots, p_n\}$.
Output: Clusters that contain the Web pages
that refer to the same person.
Method:
1. Set each element in Wp as an initial cluster.
2. Repeat the following steps for all p_i $(i = 1, 2, \cdots, n)$ in Wp
until all of the similarities between two clusters are
less than the predefined threshold.
2.1 Compute the similarity between p_i and p_{i+1}.
if the similarity is greater than the predefined threshold,
then merge p_i and p_{i+1} and recompute the centroid
of the cluster using Equation (4),
else p_i is an independent cluster.
2.2 Compute all of the similarities between two clusters.

図 1 凝集型クラスタリングアルゴリズム
Fig. 1 Agglomerative clustering algorithm.

ズムにおいては，あるクラスタを最も類似したクラスタにマージした後の，新たなクラスタの重心ベクトル $\boldsymbol{G}^{new}$ は，(4) 式のように定義される．

$$\boldsymbol{G}^{new} = \frac{\left(\sum_{\boldsymbol{w}^{p(G)} \in \boldsymbol{G}}^{n} \boldsymbol{w}^{p(G)} + \boldsymbol{w}^{p}\right)}{n+1} \tag{4}$$

ここで，$\boldsymbol{w}^{p(G)}$ と n は，それぞれ，重心となるクラスタにおける検索結果の Web ページの特徴ベクトル $\boldsymbol{w}^{p}$，検索結果の Web ページ数を表す．

3.2 提案する半教師有りクラスタリング

1. で述べたように，人物の実体について述べている seed ページを導入すれば，人名の曖昧性解消のために，より正確にクラスタリングを行なうことができると考えられる．したがって，我々は Web 検索結果における人名の曖昧性を解消するために，半教師有りクラスタリングを適用する．我々の提案する手法は，あるクラスタを，seed ページを含むクラスタにマージする際に，そのクラスタの重心の変動を抑える点において新規性がある．この処理において，ある検索結果の Web ページの特徴ベクトル $\boldsymbol{w}^{p}$ を，seed ページを含むクラスタの重心 $\boldsymbol{G}$ にマージする際，$\boldsymbol{G}$ と $\boldsymbol{w}^{p}$ 間の距離 $D(\boldsymbol{G}, \boldsymbol{w}^{p})$ によって，$\boldsymbol{w}^{p}$ の各要素 $w^{p}_{t_k}$ を重み付けする．本研究では，(i) ユークリッド距離，(ii) マハラノビス距離，(iii) 適応的マハラノビス距離，の三つの距離尺度を用いる．式 (1) と式 (3) を用いると，あるクラスタと，それに最も類似したクラスタをマージした後の新しいクラスタの重心ベクトル $\boldsymbol{G}^{new}$ は，式 (5) のように定義される．

$$\boldsymbol{G}^{new} = \frac{\left(\sum_{\boldsymbol{w}^{p(G)} \in \boldsymbol{G}}^{n} \boldsymbol{w}^{p(G)} + \frac{\boldsymbol{w}^{p}}{D(\boldsymbol{G}, \boldsymbol{w}^{p})}\right)}{n+1} \tag{5}$$

ここで，$\boldsymbol{w}^{p(G)}$ と n は，それぞれ，重心となるクラスタにおける検索結果の Web ページの特徴ベクトル $\boldsymbol{w}^{p}$，検索結果の Web ページ数を表す．上述した (i)，(ii)，(iii) の 3 つの距離は，それぞれ以下のように定義される．図 2 に，我々の提案する半教師有りクラスタリングアルゴリズムの詳細を示す．

(i) ユークリッド距離

式 (5) において，ユークリッド距離を導入した場合，クラスタの重心ベクトル $\boldsymbol{G}$ と検索結果の Web ページ p の特徴ベクトル $\boldsymbol{w}^{p}$ 間の距離 $D(\boldsymbol{G}, \boldsymbol{w}^{p})$ は，式 (6) のように定義される．

$$D(\boldsymbol{G}, \boldsymbol{w}^{p}) = \sqrt{\sum_{k=1}^{m} (g_{t_k} - w^{p}_{t_k})^2} \tag{6}$$

(ii) マハラノビス距離

マハラノビス距離は，データ集合の相関を考慮しており，尺度水準に独立であるという点において，ユークリッド距離とは異なる．したがって，ユークリッド距離を用いるよりもマハラノビス距離を用いた方が，クラスタの重心の変動を，よりうまく抑えられることが期待される．

式 (5) において，マハラノビス距離を導入した場合，クラスタの重心ベクトル $\boldsymbol{G}$ と検索結果の Web ページ p の特徴ベクトル $\boldsymbol{w}^{p}$ 間の距離 $D(\boldsymbol{G}, \boldsymbol{w}^{p})$ は，式 (7) のように定義される．

$$D(\boldsymbol{G}, \boldsymbol{w}^{p}) = \sqrt{(\boldsymbol{w}^{p} - \boldsymbol{G})^{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}^{p} - \boldsymbol{G})} \tag{7}$$

ここで，$\boldsymbol{\Sigma}$ は，クラスタの重心におけるメンバによって定義される共分散行列である．

(iii) 適応的マハラノビス距離

この手法においては，各クラスタにおけるマハラノビス距離は，次のように導出される．$R = \{p_1, p_2, \cdots, p_n\}$ を n 個の検索結果の Web ページとする．各 Web ページ p_i $(i = 1, \cdots, n)$ は，式 (1) で定義されるベクトルとして表現される．R を K 個のクラスタ $C_1, \cdots, C_K$ に分割したものを P_C とし，各クラスタ C_k $(k = 1, \cdots, K)$ は，式 (8) のベクトルとして表現される代表点 $\boldsymbol{v}^{k}$ を持つものとする．

$$\boldsymbol{v}^{k} = (v^{k}_{t_1}, v^{k}_{t_2}, \cdots, v^{k}_{t_m}) \tag{8}$$

我々の手法は，検索結果の Web ページの集合を K 個のクラスタに分割した $P_C = (C_1, \cdots, C_K)$，ならびに，対応する代表点の集合 $L = (L_1, \cdots, L_K)$ とそのクラスタに関連する K 個の異なる距離を，式 (9) で定義される適切な基準を局所的に最小化することによって求める．

$$\begin{aligned} W(P_C, L) &= \sum_{k=1}^{K} \Delta_k^2(L_k, \delta_k) \\ &= \sum_{k=1}^{K} \sum_{i \in C_k} \delta_k(\boldsymbol{w}^{p_i}, \boldsymbol{v}^{k}) \end{aligned} \tag{9}$$

ここで，$\delta_k(\boldsymbol{w}^{p_i}, \boldsymbol{v}^{k})$ は検索結果の Web ページ $p_i \in C_k$ とクラスタ C_k の代表点 L_k との間の適応的な非類似度である．

この距離を計算する際には，C_k のクラスタ内構造に応じて，検索結果の Web ページ p_i と代表点 L_k との間の適応的マハラノビス距離を考慮する．この適応的マハラノビス距離 $\delta_k(\boldsymbol{w}^{p_i}, \boldsymbol{v}^{k})$ は，式 (10) で定義される．

$$\delta_k(\boldsymbol{w}^{p_i}, \boldsymbol{v}^{k}) = (\boldsymbol{w}^{p_i} - \boldsymbol{v}^{k})^{T} \boldsymbol{M}_k^{-1} (\boldsymbol{w}^{p_i} - \boldsymbol{v}^{k}) \tag{10}$$

ここで，$\boldsymbol{M}_k$ は，クラスタ C_k に関する共分散行列である．また我々は，式 (9) における $\Delta_k^2(L_k, \delta_k)$ を，次のように最適化する．

(a) まず，クラスタ C_k と共分散行列 $\boldsymbol{M}_k (k = 1, \cdots, K)$ を固定する．式 (9) と式 (10) に基づいて，局所的に最小化するクラスタ C_k の代表点 L_k を求める．

$$\Delta_k^2(L_k, \delta_k) = \sum_{i \in C_k} (\boldsymbol{w}^{p_i} - \boldsymbol{v}^{k})^{T} \boldsymbol{M}_k^{-1} (\boldsymbol{w}^{p_i} - \boldsymbol{v}^{k}) \tag{11}$$

Algorithm: Semi-supervised clustering
Input: Set of search-result Web page $p_i (i = 1, 2, \cdots, n)$,
and seed pages $p_{s_j} (i = 1, 2, \cdots, u)$,
$Wp = \{p_1, p_2, \cdots, p_n, p_{s_1}, p_{s_2}, \cdots, p_{s_u}\}$.
Output: Clusters that contain the Web pages
that refer to the same person.
Method:
1. Set each element in Wp as an initial cluster.
2. Repeat the following steps for all p_i $(i = 1, 2, \cdots, n)$ in Wp
2.1 Compute the similarity between p_i and p_{s_j}.
if the maximum similarity is obtained between p_i and p_{s_j},
then merge p_i into p_{s_j} and recompute the centroid
of the cluster using Equation (5),
else p_i is stored as other clusters Oth, namely, $Oth = \{p_i\}$.
3. Repeat the following steps for all p_h $(h = 1, 2, \cdots, m, (m < n))$
in Oth until all of the similarities between two clusters are
less than the predefined threshold.
3.1 Compute the maximum is obtained between p_h and p_{h+1}
if the similarity is greater than the predefined threshold,
then merge p_h and p_{h+1} and recompute the centroid
of the cluster using Equation (4),
else p_h is an independent cluster.
3.2 Compute all of the similarities between two clusters.

図 2 提案する半教師有りクラスタリングアルゴリズム
Fig. 2 Our proposed semi-supervised clustering algorithm.

式 (11) の解は，クラスタ C_k の重心となる．
(b) 次に，クラスタ C_k と，その代表点 $L_k(k=1,\cdots,K)$ を固定し，$det(\boldsymbol{M}_k)=1$ の下で基準 $\Delta_k^2(L_k,\delta_k)$ を局所的に最小化するようなクラスタ C_k の距離 δ_k を求める．文献 [15] によれば，その解は $\boldsymbol{M}_k=(det\boldsymbol{Q}_k)^{1/p}\boldsymbol{Q}_k^{-1}$ となる．ここで，$\boldsymbol{Q}_k$ は，$det(\boldsymbol{Q}_k)\neq 0$ の下で，クラスタ C_k に属する適応的共分散行列である．最終的には，$\boldsymbol{M}_k$ を用いて，適応的マハラノビス距離は式 (12) のように定義される

$$D(\boldsymbol{G}_k,\boldsymbol{w}^p)=\sqrt{(\boldsymbol{w}^p-\boldsymbol{G}_k)^T\boldsymbol{M}_k^{-1}(\boldsymbol{w}^p-\boldsymbol{G}_k)} \quad (12)$$

本研究では，上述した 3 種類の距離を用いることで，我々の提案する半教師有りクラスタリング手法による人名の曖昧性解消の精度を比較する．

4. 実　　験

4.1 実験データ

我々の実験では，“Web People Search Task” [16] において作成された WePS コーパスを用いた．この WePS コーパスは 79 名の人物集合から構成され，その集合の各人物は，人名を検索語として，Yahoo!(注4) の検索 API を通じて得られた上位 100 件の検索結果に対応する．すなわち，このコーパスは約 7,900 の Web ページから構成され，訓練集合とテスト集合には，それぞれ，49 と 30 の人名を含む．これらの人名は電子図書館と計算言語学に関する会議の参加者、英語の Wikipedia における人物情報の記事，アメリカ合衆国の国勢調査から抽出されている．まず前処理として，このコーパスにおけるすべての Web ページに対して，不要語リスト(注5) に基づいて，不要語を取り除き，Porter Stemmer [17] (注6) を用いて語幹処理を行なった．

次に WePS コーパスの訓練集合を用いて類似のクラスタをマージするための最適なパラメータを決定し，これを WePS コーパスのテスト集合に適用した．

4.2 評価尺度

本研究では，“purity”，“inverse purity” と，これらの調和平均である F 値に基づいて，クラスタリングの精度を評価する．これらは，Web People Search Task において採用されている標準的な評価尺度である．“purity” は「適合率」の尺度に関連する．この尺度では，各クラスタにおいて最もよく現れるカテゴリの頻度に注目し，ノイズの少ないクラスタを高く評価する．C を評価されるべきクラスタの集合，L を人手で作成したカテゴリの集合，n を生成されたクラスタ数とすると，purity は，式 (13) に基づいて，最大となる適合率の重み付き平均をとることで計算される．

$$Purity=\sum_i \frac{|C_i|}{n}\max Precision(C_i,L_j) \quad (13)$$

ここで，ある与えられたカテゴリ L_j に対するクラスタ C_i の適合率 $Precision(C_i,L_j)$ は，式 (14) によって定義される．

$$Precision(C_i,L_j)=\frac{|C_i\bigcap L_j|}{|C_i|} \quad (14)$$

“inverse purity” は，各カテゴリに対して最大の再現率となるクラスタに着目する．ある一つのクラスタにおいて，各カテゴリで定められた要素が多く集まったクラスタを高く評価する．inverse purity は，式 (15) によって定義される．

$$InversePurity=\sum_j \frac{|L_j|}{n}\max Recall(C_i,L_j) \quad (15)$$

(注4)：http://www.yahoo.com/
(注5)：ftp://ftp.cs.cornell.edu/pub/smart/english.stop
(注6)：http://www.tartarus.org/~martin/PorterStemmer/

表 1　凝集型クラスタリングを用いて得られたクラスタリング精度
Table 1　Clustering accuracy obtained using agglomerative clustering.

Purity	Inverse purity	F
0.66	0.49	0.51

表 2　1 つの seed ページを使い，半教師有りクラスタリングを用いて得られたクラスタリング精度
Table 2　Clustering accuracy obtained using our proposed semi-supervised clustering with one seed page.

Distance measure	Seed page	Purity	Inverse purity	F
(i) Euclidean distance	(a) Wikipedia article	0.39	0.90	0.54
	(b) Top-ranked Web page	0.40	0.82	0.54
(ii) Mahalanobis distance	(a) Wikipedia article	0.44	0.96	0.55
	(b) Top-ranked Web page	0.47	0.81	0.60
(iii) Adapative Mahalanobis distance	(a) Wikipedia article	0.48	0.88	**0.62**
	(b) Top-ranked Web page	0.50	0.78	0.61

ここで，ある与えられたカテゴリ L_j に対するクラスタ C_i の再現率 $Recall(C_i,L_j)$ は，式 (16) によって定義される．

$$Recall(C_i,L_j)=\frac{|C_i\bigcap L_j|}{|L_j|} \quad (16)$$

また，purity と inverse purity の調和平均 F は，式 (17) によって定義される．

$$F=\frac{1}{\alpha\frac{1}{Purity}+(1-\alpha)\frac{1}{InversePurity}} \quad (17)$$

なお、本研究では，$\alpha=0.5$ として，評価を行なった．

4.3 実験結果

4.3.1 凝集型クラスタリングを用いた実験結果

凝集型クラスタリングによって得られた精度を表 1 に示す．

4.3.2 半教師有りクラスタリングを用いた実験結果

我々の提案する半教師有りクラスタリングの手法では，次の 2 種類の seed ページを用いた．

(a) Wikipedia [18] における各人物の記事，
(b) Web 検索結果において 1 位に順位付けされた Web ページ．

はじめに，一つの seed ページを用いて実験を行なった．しかしながら，**4.1** で述べた WePS コーパスのテスト集合における各人名が，必ずしも Wikipedia に対応する記事を有するわけではない．したがって，もしある人名が Wikipedia に記事を有するのであれば，これを seed ページとして用いた．そうでなければ，Web 検索結果において 1 位に順位付けされた Web ページを用いた．表 2 に，一つの seed ページでの半教師有りクラスタリングを用いて得られたクラスタリング精度を示す．WePS コーパスのテスト集合における 30 の人名のうち，16 の人名に対しては Wikipedia の記事を，14 の人名に対しては 1 位に順位付けされた Web ページを用いた．なお，人名の曖昧性解消に Wikipedia を利用した最近の研究として，Bunescu [19] らは，Wikipedia の構造を用いることによって，固有名を同定し，その曖昧性を解消している．

さらに，一つの seed ページを用いた実験で，最適な F 値が得られた適応的マハラノビス距離に関して，seed ページの数を変えることによって，さらなる実験を行なった．図 3, 4 は，それぞれ，複数の Wikipedia 記事，上位 5 位までに順位付けされた Web ページを用いて得られたクラスタリング精度を示す．

4.3.3 文書を部分的に用いた実験結果

4.3.1 と **4.3.2** で述べた実験では，検索結果の Web ページと seed ページの全文を用いた．しかし，その人物を特徴付ける単語は，人名の周囲にしばしば現れることが観察される．また，検索結果のスニペットにもそのような単語が出現しやすいことが観察される．そこで，少数の seed ページを用いて最適な結果が得られている場合，すなわち，図 3 において，5 つの Wikipedia 記事を用いた場合 (purity:0.53, inverse purity:0.96, F: 0.71) に，さら

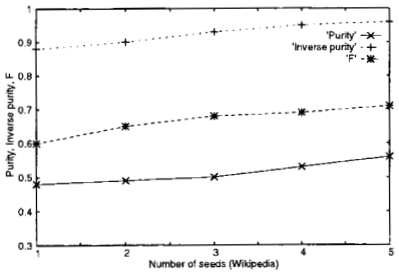

図 3　複数の seed ページを用いて得られたクラスタリング精度 (5 つまでの Wikipedia 記事)

Fig. 3　Clustering accuracy obtained using multiple seed pages (5 Wikipedia articles).

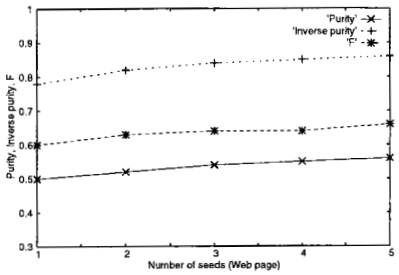

図 4　複数の seed ページを用いて得られたクラスタリング精度 (上位 5 位まで に順位付けされた Web ページ)

Fig. 4　Clustering accuracy obtained using multiple seed pages (Web pages ranked up to the top 5).

に精度が改善されるかを確認するために,

(i)　seed ページと検索結果の Web ページにおいて，人名前後の単語，および文の数を変化させる,

(ii)　検索結果のスニペットを用いる,

実験を行なった．(i) については，まず，WePS コーパスの訓練集合を用いて，最適な F 値を与える seed ページと検索結果の Web ページで用いる人名前後の単語数，または文数を求める．この結果を図 5 に示す．次に最適な F 値を与えるこれらのパラメータをテスト集合に適用し，評価を行なう．(ii) についても同様に，WePS コーパスの訓練集合を用いて，最適な F 値を与える seed ページでの人名前後の単語数，または文数を求める．この結果を図 6 に示す．次に最適な F 値を与えるパラメータをテスト集合に適用し，評価を行う．最終的に (i), (ii) によって得られるクラスタリング精度を，表 3 に示す．

4.3.4　他手法との比較

表 3 にはさらに，“Web People Search Task” において，クラスタリング精度 (F 値) が上位 5 チームの結果を示す．なお，各手法の詳細については，表 3 中の文献を参照されたい．

4.4　考　　察

3.1 で述べた凝集型クラスタリング手法において，表 1 から，purity (0.66) は，inverse purity (0.49) よりも高いことがわかる．このように，inverse purity が低く，purity が高いことは，凝集型クラスタリングが，一つの文書しか含まないクラスタを生成する傾向にあることを示す．

3.2 で述べた半教師有りクラスタリング手法において，表 2 から，ほとんどの場合において，purity の値は，凝集型クラスタリングを用いて得られた purity の値を上回ることができなかったが，inverse purity と F 値に関しては，すべての手法が凝集型クラスタリングの結果を上回っていることがわかる．これは，種文書を含むクラスタの重心の変動を抑えられたことによる効果であると考えられる．我々の提案する半教師有りクラスタリング手法にお

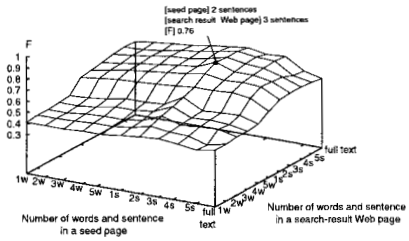

図 5　図 3 における 5 つの seed ページ (Wikipedia 記事) の場合に，seed ページと検索結果の Web ページで用いる人名前後の単語数と文数を変化させて得られるクラスタリング精度 (“w” と “s” は，それぞれ「単語」と「文」を表す)

Fig. 5　Clustering accuracy obtained varying the number of words and sentences backward and forward from a personal name in the seed pages and a search-result Web page in the case of the 5 seed pages (Wikipedia article) shown in Fig. 3 (“w” and “s” denote “word” and “sentence,” respectively).

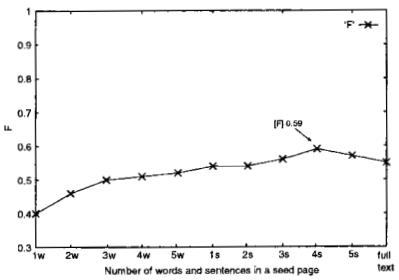

図 6　図 3 における 5 つの seed ページ (Wikipedia 記事) の場合に，検索結果のスニペットを用い，seed ページ中の人名前後の単語数と文数を変化させて得られるクラスタリング精度 (“w” と “s” は，それぞれ「単語」と「文」を表す)

Fig. 6　Clustering accuracy obtained using snippets, and varying the number of words and sentences backward and forward from a personal name in the seed page in the case of the 5 seed pages shown in Fig. 3 (“w” and “s” denote “word” and “sentence,” respectively).

いては，seed ページとして Wikipedia の記事を用い，適応的マハラノビス距離を適用した場合において，最適な F 値 (0.62) が得られた．さらに，複数の seed ページを用いた半教師ありクラスタリング手法においては，図 3, 4 によれば，種文書の数が増加するにつれて，purity と inverse purity の両方の値とも改善されることを示す．これは，seed ページを導入することによって，クラスタリングをより適切に導くことが可能になることを示している．

また，文書を部分的に用いた場合には，次のような傾向が観察される．

まず，WePS コーパスの訓練データにおいて，seed ページ，および検索結果の Web ページ中の人名前後の単語数または文数を変化させた場合，図 5 から，検索結果の Web ページに関して，単語よりも文を用いることで，より良いクラスタリング精度が得られることが観察される．これは，人名前後数語の少ない情報では，人物の実体を識別することは難しいが，人名前後の数文を用いることで，その人物を特徴付ける情報を獲得でき，人物の実体が識別しやすくなったことによるものであると考えられる．また，図 5 からは，seed ページ，検索結果の Web ページについて，それぞれ，人名前後の 2 文，3 文を用いた場合に最適な結果 (F:0.76) が得られることがわかった．これらの文数を WePS コーパスのテスト集合に適用した場合，[purity:0.72, inverse puity: 0.81, F:0.76] の結果が得られた．これは，表 3 に示した “Web People Search Task” [16] の上位 5 チームの結果の中での最適な

表 3 Web People Search Task における上位 5 チーム，および提案手法とのクラスタリング精度の比較

Table 3 Comparison of clustering accuracy obtained by the top 5 participants and our proposed.

Team-ID	Purity	Inverse purity	F
CU_COMSEM [20]	0.72	0.88	**0.78**
IRST-BP [21]	0.75	0.80	0.75
PSNUS [22]	0.73	0.82	0.75
UVA [23]	0.81	0.60	0.67
SHEF [24]	0.60	0.82	0.66
Our proposed method			
(i) 2 and 3 sentences in 5 Wikipedia seed pages and a search result Web page, respectively	0.72	0.81	**0.76**
(ii) Snippet and 4 sentences in 5 Wikipedia seed pages	0.64	0.52	0.57

結果 [purity:0.72, inverse purity:0.88, F:0.78] [20] に続くものであることがわかる．

次に，検索結果のスニペットを用い，WePS コーパスの訓練データにおいて，seed ページ中の人名前後の単語数または文数を変化させた場合 (図 6)，同様に，seed ページ中の人名前後の単語ではなく，文を用いたときに，より良いクラスタリング精度が得られることが観察される．これもやはり，人名前後数語の少ない情報よりも，人名前後の数文を用いることで，その人物を特徴付ける情報が獲得でき，人物の実体が識別しやすくなったことによるものと考えられる．また，図 6 からは，seed ページについて，人名前後の 4 文を用いた場合に最適な結果 (F:0.59) が得られることがわかった．この文数を WePS コーパスのテスト集合に適用した場合，[purity:0.64, inverse puity: 0.52, F:0.57] の結果が得られた．この結果は，Web People Search Task の上位 5 チームの結果，および本研究で行なった他の実験結果と比較して，かなり劣っている．これは，スニペットのような数語程度の情報だけでは，たとえ seed ページで人名前後の 4 文という内容を用いたとしても，その seed ページには人物の実体を述べている検索結果の Web ページが適切に集まらないため，クラスタリングの精度が悪くなることによると考えられる．

5. おわりに

本論文では，Web 検索結果における人名の曖昧性を解消するための半教師有りクラスタリングの手法を提案した．我々の手法は，seed ページを含むクラスタの重心の変動を抑えることによって，より正確な半教師有りクラスタリングを実現する点において新規性がある．また，クラスタの重心の変動を抑えるために，いくつかの距離尺度を導入した．実験の結果，適応的マハラノビス距離を用い，seed ページにおいては曖昧性のある名前の前後 2 文を，検索結果の Web ページにおいては，曖昧性のある名前の前後 3 文を同時に用いた場合に，最適な F 値 (0.76) を達成することができた．

今後の課題として，Web 検索結果における人名の曖昧性を解消するために，対象ページからハイパーリンクで結ばれた Web ページを用いることや，提案手法を Web 検索結果における地名の曖昧性解消に拡張することが挙げられる．

文　献

[1] G. S. Mann and D. Yarowsky. Unsupervised Personal Name Disambiguation. In *Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pp. 33–40, 2003.

[2] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[3] X. Wan, J. Gao, M. Li, and B. Ding. Person Resolution in Person Search Results: WebHawk. In *Proc. of the 14th International Conference on Information and Knowledge Management (CIKM '05)*, pp. 163–170, 2005.

[4] T. Pedersen, A. Purandare, and A. Kulkarni. Name Discrimination by Clustering Similar Contexts. In *Proc. of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2005)*, pp. 226–237, 2005.

[5] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way Distributional Clustering via Pairwise Interactions. In *Proc. of the 22nd International Conference on Machine Learning (ICML2005)*, pp. 41–48, 2005.

[6] R. Bekkerman and A. McCallum. Disambiguating Web Appearances of People in a Social Network. In *Proc. of the 14th International World Wide Web Conference (WWW2005)*, pp. 463–470, 2005.

[7] D. Bollegala, Y. Matsuo, and M. Ishizuka. Extracting Key Phrases to Disambiguate Personal Names on the Web. In *Proc. of the 7th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2006)*, pp. 223–234, 2006.

[8] K. Wagstaff and C. Cardie. Clustering with Instance-level Constraints. In *Proc. of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1103–1110, 2000.

[9] K. Wagstaff and S. Rogers and S. Schroedl. Constrained K-means Clustering with Background Knowledge. In *Proc. of the 17th International Conference on Machine Learning (ICML 2001)*, pp. 577–584, 2001.

[10] S. Basu and A. Banerjee and R. Mooney. Semi-supervised Clustering by Seeding. In *Proc. of the 19th International Conference on Machine Learning (ICML 2002)*, pp. 27–34, 2002.

[11] D. Klein and S. D. Kamvar and C. D. Manning. From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering. In *Proc. of the 19th International Conference on Machine Learning (ICML 2002)*, pp. 307–314, 2002.

[12] E. P. Xing and A. Y. Ng and M. I. Jordan and S. J. Russell. Distance Metric Learning with Application to Clustering with Side-Information. *Advances in Neural Information Processing Systems*, Vol. 15, pp. 521–528, 2003.

[13] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability*, pp. 281–297, 1967.

[14] K. Papineni. Why Inverse Document Frequency? In *Proc. of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pp. 25–32, 2001.

[15] E. Diday and G. Govaert. Classification Automatique Avec Distances Adaptatives. *R.A.I.R.O. Informatique Computer Science*, Vol. 11, No. 4, pp. 329–349, 1977.

[16] J. Artiles and J. Gonzalo and S. Sekine. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 64–69, 2007.

[17] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, Vol. 14, No. 3, pp. pages 130–137, 1980.

[18] M. Remy. Wikipedia: The Free Encyclopedia. *Online Information Review*, Vol. 26, No. 6, p. 434, 2002.

[19] R. Bunescu and M. Paşca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pp. 9–16, 2006.

[20] Y. Chen and J. Martin. CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 125–128, 2007.

[21] O. Popescu and B. Magnini. IRST-BP: Web People Search Using Name Entities. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 195–198, 2007.

[22] E. Elmacioglu and Y. F. Tan and S. Yan and M.-Y. Kan and D. Lee. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 268–271, 2007.

[23] K. Balog and L. Azzopardi and M. Rijke. UVA: Language Modeling Techniques for Web People Search. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 468–471, 2007.

[24] H. Saggion. SHEF: Semantic Tagging and Summarization Techniques Applied to Cross-document Coreference. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 292–295, 2007.