半教師有りクラスタリングを用いたWeb検索結果における人名の曖昧性解消

杉山　一成[†]・奥村　　学[††]

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．この問題を解決するために，Web検索結果における人名の曖昧性解消を目的とした従来研究の多くは，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，種文書を含む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．

キーワード：Web情報検索，半教師有りクラスタリング，人名の曖昧性解消

Personal Name Disambiguation in Web Search Results Using a Semi-Supervised Clustering Approach

Kazunari Sugiyama[†] and Manabu Okumura[††]

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. In order to address this problem, 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.

Key Words: *Web information retrieval, Semi-supervised clustering, Personal name disambiguation*

1　はじめに

検索エンジン *ALLTheWeb*[1] において，英語の検索語の約1割が人名を含むという報告[2] があるように，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，その検索結果としては，その人名を有する同姓同名人物についてのWebページを含む長いリストが返されるの

[†] シンガポール国立大学計算機科学科, Department of Computer Science, National University of Singapore
[††] 東京工業大学精密工学研究所, Precision and Intelligence Laboratory, Tokyo Institute of Technology
[1] http://www.alltheweb.com/
[2] http://tap.stanford.edu/PeopleSearch.pdf

みである．例えば，ユーザが検索エンジン Google[3] に “William Cohen” という人名を入力すると，その検索結果には，この名前を有する情報科学の教授，アメリカ合衆国の政治家，外科医，歴史家などの Web ページが，各人物の実体ごとに分類されておらず，混在している．

こうした Web 検索結果における人名の曖昧性を解消する従来研究の多くは，凝集型クラスタリングを利用している (Mann and Yarowsky 2003)，(Pedersen, Purandare, and Kulkarni 2005)，(Bekkerman, El-Yaniv, and McCallum 2005)，(Bollegala, Matsuo, and Ishizuka 2006)．しかし，一般に人名の検索結果では，その上位に，少数の同姓同名だが異なる人物のページが集中する傾向にある．したがって，上位に順位付けされたページを種文書として，クラスタリングを行えば，各人物ごとに検索結果が集まりやすくなり，より正確にクラスタリングができると期待される．以下，本論文では，このような種文書となる Web ページを「seed ページ」と呼ぶことにする．本研究では，この seed ページを用いた半教師有りクラスタリングを，Web 検索結果における人名の曖昧性解消のために適用する．

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，の二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，より正確なクラスタリングを可能にする．例えば，Wagstaff ら (Wagstaff and Cardie 2000)，(Wagstaff, Rogers, and Schroedl 2001) の半教師有り K-means アルゴリズムでは，“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て，データのクラスタリングを行なう．Basu ら (Basu, Banerjee, and Mooney 2002) もまた，ラベルの付与されたデータから初期の種クラスタを生成し，これらの間に制約を導入する半教師有り K-means アルゴリズムを提案している．また，距離に基づいた手法では，教師付きデータとして付与されたラベルや制約を満たすための学習を必要とする．例えば，Klein ら (Klein, Kamvar, and Manning 2002) の研究では，類似した 2 点 (x_i, x_j) 間には “0”，類似していない 2 点間には $(\max_{i,j} D_{ij}) + 1$ と設定した隣接行列を作成して，クラスタリングを行なう．また，Xing ら (Xing, Ng, Jordan, and Russell 2003) の研究では，特徴空間を変換することで，マハラノビス距離の最適化を行う．さらに，Bar-Hillel ら (Bar-Hillel, Hertz, and Shental 2003) の研究では，適切な特徴には大きな重みを，そうでない特徴には小さな重みを与える RCA (Relevant Component Analysis) (Shental, Hertz, Weinshall, and Pavel 2002) により，特徴空間を変換する．一方，我々の提案する半教師有りクラスタリングでは，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．

本論文の構成は次のとおりである．2 章では，我々の提案する新たな半教師有りクラスタリングの手法について説明する．3 章では，提案手法を評価するための実験結果を示し，その結

[3] http://www.google.com/

果について考察する．最後に 4 章では，本論文のまとめと今後の課題について述べる．

2 提案手法

1 章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，クラスタリングを適切に導いていく基準がないため，正確なクラスタリングを行うことは難しい．一方，これまでに提案されている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 (MacQueen 1967) を改良することを目的としている．しかし，本研究においては，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．したがって，我々の手法においては，事前にクラスタ数を設定するのではなく，新たに生成されたクラスタと，すでに生成されているクラスタ間の類似度を計算し，これらの値がすべて，あらかじめ設定した閾値よりも小さくなった場合に，クラスタリングの処理を終え，その時点で生成されているクラスタ数を最終的な同姓同名人物の数とする．

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 (Wagstaff and Cardie 2000)，(Wagstaff et al. 2001)，(Basu et al. 2002)，距離を学習したり (Klein et al. 2002)，(Xing et al. 2003)，(Bar-Hillel et al. 2003) することにのみ着目していた．しかし，半教師有りクラスタリングにおいて，より正確なクラスタリング結果を得るためには，seed ページ間への制約の導入とともに，seed ページを含むクラスタの重心の変動の抑制も重要である．これは，(1) seed ページを導入して半教師有りクラスタリングを行なう場合，通常の重心の計算法では重心の変動が大きくなる傾向にあり，クラスタリングの基準となる seed ページを導入する効果が得られない，(2) 重心を完全に固定して半教師有りクラスタリングを行なう場合，その重心と類似度が高い Web ページしかマージされなくなり，多数の独立したクラスタが生成されやすくなる，という二つの考えに基づく．したがって，seed ページを含むクラスタの重心の変動を抑えることができれば，より適切なクラスタリングが実現できると期待される．

本章では，我々の提案する半教師有りクラスタリングの手法について説明する．

以下，検索結果集合 W_p 中の Web ページ p_i の特徴ベクトル $\boldsymbol{w}^{p_i}$ $(i = 1, \cdots, n)$ を式 (1) のように表す．

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

ここで，m は検索結果集合 W_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．予備実験として，(a) Term Frequency (TF)，(b) Inverse Document Frequency (IDF)，(c) residual IDF (RIDF)，(d) TF-IDF，(e) x^I-measure，(f) gain の 6 つの単語重み付け法を比較した．これらの単語重み付け法は，それぞれ，次のように定義される．

(a) Term Frequency (TF)

TF は，与えられた文書において，ある単語がどれだけ顕著に出現するかを示し，この値が大きければ大きいほど，その単語が文書の内容をよく表現していることを示す．$tf(t_k, p_i)$ を Web ページ p_i における単語 t_k の頻度とする．このとき，$\boldsymbol{w}^{p_i}$ の各要素 $w_{t_k}^{p_i}$ は，式 (2) によって定義される．

$$w_{t_k}^{p_i} = \frac{tf(t_k, p_i)}{\sum_{s=1}^{m} tf(t_s, p_i)} \tag{2}$$

(b) Inverse Document Frequency (IDF)

(Jones 1973) によって導入された IDF は，その単語が出現する文書数が少なければ少ないほど，その単語が出現する文書にとっては，有用であることを示すスコアである．このとき，$\boldsymbol{w}^{p_i}$ の各要素 $w_{t_k}^{p_i}$ は，式 (3) によって定義される．

$$w_{t_k}^{p_i} = \log \frac{N}{df(t_k)} \tag{3}$$

ここで，N は Web ページの総数，$df(t_k)$ は単語 t_k が現れる Web ページ数である．

(c) Residual Inverse Document Frequency (RIDF)

Church and Gale (Church and Gale 1995a, 1995b) は，ほとんどすべての単語は，ポアッソンモデルのような独立性に基づいたモデルに応じて，非常に大きな IDF スコアを持つことを示した．また，単語の有用性は，推定されるスコアからは大きな偏差を持つ傾向があるという考えに基づいて導入したスコアが residual IDF である．このスコアは，実際の IDF とポアッソン分布によって推定される IDF との差として定義される．cf_k を文書集合中における単語 t_k の総出現数，N を Web ページの総数としたとき，1 つの Web ページあたりの単語 t_k の平均出現数は，$\lambda_k = \frac{cf_k}{N}$ と表される．このとき，$\boldsymbol{w}^{p_i}$ の各要素 $w_{t_k}^{p_i}$ は，式 (4) によって定義される．

$$\begin{aligned} w_{t_k}^{p_i} &= IDF - \log \frac{1}{1 - p(0; \lambda_i)} \\ &= \log \frac{N}{df(t_k)} + \log(1 - p(0; \lambda_k)) \end{aligned} \tag{4}$$

ここで，p は，パラメータ λ_k を伴うポアッソン分布である．この手法は，少数の文書のみに出現する単語は，より大きな RIDF スコアを持つ傾向がある．

(d) TF-IDF

TF-IDF 法 (Salton and McGill 1983) は，文書中の単語を重み付けするために，情報検索の研究において広く使われている．TF-IDF は，上述した (a) TF と (b) IDF に基づいて，式 (5) のように定義される．

$$w_{t_k}^{p_i} = \frac{tf(t_k, p_i)}{\sum_{s=1}^{m} tf(t_s, p_i)} \cdot \log \frac{N}{df(t_k)} \tag{5}$$

ここで，$tf(t_k, p_i)$ と $df(t_k)$ は，それぞれ，Web ページ p_i における単語 t_k の頻度と，単語 t_k が出現する Web ページ数を表す．また，N は Web ページの総数である．

(e) x^I-measure

Bookstein and Swanson (Bookstein and Swanson 1974) は，単語 t_k に対する x^I-measure というスコアを導入した．$tf(t_k, p_i)$ を Web ページ p_i における単語 t_k の頻度，$df(t_k)$ を単語 t_k が現れる Web ページ数とすると，$\boldsymbol{w}^{p_i}$ の各要素 $w_{t_k}^{p_i}$ は，式 (6) によって定義される．

$$w_{t_k}^{p_i} = tf(t_k, p_i) - df(t_k) \tag{6}$$

この手法は，同程度の出現頻度である 2 つの単語のうち，少数の文書に集中して出現する単語ほど，高いスコアを示す．

(f) gain

一般に，IDF は単語の重要性を表すと考えられているが，Papineni (Papineni 2001) は，IDF は単語の特徴を表す最適な重みに過ぎず，単語の重要性とは異なるものであるため，利得を単語の重要性と考え，gain を提案した．本手法では，$\boldsymbol{w}^{p_i}$ の各要素 $w_{t_k}^{p_i}$ は，式 (7) によって定義される．

$$w_{t_k}^{p_i} = \frac{df(t_k)}{N}\left(\frac{df(t_k)}{N} - 1 - \log\frac{df(t_k)}{N}\right) \tag{7}$$

ここで，$df(t_k)$ は，単語 t_k が現れる Web ページ数を，N は Web ページの総数を示す．本手法では，ほとんど出現しない単語と，非常に頻出する単語は，両方とも低いスコアとなり，中頻度の単語は高いスコアとなる．

上述した (a)〜(f) の単語重み付け手法の中で，本研究においては，“(f) gain” が最も効果的な単語の重み付け法であることがわかったため，これを本研究における単語の重み付け法として用いる．さらに，クラスタ C の重心ベクトル $\boldsymbol{G}^C$ を式 (8) のように定義する．

$$\boldsymbol{G}^C = (g_{t_1}^C, g_{t_2}^C, \cdots, g_{t_m}^C) \tag{8}$$

ここで，$g_{t_k}^C$ は $\boldsymbol{G}^C$ 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．なお，以下で述べるクラスタリング手法では，2 つのクラスタ C_i，C_j 間の類似度 $sim(C_i, C_j)$ を，式 (9) によって計算する．

$$sim(C_i, C_j) = \frac{\boldsymbol{G}^{C_i} \cdot \boldsymbol{G}^{C_j}}{|\boldsymbol{G}^{C_i}| \cdot |\boldsymbol{G}^{C_j}|} \tag{9}$$

ただし，$\boldsymbol{G}^{C_i}$，$\boldsymbol{G}^{C_j}$ は，それぞれ，クラスタ C_i，C_j の重心ベクトルを表す．

2.1 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 ページを，個々のクラスタとして設定す

る．次に，二つのクラスタ間の類似度が，あらかじめ設定された閾値より小さくなるまで，類似度が最大と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のアルゴリズムを示す．

このアルゴリズムでは，あるクラスタ C_i (要素数 n_i) を最も類似したクラスタ C_j (要素数 n_j) にマージした後の，新たなクラスタ C^{new} の重心ベクトル $\boldsymbol{G}^{new}$ は，式 (10) のように定義される．

$$\boldsymbol{G}^{new} = \frac{\sum_{\boldsymbol{w}^p \in C_i} \boldsymbol{w}^p + \sum_{\boldsymbol{w}^p \in C_j} \boldsymbol{w}^p}{n_i + n_j} \tag{10}$$

2.2 提案する半教師有りクラスタリング

一般に，seed ページを含むクラスタ C_{s_j} と，seed ページを含まないクラスタ C_i の類似度が大きい場合には，両者を新たなクラスタとしてマージすべきであるが，両者の距離が大きい場合には，通常の重心の計算法では，重心の変動が大きくなる傾向にある．そこで，はじめに，あるクラスタ C_i(重心ベクトル $\boldsymbol{G}^{C_i}$) を，seed ページを含むクラスタ C_{s_j}(重心ベクトル $\boldsymbol{G}^{C_{s_j}}$) にマージする際，これらのクラスタの重心間の距離 $D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}})$ に基づいて，Web ページ p の特徴ベクトル $\boldsymbol{w}^p \in C_i$ を重み付けする．次に，この重み付けした特徴ベクトルを用いて重心の計算を行なうことで上述した傾向を防ぎ，重心の変動を抑える．

まず，これまでに k_j 個のクラスタがマージされた seed ページを含むクラスタ $C_{s_j}^{(k_j)}$ (要素数 n_{s_j}) に対して，クラスタ C_i (要素数 n_i) が $k_j + 1$ 回目にマージされるクラスタであるとする．

Algorithm: Agglomerative clustering

Input: Set of feature vectors of search-result Web pages $\boldsymbol{w}^{p_i} (i = 1, 2, \cdots, n)$,

$Wp = \{\boldsymbol{w}^{p_1}, \boldsymbol{w}^{p_2}, \cdots, \boldsymbol{w}^{p_n}\}$.

Output: Set of clusters $\mathcal{C} = \{C_1, C_2, \cdots\}$ that contain the Web pages that refer to the same person.

Method:

1. Set each element in Wp as an initial cluster C_i.

   $C_i = \{\boldsymbol{w}^{p_i}\}$, thus set of clusters $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$.

2. **do**

   2.1 Find the pair of clusters (C_i, C_j) with the maximum similarity,

   then merge C_i and C_j to form a new cluster C^{new}, add C^{new} to $\mathcal{C}$, remove C_i and C_j from $\mathcal{C}$,

   and recompute the centroid $\boldsymbol{G}^{new}$ of the cluster C^{new} using Equation (10).

   2.2 Compute similarity between C^{new} and all $C_i \in \mathcal{C}$ $(C_i \neq C^{new})$.

3. **until** All of the similarities computed in 2.2 are less than the predefined threshold.

4. **return** Set of clusters $\mathcal{C}$

図 1 凝集型クラスタリングアルゴリズム

なお，クラスタ $C_{s_j}^{(0)}$ の要素は，初期の seed ページとなる．

(1) この $C_{s_j}^{(k_j)}$ にマージされるクラスタ C_i に含まれる各要素について，$C_{s_j}^{(k_j)}$ の重心 $\boldsymbol{G}^{C_{s_j}^{(k_j)}}$ と，クラスタ C_i の重心 $\boldsymbol{G}^{C_i}$ 間の距離尺度 $D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}^{(k_j)}})$ を用いて，クラスタ C_i に含まれる Web ページの特徴ベクトル $\boldsymbol{w}_{C_i}^{p_l}$ $(l = 1, \cdots, n_i)$ を重み付けし，その後に生成されるクラスタを $C_{i'}$ (要素数 $n_{i'}$) とする．このとき，$C_{i'}$ の要素となる重み付けした後の Web ページの特徴ベクトル $\boldsymbol{w}_{C_{i'}}^{p_l}$ は，式 (11) で表される．

$$\boldsymbol{w}_{C_{i'}}^{p_l} = \frac{\boldsymbol{w}_{C_i}^{p_l}}{D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}^{(k_j)}}) + c} \tag{11}$$

本研究では，$D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}^{(k_j)}})$ として，(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比較する．また，c は $D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}^{(k_j)}})$ が 0 に非常に近い値となったとき，$\boldsymbol{w}^p$ の各要素が極端に大きな値となることを防ぐために導入した定数である．この c の値の影響については，3.3.1 節で述べる．

(2) 次に，seed ページを含むクラスタ $C_{s_j}^{(k_j)}$ (要素数 n_{s_j}) に $C_{i'}$(要素数 $n_{i'}$) の要素を追加し，クラスタ $C_{s_j}^{(k_j+1)}$ (要素数 $n_{s_j} + n_{i'}$) を作成する．

$$C_{s_j}^{(k_j+1)} = \{\boldsymbol{w}_{C_{s_j}^{(k_j)}}^{p_1}, \cdots, \boldsymbol{w}_{C_{s_j}^{(k_j)}}^{p_{n_{s_j}}}, \boldsymbol{w}_{C_{i'}}^{p_1}, \cdots, \boldsymbol{w}_{C_{i'}}^{p_{n_{i'}}}\}$$

(3) このとき，$k_j + 1$ 回目のクラスタをマージしたクラスタ $C_{s_j}^{(k_j+1)}$ の重心 $\boldsymbol{G}^{C_{s_j}^{(k_j+1)}}$ は，式 (12) のように計算される．ここで，式 (11) において，マージされるクラスタの特徴ベクトル $\boldsymbol{w}_{C_i}^{p_l}$ に重み付けをしているため，重み付き平均の計算となるように，$n_{i'}$ にも同様の重みを乗じている．

$$\boldsymbol{G}^{C_{s_j}^{(k_j+1)}} = \frac{\sum_{\boldsymbol{w}^p \in C_{s_j}^{(k_j+1)}} \boldsymbol{w}^p}{n_{s_j} + n_{i'} \times \frac{1}{D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}^{(k_j)}}) + c}} \tag{12}$$

このように本研究では，seed ページを含むクラスタを重視してクラスタリングの基準を明確にし，正確なクラスタリングを行うことを目的とする．もし，2 つのクラスタが種用例を含まないのであれば，新たなクラスタの重心ベクトル $\boldsymbol{G}^{new}$ は，式 (13) のように計算される．

$$\boldsymbol{G}^{new} = \frac{\sum_{\boldsymbol{w}^p \in C_i} \boldsymbol{w}^p + \sum_{\boldsymbol{w}^p \in C_j} \boldsymbol{w}^p}{n_i + n_j} \tag{13}$$

本研究では，seed ページを含むクラスタに，それと最も類似したクラスタをマージする際，seed ページを含むクラスタの重心の変動を抑える半教師有りクラスタリングを適用して，Web 検索結果における人名の曖昧性を解消する．従来の半教師有りクラスタリングの手法のうち，制約を導入する手法では，クラスタの基準となる重心についての検討は見逃されており，また，距離を学習する手法では，特徴空間が大域的に変換される．一方，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，その重心を局所的に調整できる効果が期待される．なお，seed ページを導入することで，検索結果を改善することは，適合性フィードバック (Rocchio 1971) に類似した手法であると考えられる．しかし，適合性フィードバックでは，検索結果中の文書に対して，ユーザが判断した適合文書・非適合文書に基づいた検索語の修正を目的としているのに対し，本手法は，あらかじめ設定した seed ページに基づいて，検索結果の改善，特に本研究においては，検索結果のクラスタリング精度の改善を目的としている点が異なる．

また，検索結果をクラスタリングする検索エンジンとして，“Clusty”[4] が挙げられる．しかし，そのクラスタリングされた検索結果には，適合しない Web ページが含まれることも多く，クラスタリングを行う上で，何らかの基準が必要である．すなわち，本研究のように，seed ページをクラスタリングの基準として導入し，かつ，その seed ページを含むクラスタの重心を抑えることで，その基準を保つような手法が必要であると考えられる．

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．なお，提案する半教師有りクラスタリングでは，対象とするすべての Web ページが，いずれかの seed ページを含むクラスタにマージされるのではなく，seed ページを含まないクラスタにもマージされることに，注意されたい（図 2 下から 7 行目，“else if” 以降参照）．

ここで，本研究において比較する式 (11) 直後に述べた (i)，(ii)，(iii) の 3 つの距離尺度は，それぞれ，以下のように定義される．

(i) ユークリッド距離

式 (11) において，ユークリッド距離を導入した場合，seed ページを含むクラスタの重心ベクトル $\boldsymbol{G}^{C_s}$ と，あるクラスタ C の重心ベクトル $\boldsymbol{G}^{C}$ 間の距離 $D(\boldsymbol{G}^{C_s}, \boldsymbol{G}^{C})$ は，式 (8) に基づいて，式 (14) のように定義される．

$$D(\boldsymbol{G}^{C_s}, \boldsymbol{G}^{C}) = \sqrt{\sum_{k=1}^{m} (g_{t_k}^{C_s} - g_{t_k}^{C})^2} \tag{14}$$

(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た尺度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，より効果的に抑えられることが期待される．

式 (11) において，マハラノビス距離を導入した場合，seed ページを含むクラスタ C_s の重心ベクトル $\boldsymbol{G}^{C_s}$ と，あるクラスタ C の重心ベクトル $\boldsymbol{G}^{C}$ 間の距離 $D(\boldsymbol{G}^{C_s}, \boldsymbol{G}^{C})$ は，式 (15) のように定義される．

$$D(\boldsymbol{G}_{\boldsymbol{C_{(s)}}}, \boldsymbol{G}_{\boldsymbol{C}}) = \sqrt{(\boldsymbol{G}^{C_s} - \boldsymbol{G}^{C})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{G}^{C_s} - \boldsymbol{G}^{C})} \tag{15}$$

[4] http://clusty.com

Algorithm: Semi-supervised clustering

Input: Set of feature vectors of search-result Web pages $\boldsymbol{w}^{p_i}(i = 1, 2, \cdots, n)$,

and seed pages $\boldsymbol{w}^{p_{s_j}}(j = 1, 2, \cdots, u)$,

$Wp = \{\boldsymbol{w}^{p_1}, \boldsymbol{w}^{p_2}, \cdots, \boldsymbol{w}^{p_n}, \boldsymbol{w}^{p_{s_1}}, \boldsymbol{w}^{p_{s_2}}, \cdots, \boldsymbol{w}^{p_{s_u}}\}$.

Output: Set of clusters $\mathcal{C} = \{C_1, C_2, \cdots\}$ that contain the Web pages that refer to the same person.

Method:

1. Set feature vector of each Web page $\boldsymbol{w}^{p_i}$ and each seed pages $\boldsymbol{w}^{p_{s_j}}$ in W_p

as an initial cluster C_i and $C_{s_j}^{(k_j)}$, respectively.

$C_i = \{\boldsymbol{w}^{p_i}\}, C_{s_j}^{(k_j)} = \{\boldsymbol{w}^{p_{s_j}}\}$,

thus, set of clusters

$\mathcal{C} = \{C_1, C_2, \cdots, C_n, C_{s_1}^{(k_1)}, \cdots, C_{s_u}^{(k_u)}\}$,

where "cannot-link" constraints are introduced between $C_{s_q}^{(k_q)}$ and $C_{s_r}^{(k_r)}(q \neq r)$.

k_h $(h = 1, \cdots, u) \leftarrow 0$,

where k_h denotes the frequency of merging a cluster that do not contain a seed page into $C_{s_h}^{(k_h)}$.

2. **do**

2.1 Find the pair of clusters (C_i, C_j), or $(C_i, C_{s_h}^{(k_h)})$ with the maximum similarity.

if the maximum similarity is obtained in $(C_i, C_{s_h}^{(k_h)})$,

then compute the distance $D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_h}^{(k_h)}})$

between the centroids $\boldsymbol{G}^{C_i}$ and $\boldsymbol{G}^{C_{s_h}^{(k_h)}}$ of clusters C_i and $C_{s_h}^{(k_h)}$, respectively.

for $l = 1$ **to** n_i **do**

transform the feature vector $\boldsymbol{w}_{C_i}^{p_l}$ in C_i to $\boldsymbol{w}_{C_{i'}}^{p_l}$ using Equation (11),

add $\boldsymbol{w}_{C_{i'}}^{p_l}$ to $C_{s_h}^{(k_h)}$

end for

$k_h \leftarrow k_h + 1$

recompute the centroid $\boldsymbol{G}^{C_{s_h}^{(k_h)}}$ using Equation (12), and remove C_i from $\mathcal{C}$.

else if

the maximum similarity is obtained in (C_i, C_j),

then merge C_i and C_j to form a new cluster C^{new}, add C^{new} to $\mathcal{C}$, remove C_i and C_j from $\mathcal{C}$,

and recompute the centroid $\boldsymbol{G}^{new}$ of the cluster C^{new} using Equation (13).

2.2 Compute similarities between C^{new} and all $C_i \in \mathcal{C}$ $(C_i \neq C^{new})$.

3. **until** All of the similarities computed in 2.2 are less than the predefined threshold.

4. **return** Set of clusters $\mathcal{C}$.

図 2　提案する半教師有りクラスタリングアルゴリズム

ここで，$\mathbf{\Sigma}$ は，seed ページを含むクラスタ C_s の要素によって定義される共分散行列である．

すなわち，クラスタ C_s 内の要素を，

$$C_s = \{\boldsymbol{w}_{C_s}^{p_1}, \boldsymbol{w}_{C_s}^{p_2}, \cdots, \boldsymbol{w}_{C_s}^{p_m}\}$$

と表せば，重心ベクトル $\boldsymbol{G}^{C_s}$，

$$\boldsymbol{G}^{C_s} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{w}_{C_s}^{p_i}$$

を用いて，共分散 Σ_{ij} を式 (16) のように定義することができる．

$$\Sigma_{ij} = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}_{C_s}^{p_i} - \boldsymbol{G}^{C_s})(\boldsymbol{w}_{C_s}^{p_j} - \boldsymbol{G}^{C_s})^T \tag{16}$$

以上から，共分散行列 $\mathbf{\Sigma}$ は，

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{mm} \end{bmatrix}$$

と表すことができる．

(iii) 適応的マハラノビス距離

(ii) のマハラノビス距離は，クラスタ内の要素数が少ないときに，共分散が大きくなる傾向がある．そこで，seed ページを含むあるクラスタ C_{s_j} について，このクラスタに含まれる Web ページの特徴ベクトル間の非類似度を局所最小化することを考える．この局所最小化で得られる分散共分散行列を用いて計算した C_{s_j} の重心ベクトル $\boldsymbol{G}^{C_{s_j}}$ と，このクラスタにマージされるクラスタ C_l の重心ベクトル $\boldsymbol{G}^{C_l}$ 間の距離が，適応的マハラノビス距離 (Diday and Govaert 1977) である．この分散共分散行列は，次のように導出される．

(1) まず，クラスタ C_{s_j} において，このクラスタに含まれる Web ページの特徴ベクトル $\boldsymbol{w}^{p_i}$ と，それ以外の特徴ベクトル $\boldsymbol{v}$ ($\boldsymbol{w}^{p_i} \neq \boldsymbol{v}$) との非類似度 $d_{s_j}(\boldsymbol{w}^{p_i}, \boldsymbol{v})$ を，式 (17) により定義する．

$$d_{s_j}(\boldsymbol{w}^{p_i}, \boldsymbol{v}) = (\boldsymbol{w}^{p_i} - \boldsymbol{v})^T \boldsymbol{M}_{s_j}^{-1}(\boldsymbol{w}^{p_i} - \boldsymbol{v}) \tag{17}$$

ただし，$\boldsymbol{M}_{s_j}$ は C_{s_j} の分散共分散行列を表す．すなわち，クラスタ C_{s_j} 内の要素を，

$$C_{s_j} = \{\boldsymbol{w}_{C_{s_j}}^{p_1}, \boldsymbol{w}_{C_{s_j}}^{p_2}, \cdots, \boldsymbol{w}_{C_{s_j}}^{p_m}\}$$

と表せば，重心ベクトル $\boldsymbol{G}^{C_{s_j}}$，

$$\boldsymbol{G}^{C_{s_j}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{w}_{C_{s_j}}^{p_i}$$

を用いて，共分散 M_{ij} を式 (18) のように定義することができる．

$$M_{ij} = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}_{C_{s_j}}^{p_i} - \boldsymbol{G}^{C_{s_j}})(\boldsymbol{w}_{C_{s_j}}^{p_j} - \boldsymbol{G}^{C_{s_j}})^T \tag{18}$$

以上から，共分散行列 $\boldsymbol{M_{s_j}}$ は，

$$\boldsymbol{M_{s_j}} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1m} \\ M_{21} & M_{22} & \cdots & M_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m1} & M_{m2} & \cdots & M_{mm} \end{bmatrix}$$

と表すことができる．

(2) 次に，目的関数

$$\begin{aligned}\Delta_{s_j}(\boldsymbol{v}, \boldsymbol{M}_{s_j}) &= \sum_{\boldsymbol{w}^{p_i} \in C_{s_j}} d_{s_j}(\boldsymbol{w}^{p_i}, \boldsymbol{v}) \\ &= \sum_{\boldsymbol{w}^{p_i} \in C_{s_j}} (\boldsymbol{w}^{p_i} - \boldsymbol{v})^T \boldsymbol{M}_{s_j}^{-1}(\boldsymbol{w}^{p_i} - \boldsymbol{v})\end{aligned}$$

を定義し，これを局所最小化するような C_{s_j} の代表点の特徴ベクトル $\boldsymbol{L}_{s_j}$ と分散共分散行列 $\boldsymbol{S}_{s_j}$ を求める．

(i) まず，クラスタ C_{s_j} の要素により定義される共分散行列 $\boldsymbol{M}_{s_j}$ を固定し，Δ_{s_j} を最小化する $\boldsymbol{L}_{s_j}$ を求める．

$$\boldsymbol{L}_{s_j} = \arg\min_{\boldsymbol{v}} \sum_{\boldsymbol{w}^{p_i} \in C_{s_j}} (\boldsymbol{w}^{p_i} - \boldsymbol{v})^T \boldsymbol{M}_{s_j}^{-1}(\boldsymbol{w}^{p_i} - \boldsymbol{v}) \tag{19}$$

式 (19) において，クラスタ C_{s_j} の重心 G に最も近い点 G' の特徴ベクトルを $\boldsymbol{v}_{G'}$ と表せば，$\boldsymbol{L}_{s_j} = \boldsymbol{v}_{G'}$ と求めることができる．

(ii) 次に，(i) で求めた代表点の特徴ベクトル $\boldsymbol{L}_{s_j} = \boldsymbol{v}_{G'}$ を固定する．ここで，$det(\boldsymbol{M}_{s_j}) = 1$ のもとで，Δ_{s_j} を局所最小化する $\boldsymbol{S}_{s_j}$ を求める．

$$\boldsymbol{S}_{s_j} = \arg\min_{\boldsymbol{M}_{s_j}} \sum_{\boldsymbol{w}^{p_i} \in C_{s_j}} (\boldsymbol{w}^{p_i} - \boldsymbol{v}_{G'})^T \boldsymbol{M}_{s_j}^{-1}(\boldsymbol{w}^{p_i} - \boldsymbol{v}_{G'}) \tag{20}$$

この $\boldsymbol{S}_{s_j}$ は，クラスタ C_{s_j} の共分散行列 $\boldsymbol{M}_{s_j}$ を用いて，式 (21) によって与えられることが，文献 (Diday and Govaert 1977) により示されている．

$$\boldsymbol{S}_{s_j} = (det(\boldsymbol{M}_{s_j}))^{1/m}\boldsymbol{M}_{s_j}^{-1} \tag{21}$$

ただし，$det(\boldsymbol{M}_{s_j}) \neq 0$ であり，m は検索結果集合における単語の異なり数を表す．

以上から，seed ページを含むあるクラスタ C_{s_j} において，Web ページ間の非類似度を局所最小化することを考慮した分散共分散行列 $\boldsymbol{S}_{s_j}$ を求めることができる．この $\boldsymbol{S}_{s_j}$ を用いて，C_{s_j} の重心ベクトル $\boldsymbol{G}^{C_{s_j}}$ と，このクラスタにマージされるべきクラスタ C_l の重心ベクトル $\boldsymbol{G}^{C_l}$ 間の適応的マハラノビス距離は，式 (22) のように定義される．

$$D(\boldsymbol{G}^{C_{s_j}}, \boldsymbol{G}^{C_l}) = \sqrt{(\boldsymbol{G}^{C_{s_j}} - \boldsymbol{G}^{C_l})^T \boldsymbol{S}_{s_j}^{-1}(\boldsymbol{G}^{C_{s_j}} - \boldsymbol{G}^{C_l})} \tag{22}$$

なお，式 (22) は，上述した **(1)**〜**(2)** によるクラスタ C_{s_j} における Web ページ間の非類似度を考慮して得られた式 (21) の分散共分散行列 $\boldsymbol{S}_{s_j}$ を適用している点で，式 (15) とは異なる．

3 実験

3.1 実験データ

本研究では，“Web People Search Task” (Artiles, Gonzalo, and Sekine 2007) において作成された「WePS コーパス」を，実験に用いた．この WePS コーパスは，訓練集合とテスト集合から構成され，それぞれ 49，30，合計で 79 の人名が含まれる．これらは，人名を検索語として，Yahoo![5] の検索 API を通じて得られた上位 100 件の検索結果から取得されたものである．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る．具体的な統計量を表 1 に示す．

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[6] に基づいて，不要語を取り除き，Porter Stemmer (Porter 1980)[7] 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した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．

3.2 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 (Hotho, Nürnberger, and Paaß 2005) 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．以下，生成されたクラスタに割り当てられるべき，人手で定めた正解を「カテゴリ」と呼ぶことにする．“purity” 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

[5] http://www.yahoo.com/
[6] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[7] http://www.tartarus.org/~martin/PorterStemmer/

表 1　WePS コーパスにおける統計量

	同姓同名人物選択の情報源	対象人名数 (A)	(A) における 1 人名あたりの同姓同名人物数
訓練集合	Wikipedia における人物情報記事	7	23.1
	ECDL'06 プログラム委員リスト	10	15.3
	アメリカ合衆国・国勢調査	32	5.90
		合計 49	平均 10.8
テスト集合	Wikipedia における人物情報記事	10	56.5
	ACL'06 参加者リスト	10	31.0
	アメリカ合衆国・国勢調査	10	50.3
		合計 30	平均 45.9

*ECDL: European Conference on Digital Libraries, ACL: Association for Computational Linguistics

を評価対象となるクラスタの集合，L を人手で作成したカテゴリの集合，n をクラスタリング対象の文書数とすると，purity は，式 (23) 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．

$$Purity = \sum_i \frac{|C_i|}{n} \max Precision(C_i, L_j) \tag{23}$$

ここで，あるカテゴリ L_j に対するクラスタ C_i の適合率 $Precision(C_i, L_j)$ は，式 (24) によって定義される．

$$Precision(C_i, L_j) = \frac{|C_i \bigcap L_j|}{|C_i|} \tag{24}$$

"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を多く含むクラスタを高く評価する．inverse purity は，式 (25) によって定義される．

$$InversePurity = \sum_j \frac{|L_j|}{n} \max Recall(C_i, L_j) \tag{25}$$

ここで，あるカテゴリ L_j に対するクラスタ C_i の再現率 $Recall(C_i, L_j)$ は，式 (26) によって定義される．

$$Recall(C_i, L_j) = \frac{|C_i \bigcap L_j|}{|L_j|} \tag{26}$$

また，purity と inverse purity の調和平均 F は，式 (27) によって定義される．

$$F = \frac{1}{\alpha \frac{1}{Purity} + (1-\alpha) \frac{1}{InversePurity}} \tag{27}$$

なお，本研究では，$\alpha = 0.5$，0.2 として，評価を行なった．以下，$\alpha = 0.5$，0.2 のときの F 値を，それぞれ，$F_{0.5}$，$F_{0.2}$ と示すことにする．

3.3 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 ページを用いた実験を行なった．

(a) Wikipedia (Remy 2002) における各人物の記事，

(b) Web 検索結果において上位に順位付けされた Web ページ．

3.3.1 パラメータ c の設定

我々の提案する手法では，seed ページを含むクラスタ C_{s_j} と，それに最も類似したクラスタ C_i をマージした後の新しいクラスタの重心ベクトルは，2 章で述べたように，式 (11) に基づいてクラスタ C_i に含まれる Web ページの特徴ベクトル $\boldsymbol{w}_{C_i}^{p_l}$ $(l = 1, \cdots, n_i)$ を重み付けし，この重み付けした特徴ベクトルを用いて，式 (12) によって計算される．

式 (11) における c は，$D(\boldsymbol{G}^{C_i}, \boldsymbol{G}^{C_{s_j}})$ が 0 に非常に近い値となったとき，$\boldsymbol{w}^p$ の各要素が極端に大きな値となることを防ぐために導入した定数であるが，この値によっては，クラスタリングの精度にも影響が及ぶものと考えられる．そこで，WePS コーパスの訓練集合を用いて，上述した 2 種類の seed ページ (a)，(b) ともに 7 個までの seed ページを用いた場合について，$0.1 \leq c \leq 50$ として得られるクラスタリング精度について検証した．ここで，seed ページの数を 7 個までと定めたのは，少数の seed ページでの効果を確認するためである．この結果，表 2 〜4 に示す c の値のときに，$F_{0.5}$，$F_{0.2}$ ともに，最良なクラスタリング精度が得られた．

なお，以下の 3.3.3 節では，距離尺度，seed ページの種類とその数，に応じて，表 2 〜4 に示した c の値を，WePS コーパスのテスト集合に適用して得られた実験結果を示している．

表 2 ユークリッド距離を用いたときの最良なクラスタリング精度を与える c の値

Seed page	c	$F_{0.5}$	$F_{0.2}$	Seed page	c	$F_{0.5}$	$F_{0.2}$
1 Wikipedia article	0.94	0.63	0.61	1 Web page	0.95	0.61	0.58
2 Wikipedia articles	0.93	0.65	0.64	2 Web pages	0.94	0.63	0.61
3 Wikipedia articles	0.95	0.66	0.66	3 Web pages	0.96	0.64	0.65
4 Wikipedia articles	0.97	0.67	0.68	4 Web pages	0.93	0.64	0.66
5 Wikipedia articles	0.98	0.68	0.69	5 Web pages	0.97	0.66	0.67
6 Wikipedia articles	0.96	0.67	0.70	6 Web pages	0.94	0.64	0.67
7 Wikipedia articles	0.97	0.66	0.68	7 Web pages	0.96	0.62	0.65

表 3　マハラノビス距離を用いたときの最良なクラスタリング精度を与える c の値

Seed page	c	$F_{0.5}$	$F_{0.2}$	Seed page	c	$F_{0.5}$	$F_{0.2}$
1 Wikipedia article	0.95	0.64	0.61	1 Web page	0.93	0.62	0.59
2 Wikipedia articles	0.94	0.66	0.63	2 Web pages	0.96	0.63	0.61
3 Wikipedia articles	0.96	0.68	0.66	3 Web pages	0.95	0.66	0.64
4 Wikipedia articles	0.98	0.70	0.68	4 Web pages	0.95	0.67	0.68
5 Wikipedia articles	0.97	0.71	0.70	5 Web pages	0.94	0.69	0.70
6 Wikipedia articles	0.95	0.69	0.71	6 Web pages	0.93	0.68	0.69
7 Wikipedia articles	0.96	0.67	0.70	7 Web pages	0.95	0.66	0.68

表 4　適応的マハラノビス距離を用いたときの最良なクラスタリング精度を与える c の値

Seed page	c	$F_{0.5}$	$F_{0.2}$	Seed page	c	$F_{0.5}$	$F_{0.2}$
1 Wikipedia article	0.97	0.66	0.67	1 Web page	0.96	0.65	0.63
2 Wikipedia articles	0.95	0.68	0.68	2 Web pages	0.98	0.66	0.65
3 Wikipedia articles	0.94	0.70	0.71	3 Web pages	0.93	0.68	0.69
4 Wikipedia articles	0.96	0.72	0.73	4 Web pages	0.95	0.70	0.72
5 Wikipedia articles	0.98	0.75	0.74	5 Web pages	0.98	0.73	0.73
6 Wikipedia articles	0.97	0.73	0.72	6 Web pages	0.96	0.71	0.72
7 Wikipedia articles	0.95	0.71	0.70	7 Web pages	0.97	0.69	0.71

表 5　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

Purity	Inverse purity	$F_{0.5}$	$F_{0.2}$
0.67	0.48	0.52	0.49

3.3.2　文書全体を用いた実験結果

(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 5 に示す．

(2) 半教師有りクラスタリングを用いた実験結果

seed ページを導入することによる効果を確かめるため，はじめに一つの seed ページを用いて実験を行なった．この際，3.3 節はじめに述べた 2 種類の seed ページに関して，(a) は検索結果の上位にある Wikipedia の記事を，(b) は第 1 位に順位付けされた Web ページを用いた．しかしながら，3.1 節で述べた WePS コーパスのテスト集合におけるすべての人名が，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，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 検索結果において第 1 位に順位付けされた Web ページを用いた．この方針に基づき，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 の人名に対しては第 1 位に順位付けされた Web ページを seed ページとして用いた．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 (Bunescu and Pasca 2006) らは，Wikipedia の構造を用いることによって固有名を同定するとともに，その固有名の曖昧性を解消している．表 6 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．

さらに，一つの seed ページを用いた実験において，最も良い F 値 ($F_{0.5} = 0.68$，$F_{0.2} = 0.66$) 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．3.3.1 節でも述べたように，少数の seed ページでの効果を確認するために，導入する seed ページの数は 7 個までとした．また，図 2 に示したように，これらの seed ページの間には，“cannot-link” の制約を導入している．これは，上位に順位付けされる検索エンジンの出力結果を信頼し，それぞれの Web ページが異なる人物について記述していると想定していることに基づく．図 3，4 は，それぞれ, 複数の Wikipedia 記事，上位 7 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 (F 値) を示す．

また，この実験では，2 章で述べたように，seed ページを含むクラスタの重心と，それにマージされるクラスタの重心間の距離を考慮する．この提案手法の有効性を確認するために，1 章で述べた距離を学習する半教師有りクラスタリング手法である Klein ら (Klein et al. 2002)，Xing ら (Xing et al. 2003)，Bar-Hillel ら (Bar-Hillel et al. 2003) の手法を用いて得られた結果との比較を示す．また，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る効果を確認するために，重心を固定する手法との比較も示す．

3.3.3 文書を部分的に用いた実験結果

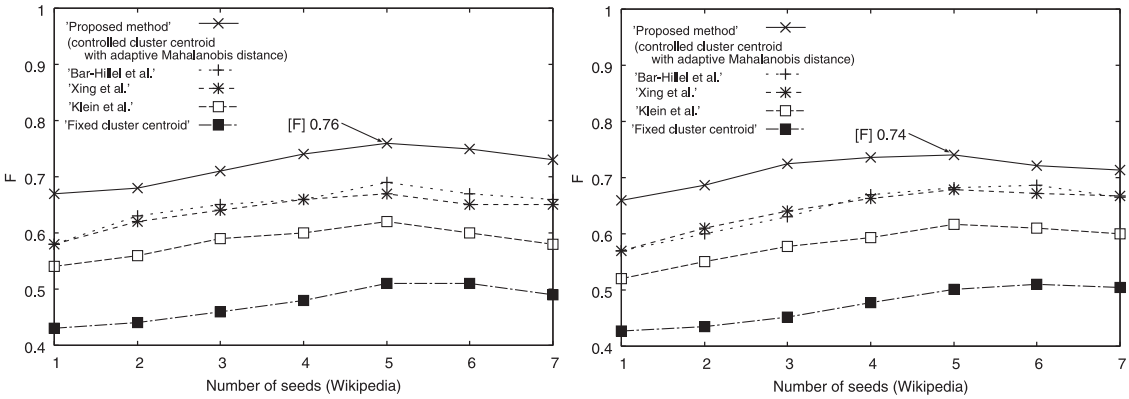
3.3.2 節で述べた実験では，検索結果の Web ページと seed ページの全文を用いた．しかし，人物について記述された Web ページにおいて，その人物を特徴付ける単語は，人名の周囲にしばしば現れること，また，検索結果のスニペットにおいても，同様の傾向が観察される．

そこで，seed ページを用いて最も良い結果が得られている場合，すなわち，図 3 において，5

表 6 1 つの seed ページを使い，提案する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

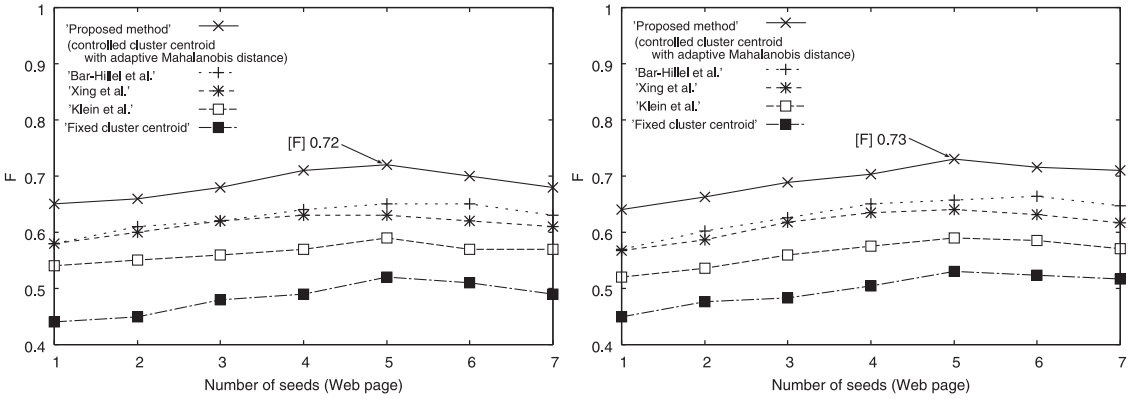
Distance measure	Seed page*	Purity	Inverse purity	$F_{0.5}$	$F_{0.2}$
(i) Euclidean distance	(a)	0.47	0.82	0.60	0.62
	(b)	0.49	0.77	0.58	0.61
(ii) Mahalanobis distance	(a)	0.50	0.85	0.62	0.64
	(b)	0.53	0.75	0.64	0.65
(iii) Adapative Mahalanobis distance	(a)	0.57	0.88	**0.68**	**0.66**
	(b)	0.55	0.76	0.65	0.64

*(a) and (b) in “Seed page” denote “Wikipedia article” and “top-ranked Web page,” respectively.



$[\alpha = 0.5]$　　$[\alpha = 0.2]$

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 (7 つまでの Wikipedia 記事)



$[\alpha = 0.5]$　　$[\alpha = 0.2]$

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 7 位までに順位付けされた Web ページ)

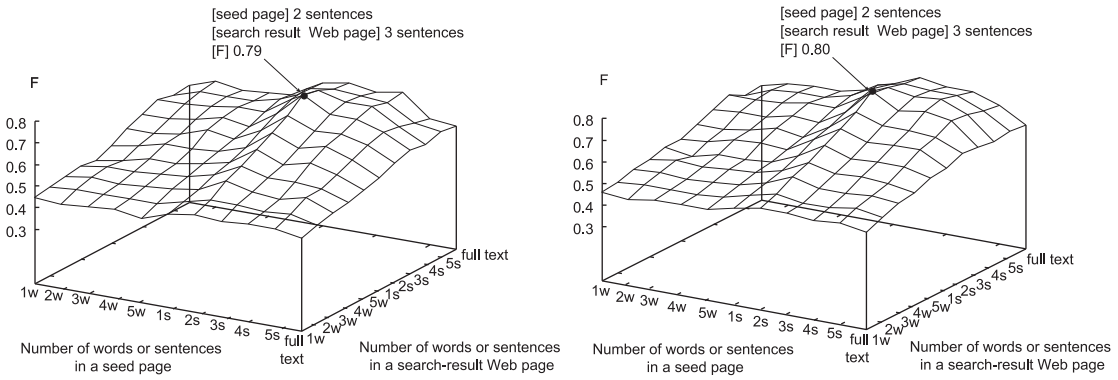
つの Wikipedia 記事を用いた場合 ($F_{0.5} = 0.76, F_{0.2} = 0.74$) に，さらに精度が改善されるかを確認するために，

(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，

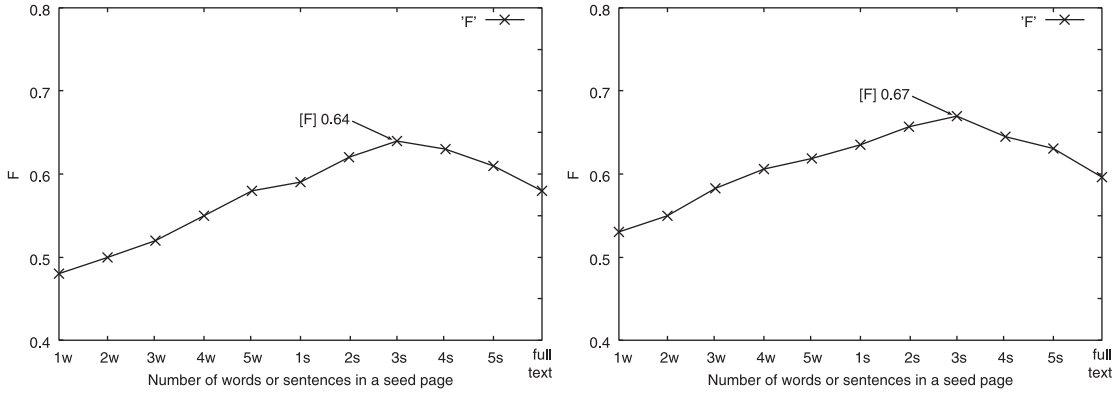
実験を行なった．

(i) については，まず，WePS コーパスの訓練集合を用いて，最も良い F 値を与える seed ページと検索結果の Web ページのそれぞれにおいて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，これらのパラメータをテスト集合に適用し，評価する．(ii)



$[\alpha = 0.5]$　　$[\alpha = 0.2]$

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)



$[\alpha = 0.5]$　　$[\alpha = 0.2]$

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)

についても同様に，WePS コーパスの訓練集合を用いて，最も良い 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，これらのパラメータをテスト集合に適用し，評価する．最終的に (i)，(ii) の実験によって得られたクラスタリング精度を，表 7 に示す．

3.3.4　他手法との比較

“Web People Search Task” における上位 3 チームのクラスタリング精度 (F 値) を，表 7 に示す．なお，これらのチームで採用している手法の詳細については，表 7 に示した文献を参照さ

表 7　Web People Search Task における上位 3 チームと提案手法とのクラスタリング精度の比較

Team-ID	Purity	Inverse purity	$F_{0.5}$	$F_{0.2}$
CU_COMSEM (Chen and Martin 2007)	0.72	0.88	**0.78**	**0.83**
IRST-BP (Popescu and Magnini 2007)	0.75	0.80	0.75	0.77
PSNUS (Elmacioglu, Tan, Yan, Kan, and Lee 2007)	0.73	0.82	0.75	0.78
Our proposed method (with adaptive Mahalanobis distance)				
Using full text (Sec. 3.3.2)				
1 Wikipedia article	0.58	0.84	0.68	0.66
1 Web page	0.60	0.76	0.65	0.64
5 Wikipedia articles	0.75	0.78	0.76	0.74
5 Web pages	0.77	0.69	0.72	0.73
Using fragments (Sec. 3.3.3)				
(i) 2 and 3 sentences in 5 Wikipedia seed pages and a search result Web page, respectively	0.80	0.83	**0.81**	**0.82**
(ii) Snippet and 3 sentences in 5 Wikipedia seed pages	0.70	0.62	0.66	0.68

れたい．基本的には，凝集型クラスタリングの手法が採用されている．また，提案手法によって得られた結果も，比較のために示す．

3.4　処理時間に関する検討

3.3.2 節で述べたように，式 (11) において，適応的マハラノビス距離を用いて，seed ページを含むクラスタにマージされるクラスタに含まれる Web ページの特徴ベクトルを重み付けし，この変換された特徴ベクトルを用いて重心の計算を行なった場合に，最良なクラスタリング精度が得られることがわかった．この場合について，7 つまでの Wikipedia 記事，上位 7 位までに順位付けされた Web ページを seed ページとして用い，最も処理時間を要すると考えられる 3.3.2 節の文書全体を用いた場合についての処理時間を測定した．なお，提案手法は，PC (CPU: Intel Pentium M・2.0 GHz，Memory: 2 GByte，OS: Windows XP) 上に Perl を用いて実装されている．図 7 に，その結果を示す．

3.5　考察

式 (11) における c の値について，特徴ベクトルを重み付けする際には，表 2〜4 から $c = 0.95$ 前後の値を用いたときに，最良なクラスタリング精度が得られることがわかった．なお，$5 \leq c \leq 50$ の大きな値のときには，それほど高いクラスタリング精度が得られないことも観察された．これは，式 (11) において，距離尺度よりも c が支配的になることにより，クラスタにマージすべき Web ページの特徴ベクトルの各要素の値が小さくなりすぎることによる影響であると考えられる．

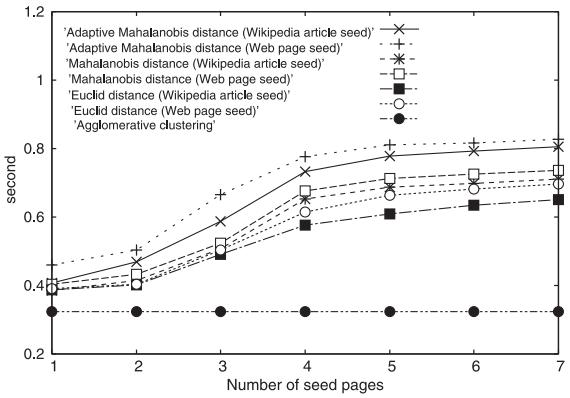


図 7　seed ページ数を変化させたときのクラスタリングに要する処理時間

凝集型クラスタリングの手法においては，表 5 から，purity (0.67) は，inverse purity (0.48) よりも高いことがわかる．このように，purity が高いことは，凝集型クラスタリングが，一つの要素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．また，F 値が $F_{0.5}$=0.52，$F_{0.2}$=0.49 であり，それほど高い精度が得られていないことは，凝集型クラスタリングでは，クラスタリングを適切に行なうことが難しいことを改めて確認できたといえる．

2 章で述べた半教師有りクラスタリングの手法において，表 6 から purity の値 (0.47〜0.57) は，表 5 の凝集型クラスタリングを用いて得られた purity の値 (0.67) を上回ることができなかったが，inverse purity の値 (0.75〜0.88) は，すべての手法が凝集型クラスタリングの値 (0.48) を上回っていることがわかる．また，良好な inverse purity の値によって，F 値においても，良い結果が得られている．これは，seed ページを導入したこと，ならびに，その seed ページ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．さらに，表 6 からは，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，最も良い F 値 ($F_{0.5} = 0.68$，$F_{0.2} = 0.66$) が得られたことがわかる．

複数の seed ページを用いた半教師有りクラスタリング手法においては，図 3，4 から，次の内容が観察される．まず，いずれの seed ページを用いても，また，いずれの手法においても，導入する種文書数の増加とともに，クラスタリング精度 (F 値) が改善されている．seed ページの数について，7 個まで導入したが，いずれの seed ページとも 5 個の時点でのクラスタリング精度が最も良いことが観察される．さらに，重心を固定する方法は，他の手法に比べて非常に精度が劣る結果となった．これは，重心を完全に固定してしまうと，その重心と類似度が高い Web ページしかマージされなくなるため，本来クラスタにマージされるべき Web ページが独立したクラスタとなってしまうことが原因であると考えられる．この実験においては，高い purity の値が得られていたことからも，上述した原因が裏付けられるといえる．

一方，距離を学習するクラスタリング手法では，Bar-Hillel ら (Bar-Hillel et al. 2003)，Xing ら (Xing et al. 2003)，Klein ら (Klein et al. 2002) の手法の順に良いクラスタリング精度が得られている．1 章で述べたように，Klein らの手法では，類似した 2 点 (x_i, x_j) 間を 0，類似していない 2 点間を $(\max_{i,j} D_{ij})$+1 と設定した単純な隣接行列を作成した上で，クラスタリングを行なうのに対し，Xing ら，Bar-Hillel らの方法では，特徴空間を適切に変換する手法が用いられている．後者の二つの手法では，この変換手法が有効に作用しているものと考えられる．しかし，これらの距離を学習する手法と比較しても，重心の変動を抑えたクラスタリングを行なう我々の提案手法が，最も良いクラスタリング精度を示した．これは，あるクラスタを seed ページを含むクラスタにマージするたびに，その seed ページを含むクラスタの重心を局所的に調整できることによる効果であると考えられる．

さらに，seed ページについては，Wikipedia における各人物の記事を用いたほうが，Web 検索結果の上位に順位付けされた Web ページを用いるよりも良い精度が得られた．これは，クラスタリングのための seed ページとして，Wikipedia の記述内容を用いることが有効であることを示す事例であると考えられる．

また，文書を部分的に用いた場合には，以下に述べるような傾向が観察される．まず，WePS コーパスの訓練集合において，3.3.3 節 (i) で述べたように，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，人名前後の数語のみでは，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，人物の実体を識別しやすくなったことによる効果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も良い F 値 ($F_{0.5} = 0.79$, $F_{0.2} = 0.80$) 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:0.80，inverse purity:0.83, $F_{0.5} = 0.81$，$F_{0.2} = 0.82$] の結果が得られた．特に F 値は，$\alpha = 0.5$ のとき，表 7 に示した “Web People Search Task” (Artiles et al. 2007) の第 1 位のチーム (CU_COMSEM) の結果を 0.03 上回り，提案手法が有効であることが確認される．なお，3.3.2 節 (2) で述べたように，Wikipedia に記事のある 16 人名のうち，Wikipedia から取得した人名数は 10 (表 1 参照，以下 (A) とする)，ACL’06 参加者リスト，アメリカ合衆国・国勢調査の人名のうち，Wikipedia にも記事のある人名数は 6（表 1 参照，以下 (B) とする）である．これらの人名について，Wikipedia を seed ページとしてクラスタリングした場合に，その精度に差があるか否かを検証した．その結果を表 8 に示す．(A) の方が (B) よりも，0.02〜0.04 上回る結果が得られているが，それほど大きな差ではない．このことから，seed ページとして Wikipedia の記述内容を用いることは，(B) のように他分野から取得した人物の Web ページに対しても有効であり，Wikipedia の記述内容の汎用性が特徴付けられる結果であると考えられる．

また，クラスタ数については，seed ページを導入したことで，この seed ページを中心に，Web ページのグループが形成され，実際の正解クラスタ数よりも少ない数のクラスタが生成される傾向が観察された．これは，表 7 において，inverse purity の値が高いことからも裏付けられる．

なお，“Web People Search Task” の上位 3 チームは，凝集型クラスタリングの手法を採用しているが，これらの手法は素性を工夫することで，比較的高い精度を得ている．一方，我々の提案する半教師有りクラスタリングでは，seed ページを含むクラスタの重心の変動を抑えることで，表 5 に示した凝集型クラスタリングよりも精度が改善されている．我々が導入した素性は，2 章で述べたように，gain によって単語を重み付けする簡単なものであるが，“Web People Search Task” の上位 3 チームが使用した素性を我々の手法に適用すれば，さらなる精度の向上が期待される．そこで，これらの 3 チームの素性を，我々の手法で用いた結果を表 9 に示す．なお，表 7 に示した我々の提案手法で得られた最良の結果と比較するため，seed ページとして Wikipedia における各人物の記事を 5 つ導入した場合についての比較を行った．まず，CU_COMSEM について，表 7 に示した凝集型クラスタリングの F 値 ($F_{0.5} = 0.78$，$F_{0.2} = 0.83$) と比較して，半教師有りクラスタリングの F 値も高め ($F_{0.5} = 0.81$，$F_{0.2} = 0.84$) となっている．しかし，$F_{0.5}$ で 0.03，$F_{0.2}$ で 0.01 程度の改善に過ぎない．これは，文中の単語，URL のトークン，名詞句など，すでに多くの素性を導入しているため，半教師有りクラスタリングを適用しても，それほど効果は得られないことによると考えられる．IRST-BP については，表 7 に示した凝集型クラスタリングの F 値 ($F_{0.5} = 0.75$，$F_{0.2} = 0.77$) と比較しても，半教師有りクラスタリングの精度は ($F_{0.5} = 0.76$，$F_{0.2} = 0.81$) であり，改善の程度は $F_{0.5}$ で 0.01，$F_{0.2}$ で 0.04 であった．このチームが使用している固有名詞，時制表現，人名のある段落で最も良く出現する単語といった素性は，あまり有効

表 8 Wikipedia を seed ページとした場合，(A) Wikipedia から取得した 10 人名と，(B) Wikipedia に記事はあるが，ACL’06 参加者リスト，アメリカ合衆国・国勢調査から取得した 6 人名のクラスタリング精度の比較

	文書全体を用いた場合		文書を部分的に用いた場合	
	$F_{0.5}$	$F_{0.2}$	$F_{0.5}$	$F_{0.2}$
(A)	0.65	0.75	0.72	0.77
(B)	0.63	0.71	0.69	0.74

表 9 “Web People Search Task” の上位 3 チームが使用した素性を，提案する半教師有りクラスタリングに適用して得られたクラスタリング精度

Team-ID	Purity	Inverse purity	$F_{0.5}$	$F_{0.2}$
CU_COMSEM	0.70	0.92	0.81	0.84
IRST-BP	0.69	0.82	0.76	0.81
PSNUS	0.67	0.84	0.78	0.82

な素性ではないと考えられる．PSNUS については，NE 素性を TF-IDF で重み付けしたのみの単純な素性であるが，表 7 に示した凝集型クラスタリングの F 値 ($F_{0.5} = 0.75$，$F_{0.2} = 0.78$) と比較して，半教師有りクラスタリングで得られた F 値は $F_{0.5} = 0.78$，$F_{0.2} = 0.82$ であり，$F_{0.5}$ で 0.03，$F_{0.2}$ で 0.04 の改善が観察される．一方，我々の手法では素性として gain を用い，表 7 に示したとおり，$F_{0.5} = 0.81$，$F_{0.2} = 0.82$ の F 値を得ている．これは，CU_COMSEM で使用されている多数の素性で得られた F 値とほぼ同じ値が得られていることから，gain によって単純に Web ページ中の単語を重み付けした素性だけでも，我々の提案する半教師有りクラスタリングを適用することで，高い精度が得られることが確認された．また，表 5 に示した凝集型クラスタリングによる F 値 ($F_{0.5} = 0.52$，$F_{0.2} = 0.49$) と比較しても，$F_{0.5}$ で 0.29，$F_{0.2}$ で 0.33 の改善が観察されたことから，我々の提案する半教師有りクラスタリングの有効性が確認される．

次に，WePS コーパスの訓練集合において，3.3.3 節 (ii) で述べたように，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または文数を変化させた場合，図 6 から，seed ページ中の人名前後の単語ではなく，同様に文を用いたときに，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．この場合も同様に，人名前後の数語の情報よりも，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効果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 3 文を用いた場合に最も良い F 値 ($F_{0.5} = 0.64$，$F_{0.2} = 0.67$) 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:0.70，inverse purity:0.62，$F_{0.5} = 0.66$，$F_{0.2} = 0.68$] の結果が得られた．この結果は，Web People Search Task の上位 3 チームの結果，および本研究における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，seed ページで人名前後の 3 文という情報を用いたとしても，該当する人物について述べた適切な Web ページが，その seed ページには集まらず，結果として，クラスタリング精度が悪くなったことによるためであると考えられる．

以上から，提案手法では Wikipedia の記事を seed ページとして利用し，人名前後の 2 文を，また，検索結果の Web ページについては人名前後の 3 文を用いた場合に，良好な検索結果が得られることがわかった．

さらに，処理時間に関して，最良なクラスタリング精度が得られた適応的マハラノビス距離の式 (22) における分散共分散行列の計算には，単語数の 2 乗の計算量が必要となるが，1 人名について 100 件の Web ページのクラスタリングを行なうのに，最も多い 5 つの seed ページを用い，seed ページと検索結果の Web ページの双方ともに文書全体を用いた場合でも，0.8 秒余りで処理できることが図 7 から観察され，妥当な応答性を実現できていると考えられる．

4 むすび

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，seed ページを含むクラスタの重心の変動を抑える半教師有りクラスタリングの手法を提案した．実験の結果，最良な場合において，[purity:0.80，inverse purity:0.83，$F_{0.5}$:0.81，$F_{0.2}$:0.82] の評価値が得られた．今回は，上位に順位付けされる検索エンジンの出力結果が異なることを想定して実験を行った．すなわち，同一人物の seed ページ間にも “cannot-link” の制約が導入されている可能性がある．しかし，クラスタが生成される過程で，seed ページ以外の人物のページがクラスタ内の要素として支配的になり，最終的には比較的正確なクラスタが生成されることが観察された．同一人物の seed ページ間でも，その人物を正確に表現しているページ，そうでないページがあることによるためであると考えられる．したがって，その人物についてより正確に記述された Web ページを seed ページとして選択することが，今後の課題の一つとして挙げられる．また，Web 検索結果における人名の曖昧性解消の精度を高めるには，その人物を特徴付ける単語の重みが大きくなるように，Web ページの特徴ベクトルを作成して，クラスタリングを行なうことが重要である．そのために，特に，seed ページの内容に適合する人物のページが集まるように，より的確な seed ページの特徴ベクトルを作成するための手法を開発してクラスタリングを行なうことも，今後の課題として挙げられる．

参考文献

Artiles, J., Gonzalo, J., and Sekine, S. (2007). “The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task.” In *Proc. of the Semeval 2007, Association for Computational Linguistics (ACL)*, pp. 64–69.

Bar-Hillel, A., Hertz, T., and Shental, N. (2003). “Learning Distance Functions Using Equivalence Relations.” In *Proc. of the 20th International Conference on Machine Learning (ICML 2003)*, pp. 577–584.

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

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

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

Bookstein, A. and Swanson, D. R. (1974). “Probabilistic Models for Automatic Indexing.” *Journal of the American Society for Information Science*, **25** (5), pp. 312–318.

Bunescu, R. and Pasca, M. (2006). “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.

Chen, Y. and Martin, J. (2007). “CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation.” In *Proc. of the Semeval 2007, Association for Computational Linguistics (ACL)*, pp. 125–128.

Church, K. W. and Gale, W. A. (1995a). “Inverse Document Frequency (IDF): A Measure of Deviation from Poisson.” In *Proc. of the 3rd Workshop on Very Large Corpora*, pp. 121–130.

Church, K. W. and Gale, W. A. (1995b). “Poisson Mixtures.” *Journal of Natural Language Engineering*, **1** (2), pp. 163–190.

Diday, E. and Govaert, G. (1977). “Classification Automatique Avec Distances Adaptatives.” *R.A.I.R.O. Informatique Computer Science*, **11** (4), pp. 329–349.

Elmacioglu, E., Tan, Y. F., Yan, S., Kan, M.-Y., and Lee, D. (2007). “PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features.” In *Proc. of the Semeval 2007, Association for Computational Linguistics (ACL)*, pp. 268–271.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). “A Brief Survey of Text Mining.” *GLDV-Journal for Computational Linguistics and Language Technology*, **20** (1), pp. 19–62.

Jones, K. S. (1973). “Index Term Weighting.” *Information Strage and Retrieval*, **9** (11), pp. 619–633.

Klein, D., Kamvar, S. D., and Manning, C. D. (2002). “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.

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

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

Papineni, K. (2001). “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.

Pedersen, T., Purandare, A., and Kulkarni, A. (2005). “Name Discrimination by Clustering Similar Contexts.” In *Proc. of the 6th International Conference on Computational Linguistics*

and Intelligent Text Processing (CICLing 2005), pp. 226–237.

Popescu, O. and Magnini, B. (2007). "IRST-BP: Web People Search Using Name Entities." In *Proc. of the Semeval 2007, Association for Computational Linguistics (ACL)*, pp. 195–198.

Porter, M. F. (1980). "An Algorithm for Suffix Stripping." *Program*, **14** (3), pp. 130–137.

Remy, M. (2002). "Wikipedia: The Free Encyclopedia." *Online Information Review*, **26** (6), p. 434.

Rocchio, J. (1971). "Relevance Feedback in Information Retrieval." In Salton, G. (Ed.), *The Smart Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall, Englewood Cliffs, NJ.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Shental, N., Hertz, T., Weinshall, D., and Pavel, M. (2002). "Adjustment Learning and Relevant Component Analysis." In *Proc. of the 7th European Conference on Computer Vision (ECCV 2002)*, pp. 776–792.

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

Wagstaff, K., Rogers, S., and Schroedl, S. (2001). "Constrained K-means Clustering with Background Knowledge." In *Proc. of the 18th International Conference on Machine Learning (ICML 2001)*, pp. 577–584.

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

略歴

杉山　一成：1998 年横浜国立大学工学部電子情報工学科卒業．2000 年同大学院工学研究科電子情報工学専攻博士前期課程修了．KDDI（株）勤務を経て，2004 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了，博士（工学）．（株）日立製作所勤務を経て，2006 年東京工業大学精密工学研究所研究員，2009 年シンガポール国立大学計算機科学科研究員，現在に至る．情報検索，自然言語処理に関する研究に従事．電子情報通信学会，情報処理学会，人工知能学会，IEEE，ACM，AAAI 各会員．

奥村　　学：1984 年東京工業大学工学部情報工学科卒業．1989 年同大学院博士課程修了．工学博士．同年，東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，2009 年同教授，現在に至る．自然言語処理，知的情報提

示技術，語学学習支援，テキスト評価分析，テキストマイニングに関する研究に従事．電子情報通信学会，情報処理学会，人工知能学会，言語処理学会，認知科学会，計量国語学会，AAAI，ACL 各会員．

（2009 年 2 月 6 日　受付）
（2009 年 7 月 16 日　再受付）
（2009 年 8 月 19 日　採録）