クエリー拡張による特徴量抽出を用いたWeb検索における同姓同名問題解消

池田　雅紀† 　小野　真吾† 　佐藤　一誠† 　吉田　稔†† 　中川　裕志††

† 東京大学大学院情報理工学系研究科〒 113-0033 東京都文京区本郷 7-3-1
†† 東京大学情報基盤センター
E-mail: †{ikeda,ono,sato,mino}@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

あらまし　Web 検索における人名検索が重要になるにつれて，複数の同姓同名人物の存在による閲覧性の低下が問題となってきている．この解決策として，検索結果に対して人物ごとのクラスタを作成し，表示する方法が提案されている．本研究では，複数の特徴量を用いて文書の類似度を計算し，階層併合クラスタリングを用い，そのための特徴量の抽出方法や類似度計算の検討を行った．また，階層併合クラスタリング結果を元に有効な特徴量を抽出し，抽出した特徴量を用いて二段階目のクラスタリングを行った．二段階クラスタリングによって，1 文書中において複数の同姓同名の人物が扱われる問題についても対応した．提案手法は WePS-1, WePS-2 のデータセットを用いて評価を行った．
キーワード　同姓同名問題解消, クラスタリング, 固有表現

Person Name Disambiguation on the Web Using Query Expansion

Masaki IKEDA†, Shingo ONO†, Issei SATO†, Minoru YOSHIDA††, and Hiroshi NAKAGAWA††

† Graduate School of Information Science and Technology, The University of Tokyo. Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan
†† Information Technology Center, The University of Tokyo
E-mail: †{ikeda,ono,sato,mino}@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

Abstract　The more important the web search become, the bigger the same name problem in the web search. Proposed solution is forming clusters of people from search results. In this paper, we report our algorithms that disambiguates person names in web search results. Our clustering algorithm is based on hierarchical agglomerative clustering using named entities, compound key words and URLs as features for document similarity calculation. We considered extraction of the features and calculation of document similarity. Additionally, we extracted useful features from the clustering result and did second-stage clustering by the features. Two-stage clustering adapt document which several person of same name. We evaluated the proposed method with WePS-1 and WePS-2 data sets.
Key words　Person name disambiguation, Clustering, Named Entity

1. はじめに

Web 上の人物検索は Web 検索において重要な地位を占めてきている．このような状況の中，人物の検索に関する問題として人物の同姓同名問題の解消が求められている．人物の同姓同名問題とは，Web 検索において検索対象者と同姓同名の人物の存在によって検索結果から目的の人物のページを発見することが困難になるという問題である．特に困難な場合としては以下の場合が考えられる．第一に，検索対象者と同姓同名の有名人が存在する場合である．例えば，米国前大統領の “George Bush” と同姓同名の別人物を検索する場合，大統領である “George Bush” に関するページが検索結果に多く現れ，目的とするページを探すのが困難になる．第二に，検索対象者の名前が多くの同姓同名の人物を持つ場合である．例えば，“田中太郎”，“John Smith” という名前を持つ人々は非常に多い．このように，同姓同名問題は言語を問わず問題となっている．実際の問題を考えると，有名人の影響の大きさや人数によるデータのばらつき，1 文書内における複数の同姓同名の人物の存在などが混在して問題を複雑にしている．

この問題の解決方法として提案されているのが，検索結果の人名ごとのクラスタリングである．即ち，検索結果を同一人物ごとのクラスタにまとめて提示し，検索結果の閲覧性を向上さ

せることで同姓同名の存在による効率の低下を防ぐという方法である．同姓同名の人物のクラスタリングには文書中の人物に関わる名詞句を用いることが有効であるとされている．特に，人名，地名，組織名といった固有表現がクラスタリングにおいて有効であると先行研究 [1] によって示されている．我々の先行研究 [2] では，固有表現以外の特徴量として，文書中から検索クエリーの前後の文字列を取り出し，その部分から複合名詞を抽出した．抽出した複合名詞に対して，重要度を計算し，重要度が一定以上の複合語を抜き出し，重要語とした．そして，これらの各特徴量についてクラスタリングを行った後，得られたクラスタの和集合を取ることによって同姓同名の人物のクラスタを作成する方法を用いた．

本研究では，従来研究における特徴量抽出と類似度計算方法について見直しを行った．第一に，特徴量抽出範囲の検討を行った．従来研究では検索対象の人物名の周辺の文から固有名詞，重要語の抽出を行っていたが，範囲を実験的に検討した結果，文書全体を対象とした抽出を行った方が性能が良いとの結果を得た．第二に，文書間の類似度計算について見直した．従来研究では重要語の類似度計算方法として，重要語の重要度を用いた文書ベクトルによる cos 類似度を用いていたが，本研究では Overlap 係数を用いて類似度計算を行った．同時に，複数種類の特徴量に基づいてクラスタリングを行う方法として，各特徴量によるクラスタリングの結果を併合する方法から 各特徴量の類似度を組み合わせて計算した文書間の類似度を用いて階層併合クラスタリングを行う方法へと変更した．

さらに，クラスタリングに有効な特徴量を抽出する手段として，上記のクラスタリングの結果を利用して，人物に関連していると考えられる重要語を抽出し，二段階目のクラスタリングを行った．この手法は各クラスタにおいて情報検索におけるクエリー拡張を行うことを基本として構成されている．二段階目のクラスタリングにおける目的は次の二つである．第一に，第一段階のクラスタリングで分離している同一人物のクラスタをまとめることである．第二に，複数の同姓同名の人物が 1 文書中で扱われている場合への対応である．例えば，Wikipedia の曖昧性解消のページ(注1)や別の検索エンジンによる検索結果などがこの場合に当たる．第一段階で行ったクラスタリングは 1 文書中で複数の同姓同名の人物が含まれている場合を考慮していない．第二段階では，再クラスタリングを行う際に，複数の同姓同名の人物のクラスタから類似している文書については複数のクラスタに属するようにするソフトクラスタリングを行い，このような場合への対応を行った．

最後に，本稿での提案手法をまとめる．本稿では，同姓同名人物のクラスタリングに対して，以下の点を改良した．

（1） 特徴量抽出範囲
（2） 類似度計算への Overlap 係数
（3） 複数の特徴量に基づく類似度計算
（4） 階層併合クラスタリング手法
（5） 二段階クラスタリングによるソフトクラスタリング

本稿の構成は，以下のようになっている．第 2 節で先行研究を紹介する．第 3 節では特徴量抽出について説明する．第 4 節では階層併合クラスタリングとクラスタリングにおける類似度計算について説明する．第 5 節では二段階クラスタリングの処理について説明する．第 6 節では実験により本研究の手法を検証した結果について説明する．第 7 節で本稿の結論を述べる．各改良点の説明は，特徴量抽出に関する（1）は第 3 節で説明する．階層併合クラスタリングにおける改良点である（2）〜（4）は第 4 節で説明する．二段階クラスタリングの手法である（5）は第 5 節で説明する．

2. 先 行 研 究

先行研究として，以下のようなものがある．Bagga ら [3] は，文書に出現する単語を要素とする文書ベクトルを作り，ベクトル空間内において文書間の類似度を計算し，クラスタリングを行った．出現する単語に加え，文書中から人物に関する個人情報を抽出し，クラスタリングする試みとして [4] が挙げられる．佐藤ら [5] は文書中に出現する人名に注目し，共起する人名を用いて Web ページ間のリンクを作成し，それをグループ分けすることで同姓同名の分離を行っている．Wan ら [6] は高い頻度で検索される人名をクエリーとした場合の検索結果を対象にクラスタリングを行うシステムを提案している．Wan らのシステムにおいては，例えば苗字や名前だけがクエリーとして与えられる場合も想定されている．Bekkerman ら [7] は，実世界において同一コミュニティに属する複数の人物に関するページを集め，それらの中でのリンク解析や階層併合・分割ダブルクラスタリングを行うことで，同一コミュニティに属する人物を同時にそれぞれの同姓同名についての記述と分離する方法を提案した．この方法では結果的に同じコミュニティに属する人物について前提知識として情報が与えられた上でクラスタリングを行っていることになる．

本研究で用いている二段階クラスタリングに関する先行研究として次のような研究が挙げられる．Tishby ら [8] によって提案された情報ボトルネックは情報理論を用いて，最適なクラスタリングを求めるアルゴリズムである．情報ボトルネックは Slonim ら [9] によって文書クラスタリングに対して適用されている．彼らは文書クラスタリングに対して，関連すると考えられる単語クラスタリングの結果を用いてクラスタリングを行っている．Liu ら [10] はクラスタを区別するために有効な特徴量を一段階のクラスタの多数決に基づいて求め，二段階クラスタリングを行っている．

また，近年人名の曖昧性の解消を目的とした Web 上での人物検索に関するワークショップ WePS [11] が行われ，様々な知見が明らかとなっている．2006 年から 2007 年にかけて第 1 回が行われ，2008 年から 2009 年に第 2 回が開かれている [12]. WePS の上位チーム [1], [13]〜[16] が用いている方法の多くは文書ベクトル空間の類似度に基づくクラスタリングを用いたものである．

(注1)：http://ja.wikipedia.org/wiki/Category:曖昧さ回避

3. 特徴量抽出

3.1 人物に関連した単語・句の抽出方法の検討

同姓同名の人物のクラスタリングにおいて重要となるのが，文書からの特徴量の抽出である．文書から抽出する固有表現や重要語には同姓同名の人物のクラスタリングにとって不適切な特徴量が含まれており，間違ったクラスタを形成するという問題が発生する．これは文書が複数の話題を扱っているために起こる問題であり，この問題に対処するための方法として特徴量を抽出する範囲を対象人物の周辺に限定する方法が用いられている [2]．しかし，範囲を限定したことによって本来クラスタリングに用いることができる特徴量を用いることができず，クラスタを形成できないという問題が生じる．

本研究では，文書から同姓同名人物のクラスタリングに有効な特徴量抽出方法についての検討を行った．6.2 節の予備実験の結果，固有表現は文書全体から抽出し，重要語は範囲を限定して抽出することで同姓同名の人物のクラスタリングに適した重要語を得ることができることが判明した．

3.2 固有表現抽出

文書から人物に関連した固有名詞である固有表現を抽出する．固有表現として，本研究では人名，地名，組織名を扱っている．しかし，地名，組織名には特定人物との関連が弱く，複数の人物に共通する固有名詞が多く存在する．そのため，あらかじめ作成した不要語辞書を用いて，大域頻度の高い固有名詞は取り除く．

3.3 重要語抽出

文書から検索対象となる人物に関連した単語・句を抽出する方法のもう 1 つである重要語を用いた抽出について説明する．

文書に対して，形態素解析を適用した結果から，言選 Web (注2) を用いて重要語を抽出する [17]．重要語抽出は以下のようにして行う．まず，形態素解析の結果から名詞句 $\mathbf{w}$ を取り出し $\mathbf{w} = \{w_1, w_2, \cdots, w_L\}$ に存在する各単語 w_i について，単語重要度 $\mathrm{LR}(w_i)$ を式 (1) に従って計算する．

$$\mathrm{LR}(w_i) = \sqrt{(\mathrm{LF}(w_i)+1)\cdot(\mathrm{RF}(w_i)+1)} \tag{1}$$

$\mathrm{LF}(w_i), \mathrm{RF}(w_i)$ は文書内の全ての名詞句 $\{\mathbf{w}\}$ 内において単語 w_i の左側，右側に単語が存在する回数であり，これに対して 1 を加えて平滑化を行う．

単語重要度 $\mathrm{LR}(w_i)$ を元にして，名詞句の重要度 $\mathrm{FLR}(\mathbf{w})$ を式 (2) に従って計算する．

$$\mathrm{FLR}(\mathbf{w}) = \mathrm{F}(\mathbf{w}) \cdot \left(\prod_{i=0}^{L} \mathrm{LR}(w_i) \right)^{\frac{1}{L}} \tag{2}$$

$\mathrm{F}(\mathbf{w})$ は文書中の名詞句 $\mathbf{w}$ の出現回数であり，L は名詞句の長さである．

このようにして，抽出した名詞句 $\mathbf{w}$ のうち，重要度 $\mathrm{FLR}(\mathbf{w})$ が閾値 θ_{CKW} 以上の名詞句を重要語とする．

(注2)：http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html

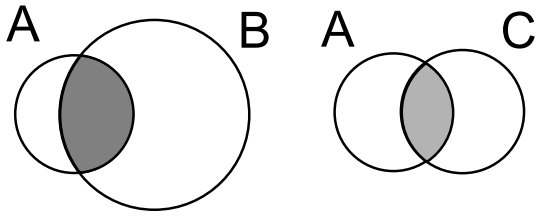

図 1　仮定 4.1 の例

3.4 リンク構造抽出

文書内に含まれる他文書へのリンクを抽出し，特徴量として用いる．文書の<a>タグに含まれる URL と文書自身の URL を抽出し，正規化を行った後，URL による特徴量とする．URL についても，あらかじめ作成した不要語辞書を用いて大域頻度の高い URL を取り除く．

4. 類似度計算とクラスタリング

本研究では，階層併合クラスタリングを用いて，第一段階のクラスタを作成する．階層併合クラスタリングは各文書間の類似度を元にクラスタを生成する．本稿では，階層併合クラスタリングにおける類似度計算を各特徴量の類似度，文書間の類似度，クラスタ間の類似度の 3 点で改良する．

4.1 Overlap 係数の導入

各特徴量の類似度に用いる Overlap 係数 [18] について説明する．Overlap 係数は式 (3) のように計算される．

$$\mathrm{Overlap}(d_x, d_y) = \frac{|\mathbf{f}_x \cap \mathbf{f}_y|}{\max(\min(|\mathbf{f}_x|, |\mathbf{f}_y|), T)} \tag{3}$$

$\mathbf{f}_x, \mathbf{f}_y$ はそれぞれ文書 d_x, d_y に含まれる特徴量の集合である．$|\mathbf{f}_x \cap \mathbf{f}_y|$ は文書 d_x，d_y の共通する特徴量の数であり，$\min(|\mathbf{f}_x|, |\mathbf{f}_y|)$ は文書 d_x，d_y の特徴量の数の最小値である．T は特徴量の極端に少ない文書の影響を減らすために定める分母の取りうる最小値であり，本研究においては $T = 4$ とする．

我々は仮定 4.1 の下で，Overlap 係数の妥当性を述べる．

［仮定 4.1］　文書 $\mathrm{A}, \mathrm{B}, \mathrm{C}(A \leqq B \wedge A \leqq C)$ において，$|\mathrm{A} \cap \mathrm{B}| > |\mathrm{A} \cap \mathrm{C}|$ が成り立つ場合（図 1 参照），A, B 間の類似度は A, C 間の類似度よりも高い状態である．

$|\mathrm{A} \cap \mathrm{B}| > |\mathrm{A} \cap \mathrm{C}|$ が成り立つ場合，Overlap 係数を用いて類似度を計算すると，$\mathrm{Overlap}(\mathrm{A}, \mathrm{B}) > \mathrm{Overlap}(\mathrm{A}, \mathrm{C})$ が成り立つ．一方，cos 類似度 $\cos(\mathrm{X}, \mathrm{Y}) = \frac{\mathrm{X}\cdot\mathrm{Y}}{|\mathrm{X}|\cdot|\mathrm{Y}|}$ を用いて計算した場合，$\cos(\mathrm{A}, \mathrm{B}) > \cos(\mathrm{A}, \mathrm{C})$ は必ず成り立つとは限らない．我々は仮定 4.1 が同姓同名人物のクラスタリングにおいて成り立つとし，短い文書の影響を反映しやすい Overlap 係数を導入する．

4.2 各特徴量ごとの類似度計算方法

- 固有表現

固有表現による類似度 $\mathrm{sim}_{\mathrm{NE}}$ は固有表現抽出を用いて抽出した人名（Person），地名（Location），組織名（Organization）を用いて式 (4) のようにして計算する．

$$\begin{aligned}\mathrm{sim}_{\mathrm{NE}}(d_x, d_y) \;=\; & \alpha_{\mathrm{P}}\,\mathrm{sim}_{\mathrm{P}}(d_x, d_y) + \alpha_{\mathrm{L}}\,\mathrm{sim}_{\mathrm{L}}(d_x, d_y) \\ & + \alpha_{\mathrm{O}}\,\mathrm{sim}_{\mathrm{O}}(d_x, d_y)\end{aligned} \tag{4}$$

式 (4) の $\mathrm{sim}_{\mathrm{P}}, \mathrm{sim}_{\mathrm{L}}, \mathrm{sim}_{\mathrm{O}}$ は各属性の Overlap 係数から計算す

る．$\alpha_{\mathrm{P}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{O}}$ は各属性（人名，地名，組織名）についての重みである（$\alpha_{\mathrm{P}} + \alpha_{\mathrm{L}} + \alpha_{\mathrm{O}} = 1$）．重みは $\alpha_{\mathrm{P}} \gg \alpha_{\mathrm{O}} > \alpha_{\mathrm{L}}$ として，訓練データを用いて定める．

- 重要語

重要語による類似度 $\mathrm{sim}_{\mathrm{CKW}}$ は重要語抽出を用いて抽出した複合語から式 (5) のようにして計算する．

$$\mathrm{sim}_{\mathrm{CKW}}(d_x, d_y) = \mathrm{Overlap}(d_x, d_y) \tag{5}$$

ここでは，抽出した複合語を特徴量として計算している．

- リンク

リンクによる類似度 $\mathrm{sim}_{\mathrm{URL}}$ は元の HTML ファイルに含まれる URL から式 (6) のようにして計算する．

$$\mathrm{sim}_{\mathrm{URL}}(d_x, d_y) = \begin{cases} 1 \text{ if } \quad d_x, d_y \text{ 間に直接リンクがある} \\ \mathrm{Overlap}(d_x, d_y) \quad \text{それ以外の場合} \end{cases} \tag{6}$$

文書 d_x, d_y のどちらか一方がもう一方の URL を特徴量として含んでいる場合は類似度を 1 とし，そうでない場合は Overlap 係数を用いて計算する．

4.3 複数の特徴量による類似度

上記に述べた各特徴量を利用したクラスタリングによって形成されるクラスタを考えると，各クラスタは正解となるクラスタの部分集合であると考えられる．この複数のクラスタリング結果を用いて正解クラスタを作成する方法として，従来研究では生成したクラスタの和集合を同一人物のクラスタとして扱う方法を取っていた．本論文ではこの複数の特徴量での類似度を元にして新たな類似度を作成し，クラスタリングを行う方法を検討する．

複数の類似度から新たな類似度を作成する方法として，各類似度に重みづけして類似度を計算する方法など様々な方法があるが，ここでは類似度の最大値を文章の類似度として扱い，式 (7) のようにして，類似度を計算する．

$$\begin{aligned} \mathrm{sim}_{\max}(d_x, d_y) \;=\; & \max(\mathrm{sim}_{\mathrm{NE}}(d_x, d_y), \\ & \mathrm{sim}_{\mathrm{CKW}}(d_x, d_y), \mathrm{sim}_{\mathrm{URL}}(d_x, d_y)) \end{aligned} \tag{7}$$

$\mathrm{sim}_{\mathrm{NE}}(d_x, d_y)$，$\mathrm{sim}_{\mathrm{CKW}}(d_x, d_y)$，$\mathrm{sim}_{\mathrm{URL}}(d_x, d_y)$ は特徴量 NE，CKW，URL についての類似度である．

式 (7) のようにして，類似度を計算する場合，元の類似度が同一の値域を持つことが必要になる．各特徴量の値域は $[0, 1]$ であり，必要条件を満たしている．この複数の類似度の最大値を用いる方法の特徴として，文書ごとに異なる特徴量の類似度を用いることが挙げられる．異なる特徴量の類似度を用いることによって，単一の特徴量では共通する特徴量が存在しない文書間の類似度についても他の特徴量を用いることで補うことができる．

4.4 階層併合クラスタリング

本研究におけるクラスタリング手法である階層併合クラスタリングについて説明する．階層併合クラスタリングは類似度を用いて類似度の高い要素を順に併合していく手法であり，閾値を用いるため，陽にクラスタ数を決める必要がない．

従来の手法では，大きなクラスタを形成しやすい最短距離法を用いてクラスタリングを行っていたが，最短距離法は分類感度が低く，外乱により誤ったクラスタを形成しやすいという欠点が存在する [19]．そのため，本研究では群間平均法を用いて，階層併合クラスタリングを行う．

群間平均法ではクラスタ C_i, C_j の類似度を次の式 (8) に従って計算する．

$$\mathrm{sim}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{d_x \in C_i} \sum_{d_y \in C_j} \mathrm{sim}_{\max}(d_x, d_y) \tag{8}$$

$\mathrm{sim}_{\max}(d_x, d_y)$ はクラスタ C_i, C_j に属する文書 d_x, d_y の類似度であり，$\mathrm{sim}(C_i, C_j)$ はクラスタ C_i, C_j 内の文書間の類似度の平均として計算される．

類似度の最も高いクラスタを結合する作業を繰り返し，類似度が閾値 θ_{HAC} を下回った時点で終了することによってクラスタリングを行う．このクラスタリングにおいて，文書は 1 人の同姓同名の人物を扱っているとし，各文書は 1 クラスタのみに属する．

本研究では，4.3 節で述べた複数の類似度の最大値を文書間の類似度とみなして，群間平均法を用いることによって，複数の特徴量全体で高い類似度を持つ文書をクラスタにまとめることができるようになった．最短距離法を用いて階層併合クラスタリングを行った場合は閾値以上の類似度を持つ文書は類似度の大きさに関わらず同一のクラスタに含まれるため，従来手法である和集合を用いて複数の特徴量を扱う方法において，各特徴量のクラスタリングでの閾値を一定にした場合と等しい結果となる．4.3 節で述べた類似度は，群間平均法を用いることによって，従来手法に比べて性能を向上させることができる．

5. 二段階クラスタリング

クラスタリングにおける特徴量抽出方法として，クラスタリング結果を利用する方法を提案する．本手法はクラスタリングを二段階で行うことによってクラスタリング結果の精緻化を行っている．一段階目のクラスタリング結果をもとに，各クラスタについて，クエリーを追加し，新たな文書を収集し，クラスタに含まれる文書数を増加させる．これは情報検索におけるクエリー拡張 [20] の考え方を基本とした方法である．本研究で扱っている文書集合は対象人物の名前（NAME）をクエリーとした検索の結果である．二段階目では，一段階目で得られたクラスタリングの結果から特定の人物に関連するクエリー Q を抽出し，“NAME+Q” をクエリーとする検索を元の文書集合に対して行い，得られた結果を新しいクラスタとして扱う．

一段階目においては一人の人物に関するクラスタが複数に分かれた状態で存在している．二段階クラスタリングではこの結果に対して，Precision を下げ，Recall を上げるように働きかけ，一段階目のクラスタリングでは分かれていたクラスタをまとめる．また，一段階目では 1 つの文書は 1 つのクラスターのみ属していたが，二段階目においてはソフトクラスタリングによって複数のクラスターに属することが可能となる．このこと

によって，検索結果のページなど 1 文書中に複数の同姓同名の人物への言及が含まれている場合にも対応できるようになる．

5.1 二段階クラスタリングによるソフトクラスタリングへの対応

全文書集合を D とし，各文書を $d_i(i = 1, \cdots, N)$ とする．また，D を第一段階でクラスタリングした結果を $\mathcal{C}$ とし，各クラスタを $C_k(k = 1, \cdots, M)$ とする．$\mathcal{C}$ に含まれるクラスタは重複する文書を持たない，すなわち，$\forall k, \forall m, (k \neq m), C_k \cap C_m \neq \emptyset$ が成り立つ．

ここでは，$\mathcal{C}$ を元に，二段階クラスタリングの結果 $\mathcal{C}'$ を求める．Algorithm 1 に一連の処理を示す．各処理について説明する．

- 5 行目では，含まれる文書数 $|C_k|$ が最大のクラスタ C_k に含まれる文書を 1 文書として，重要語抽出を行い，重要度の高い上位 m 個の重要語 $\{t_j\}$ を取り出す．
- 7〜8 行目では，他のクラスタに含まれる文書 d_i に対して，t_j が含まれているか確認する．もし，d_i に t_j が 1 つでも含まれていれば，d_i を C_k に追加する．
- 9 行目では，d_i の属するクラスタ C_l が $C_l = \{d_i\}$ となる場合，クラスタ C_l を $\mathcal{C}$ から取り除く．$C_l \neq \{d_i\}$ の場合は C_l を残す．このとき，$d_i \in C_l \wedge d_i \in C_k$ となる．この処理によって，第一段階のクラスタリングにおいて対応していなかった要素の重複を許したクラスタリングが可能となる．
- 15〜17 行目では，C_k を $\mathcal{C}$ から取り除き，$\mathcal{C}'$ に追加する．また，D から C_k に含まれる文書を取り除く．

この処理を $\mathcal{C} = \emptyset$ となるまで繰り返す．閾値 θ_s について $|C_k| < \theta_s$ となるような，文書数の少ないクラスタについては処理は行わない．また，本研究では $m = |C_k|$ とする．

二段階クラスタリングにおいてハードクラスタリングを行う場合，Algorithm 1 の 7〜12 行目の部分を Algorithm 2 に置き換える．

Algorithm 1 二段階クラスタリングによるソフトクラスタリングへの対応

```
1: C' ← ∅
2: while D ≠ ∅ do
3:    C から |C_k| が最大のクラスタ C_k を取り出す
4:    if |C_k| > θ_s  (k = 1, ··· , M) then
5:       {t_j} ← termextraction(C_k)
6:       for d_i ∈ D\C_k do
7:          if d_i has t_j  (d_i ∈ C_l) then
8:             C_k ← C_k ∪ {d_i}
9:             if |C_l| = 1 then
10:               C_l ← C_l\{d_i}
11:            end if
12:         end if
13:      end for
14:   end if
15:   C' ← C' ∪ {C_k}
16:   C ← C\{C_k}
17:   D ← D\C_k
18: end while
19: return C'
```

Algorithm 2 ハードクラスタリングの場合の二段階クラスタリング

```
1: if d_i has t_j  (d_i ∈ C_l) then
2:    C_k ← C_k ∪ {d_i}
3:    C_l ← C_l\{d_i}
4: end if
```

6. 評価実験

英語の同姓同名の人物の文書集合に対して，上記の手法を適用し，クラスタリングの結果を評価する．

実験の手法について説明する．まず，lxml (注3)，Automatic English Sentence Segmenter (注4)を用いて，HTML ファイルを 1 行 1 文形式のテキストファイルに変換する．このファイルに対して，形態素解析，固有表現抽出を行う．形態素解析には，Tree Tagger (注5)，固有表現抽出には，Stanford NER (注6)を用いた．また，形態素解析の結果を用いて，重要語抽出を行い，HTML ファイルのタグから URL の抽出を行う．これらの特徴量を元に文書間の類似度を計算し，上記の手法を適用した．文書間の類似度は特徴量から計算した類似度の最大値とした．固有表現の類似度計算に用いる重みは訓練データを用いて，$\alpha_{\mathrm{P}} = 0.78, \alpha_{\mathrm{O}} = 0.16, \alpha_{\mathrm{L}} = 0.06$ とした．

(注3)：http://codespeak.net/lxml/
(注4)：http://www.answerbus.com/sentence/
(注5)：http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
(注6)：http://nlp.stanford.edu/software/CRF-NER.shtml

実験に用いるデータセットは，WePS の第 1 回目，第 2 回目のデータセット WePS-1 (注7)，WePS-2 (注8)を用いた．各データは検索エンジンにおいて，人名での検索結果の上位ページを取ってきたものであり，取得不可能なものも合わせて，WePS-1 は最大 100 ページ，WePS-2 は最大 150 ページである．人名の数はともに 30 である．データセットには人手で作成した同一人物のクラスタの正解データが存在する．これらのデータは 1 つの文書が複数の同姓同名の人物について述べている場合を許容しており，複数のクラスタに属する文書が存在している．

6.1 評価方法

評価方法としては，Purity/Inverse Purity と extended B-Cubed 指標を用いた．どちらの指標についても F-measure により，総合的なシステムの性能を評価する．これらの評価方法は同一文書が複数のクラスタに属することを許容した場合の評価方法である．

Purity, Inverse Purity による評価方法は以下の通りである [11]．結果のクラスタ集合を $\mathcal{C} = \{C_1, \cdots, C_i, \cdots, C_N\}$, 正解のクラスタ集合を $\mathcal{L} = \{L_1, \cdots, L_j, \cdots, L_M\}$ とする．任意の 2 クラスタ C_i, L_j の精度 Precision(C_i, L_j) を，

$$\mathrm{Precision}(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \tag{9}$$

(注7)：http://nlp.uned.es/weps/weps-1-data/
(注8)：http://nlp.uned.es/weps/weps-2-data/

と定義する．このとき， Purity 及び Inverse Purity は，

$$\mathrm{P} = \sum_i \frac{|C_i|}{N} \max_j \mathrm{Precision}(C_i, L_j) \tag{10}$$

$$\mathrm{IP} = \sum_i \frac{|L_i|}{N} \max_j \mathrm{Precision}(L_j, C_i) \tag{11}$$

となり，Purity/Inverse Purity $\mathrm{F_{P\text{-}IP}}$ は，

$$\mathrm{F_{P\text{-}IP}} = \frac{1}{\frac{1}{2}\left(\frac{1}{\mathrm{P}} + \frac{1}{\mathrm{IP}}\right)} \tag{12}$$

と計算される．

Amigó ら [21] は Purity/Inverse Purity が同姓同名人物クラスタリングの評価指標として不十分であることを示し，extend B-Cubed 指標を提案した [21]．extended B-Cubed 指標について説明する．文書 e が属するクラスタリング結果のクラスタ，正解クラスタをそれぞれ $C(e), L(e)$ とする．

extended B-Cubed 指標を算出する際に用いられる Multiplicity Precision （MP），Multiplicity Recall （MR）は式 (13)，(14) にして計算される．

$$\mathrm{MP}(e, e') = \frac{\min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|} \tag{13}$$

$$\mathrm{MR}(e, e') = \frac{\min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|L(e) \cap L(e')|} \tag{14}$$

これらの指標を用いて，extended B-Cubed Precision （BEP），extended B-Cubed Recall （BER）を式 (15)，(16) のように $\mathrm{MP}(e, e'), \mathrm{MR}(e, e')$ の平均値を取ることで求められる．

$$\mathrm{BEP} = \mathop{\mathrm{Avg}}_e \left[\mathop{\mathrm{Avg}}_{e'.C(e) \cap C(e') \neq \emptyset} [\mathrm{MP}(e, e')] \right] \tag{15}$$

$$\mathrm{BER} = \mathop{\mathrm{Avg}}_e \left[\mathop{\mathrm{Avg}}_{e'.L(e) \cap L(e') \neq \emptyset} [\mathrm{MR}(e, e')] \right] \tag{16}$$

extended B-Cubed F-measure は extended B-Cubed Precision，extended B-Cubed Recall を元にして，

$$\mathrm{F_{BEP\text{-}BER}} = \frac{1}{\frac{1}{2}\left(\frac{1}{\mathrm{BEP}} + \frac{1}{\mathrm{BER}}\right)} \tag{17}$$

と求められる．

Purity/Inverse purity は WePS-1 で用いられた指標であり，extended B-Cubed は WePS-2 で用いられた指標である．本研究の実験の評価では，extended B-Cubed 指標, Purity/Inverse Purity 指標を用いた．

6.2 実験:特徴量抽出における抽出範囲決定

固有表現と重要語，それぞれを文書から抽出する際に抽出範囲を変え，クラスタリングの性能を比較した．抽出範囲は検索クエリー（人名）の前後 50 語，100 語，200 語，文書全体と変え，抽出した．

表 1，表 2 は固有表現について抽出範囲を変更した場合の結果であり，表 3，表 4 は重要語について抽出範囲を変更した場合の結果である．50，100，200 は抽出範囲を表し，All は抽出範囲が文書全体の場合である．ここでは，閾値 $\theta_{\mathrm{HAC}} = 0$ とし，

表 1　WePS-1 データセットにおける固有表現の抽出範囲変更

Topic	BEP	BER	$\mathrm{F_{BEP\text{-}BER}}$	P	IP	$\mathrm{F_{P\text{-}IP}}$
NE, 50	0.96	0.50	0.64	0.90	0.62	0.72
NE, 100	0.95	0.54	0.67	0.89	0.65	0.74
NE, 200	0.93	0.57	0.69	0.88	0.68	0.76
NE, All	0.86	0.65	**0.73**	0.76	0.76	0.74

表 2　WePS-2 データセットにおける固有表現の抽出範囲変更

Topic	BEP	BER	$\mathrm{F_{BEP\text{-}BER}}$	P	IP	$\mathrm{F_{P\text{-}IP}}$
NE, 50	0.94	0.45	0.58	0.95	0.57	0.70
NE, 100	0.96	0.48	0.62	0.96	0.61	0.73
NE, 200	0.96	0.52	0.65	0.97	0.64	0.75
NE, All	0.93	0.62	**0.73**	0.95	0.73	0.82

表 3　WePS-1 データセットにおける重要語の抽出範囲変更

Topic	BEP	BER	$\mathrm{F_{BEP\text{-}BER}}$	P	IP	$\mathrm{F_{P\text{-}IP}}$
CKW, 50	0.96	0.41	0.56	0.90	0.53	0.65
CKW, 100	0.95	0.47	0.61	0.89	0.59	0.69
CKW, 200	0.94	0.53	0.66	0.88	0.64	0.73
CKW, All	0.82	0.64	**0.70**	0.73	0.74	0.72

表 4　WePS-2 データセットにおける重要語の抽出範囲変更

Topic	BEP	BER	$\mathrm{F_{BEP\text{-}BER}}$	P	IP	$\mathrm{F_{P\text{-}IP}}$
CKW, 50	0.90	0.38	0.50	0.92	0.50	0.63
CKW, 100	0.93	0.40	0.53	0.94	0.52	0.66
CKW, 200	0.96	0.44	0.58	0.96	0.57	0.70
CKW, All	0.91	0.59	**0.71**	0.93	0.71	0.80

類似度が 0 ではない場合は全て結合するように閾値を定めた場合を扱っている．全ての結果において，抽出範囲を全体に拡張した方が $\mathrm{F_{BEP\text{-}BER}}$ が向上している．

6.3 実験:提案方法によるクラスタリング

提案手法によるクラスタリング手法を比較した．結果を表 5，表 6 に示す．URL，CKW，NE は各特徴量を単独で用いたクラスタリングの結果であり，それぞれが URL，重要語，固有表現を用いている．MAX は URL，CKW，NE の類似度の最大値から生成した類似度を用いてクラスタリングを表わしている．これらのクラスタリングには群間平均法を用いた階層併合クラスタリングを用いている．抽出範囲に関しては CKW，NE は文書全体から抽出した場合の結果を示した．また，この MAX からに対して二段階クラスタリングを適用した結果を QE1，QE2 に示した．QE1 は 1 文書が 1 クラスタに属するハードクラスタリングであり，QE2 は文書が複数のクラスタに属することを許容したソフトクラスタリングである．MAX は予備実験において，最も性能が高かった固有表現，重要語をクエリーの前後 100 語 から抽出した場合の結果を示している．

実験におけるベースラインとして， ALL IN ONE， ONE IN ONE， COMBINED を用いた．ALL IN ONE は全ての文書を 1 クラスタにする場合， ONE IN ONE は各文書を 1 文書 1 クラスタに分けた場合， COMBINED は ALL IN ONE と ONE IN ONE のクラスタを合わせた場合である． COMBINED において，各文書は ALL IN ONE と ONE IN ONE の 2 クラスタに属することになる．また，WePS-1 データセットでは WePS-1 の上位チームの結果を併記した．

MAX， QE1，QE2 の結果は閾値 θ_{HAC} を変化させて得られ

たクラスタ集合のうち，最も性能の良いものである．また，このときの抽出範囲は固有表現が文書全体，重要語がクエリー前後 100 語である．

WePS-1 データセットの結果について説明する．extended B-Cubed 指標では，NE, MAX, QE, QE2 において WePS-1 上位チームを上回る結果が得られ，QE, QE2 が最高値 $F_{\text{BEP-BER}} = 0.78$ を示した．一方，Purity/Inverse Purity 指標では WePS-1 上位チームの結果を下回る結果となった．このような結果が得られた原因として，extended B-Cubed 指標が Purity/Inverse Purity 指標では評価できなかった Rag Bag [21] と呼ばれるクラスタの評価を行っていることが挙げられる．本研究の手法では，Rag Bag への対応が改善されたために性能が向上したと考えられる．また，BEP が WePS-1 上位チームと比較して，向上している点もこのことを裏付けている．

WePS-2 データセットの結果について説明する．WePS-2 のデータセットは ALL IN ONE ベースラインが ONE IN ONE ベースラインに比べて良い性能を示していることより，平均的に大きなクラスタによって構成されていると考えられる．この結果においては QE2 が最高値 $F_{\text{BEP-BER}} = 0.82$ を示した．これは MAX での値 $F_{\text{BEP-BER}} = 0.79$ を 0.03 上回る結果であり，WePS-1 データセットでの結果に比べて，改善値が大きくなっている．これらより，QE, QE2 は大きなクラスタによって形成される文書集合に対するクラスタリングにおいて有効であると考えられる．

WePS-1 データセット，WePS-2 データセットの結果から二段階クラスタリングを用いることによって評価が改善できていることが確認できた．現在のクラスタリング手法では閾値を最も低く設定した場合も含めて全体的に Precision が高く，実際の人物に関するクラスタが十分にまとまっていないことが示唆されている．Recall の改善のためには新たな特徴量を導入し，類似度計算の方法を改良することが必要となると考えられる．

表 5　WePS-1 データセットによる評価実験

Topic	BEP	BER	$F_{\text{BEP-BER}}$	P	IP	$F_{\text{P-IP}}$
Baseline						
ALL IN ONE	0.18	0.98	0.25	0.29	1.00	0.40
ONE IN ONE	1.00	0.43	0.57	1.00	0.47	0.61
COMBINED	0.17	0.99	0.24	0.64	1.00	0.78
First-Stage Clustering						
URL	0.98	0.48	0.62	0.83	0.56	0.64
CKW	0.82	0.64	0.70	0.73	0.74	0.72
NE	0.86	0.65	0.73	0.76	0.76	0.74
MAX	0.85	0.73	0.77	0.74	0.82	0.76
Second-Stage Clustering						
QE	0.84	0.76	**0.78**	0.74	0.84	0.77
QE2	0.83	0.76	**0.78**	0.74	0.84	0.77
WePS top 5						
1st	0.67	0.81	0.71	0.72	0.88	**0.79**
2nd	0.68	0.73	0.68	0.75	0.80	0.77
3rd	0.68	0.71	0.67	0.73	0.82	0.77
4th	0.79	0.50	0.58	0.81	0.60	0.69
5th	0.43	0.84	0.53	0.53	0.90	0.67

表 6　WePS-2 データセットによる評価実験

Topic	BEP	BER	$F_{\text{BEP-BER}}$	P	IP	$F_{\text{P-IP}}$
Baseline						
ALL IN ONE	0.43	1.00	0.53	0.56	1.00	0.67
ONE IN ONE	1.00	0.24	0.34	1.00	0.24	0.34
COMBINED	0.43	1.00	0.52	0.78	1.00	0.87
First-Stage Clustering						
URL	0.98	0.29	0.39	0.47	0.99	0.34
CKW	0.91	0.59	0.71	0.93	0.71	0.80
NE	0.93	0.62	0.73	0.95	0.73	0.82
MAX	0.94	0.69	0.79	0.96	0.78	0.86
Second-Stage Clustering						
QE	0.89	0.76	0.81	0.93	0.83	0.87
QE2	0.88	0.77	0.82	0.92	0.84	0.88

7. おわりに

本研究では Web における同姓同名問題の解消策として，検索結果のクラスタリングを行い，人物を同定したクラスタを作成する手法についての検討を行った．我々は名詞句が同姓同名の人物のクラスタリングにおいて重要であると考え，クラスタリングに有効な名詞句を抽出する方法として，固有表現の抽出と重要語の抽出を行った．また，各特徴量の類似度から新しい類似度を計算し，複数の特徴量を利用したクラスタリングを行った．

さらに，重要語からクラスタリングに有効な特徴量を抽出する手段として，クラスタリング結果から各クラスタに関連した重要語を選択し，再クラスタリングを行った．この再クラスタリングによって，同一人物のクラスタのまとまりを高めることができた．また，二段階クラスタリングによって 1 文書内で複数の同姓同名の人物が取り扱われる場合を取り扱うことも可能となった．

本研究で用いた手法の評価として，Web 上での人物検索に関するワークショップ WePS のデータセットを用いて実験を行い，評価を行った．その結果， WePS-1 データセットで $F_{\text{BEP-BER}} = 0.78$，WePS-2 データセットで $F_{\text{BEP-BER}} = 0.82$ を示した．この実験により，複数の特徴量によるクラスタリング，二段階クラスタリングがともに性能を向上していることも明らかとなった．

今後の課題として，抽出した名詞句から同姓同名の抽出に有効な名詞句を選択する手法の研究を進める．

文　献

[1] E. Elmacioglu, Y. Tan, S. Yan, M. Kan and D. Lee: “PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features”, The SemEval-2007, pp. 268–271 (2007).

[2] S. Ono, I. Sato, M. Yoshida and H. Nakagawa: “Person Name Disambiguation in Web Pages Using Social Network, Compound Words and Latent Topics”, LECTURE NOTES IN COMPUTER SCIENCE, **5012**, pp. 260–271 (2008).

[3] A. Bagga and B. Baldwin: “Entity-based cross-document coreferencing using the Vector Space Model”, Proceedings of the 17th international conference on Computational linguistics-Volume 1, pp. 79–85 (1998).

[4] G. S. Mann and D. Yarowsky: “Unsupervised personal

name disambiguation", Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, pp. 33–40 (2003).
[5] 佐藤, 風間, 福田, 村上："実世界指向 web マイニングによる同姓同名人物の分離 (<特集>情報融合)", 情報処理学会論文誌. データベース, **46**, 8, pp. 26–36 (2005).
[6] X. Wan, J. Gao, M. Li and B. Ding: "Person resolution in person search results: Webhawk", CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 163–170 (2005).
[7] R. Bekkerman and A. McCallum: "Disambiguating Web appearances of people in a social network", Proceedings of the 14th international conference on World Wide Web, pp. 463–470 (2005).
[8] N. Tishby, F. C. Pereira and W. Bialek: "The information bottleneck method", Proceedings of the 37-thAnnual Allerton Conference on Communication (2000).
[9] N. Slonim and N. Tishby: "Document clustering using word clusters via the information bottleneck method", Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 208–215 (2000).
[10] X. Liu, Y. Gong, W. Xu and S. Zhu: "Document clustering with cluster refinement and model selection capabilities", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 191–198 (2002).
[11] J. Artiles, J. Gonzalo and S. Sekine: "The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task", The SemEval-2007, pp. 64–69 (2007).
[12] J. Artiles, S. Sekine and J. Gonzalo: "Web people search: results of the first evaluation and the plan for the second", WWW '08: Proceeding of the 17th international conference on World Wide Web, pp. 1071–1072 (2008).
[13] Y. Chen and J. Martin: "CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation", The SemEval-2007, pp. 125–128 (2007).
[14] O. Popescu: "IRST-BP: Web People Search Using Name Entities", The SemEval-2007, pp. 195–198 (2007).
[15] K. Balog, L. A. Azzopardi and M. de Rijke: "Uva: Language modeling techniques for web people search", The SemEval-2007, pp. 468–471 (2007).
[16] H. Saggion: "SHEF: Semantic Tagging and Summarization Techniques Applied to Cross-document Coreference", The SemEval-2007, pp. 292–295 (2007).
[17] H. Nakagawa and T. Mori: "Automatic term recognition", Terminology, **9**, 2, pp. 201–219 (2003).
[18] C. Manning and H. Schütze: "Foundations of statistical natural language processing", MIT Press (1999).
[19] S. Kamvar, D. Klein and C. Manning: "Interpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach", ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 283–290 (2002).
[20] J. Xu and W. B. Croft: "Query expansion using local and global document analysis", SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 4–11 (1996).
[21] E. Amigó, J. Gonzalo, J. Artiles and F. Verdejo: "A comparison of extrinsic clustering evaluation metrics based on formal constraints", Information Retrieval, pp. 1–26 (2008).