検索結果スニペットのクラスタリングによる同姓同名人物の特定

木村 塁[†]　戸田 浩之[‡]　田中 克己[†]

†京都大学大学院情報学研究科社会情報学専攻
〒606-8501 京都市左京区吉田本町
‡日本電信電話株式会社 NTT サイバーソリューション研究所
〒239-0847 神奈川県横須賀市光の丘 1-1

E-mail: †{kimura, tanaka}@dl.kuis.kyoto-u.ac.jp, ‡toda.hiroyuki@lab.ntt.co.jp

あらまし 多様な情報が WWW に存在する現在，人物や組織等の実世界オブジェクトの情報を知るため，しばしば Web 検索エンジンが利用される．この場合，ユーザは検索エンジンに所望のオブジェクトの名称を入力し，その名称を含む文書のリストを取得する．しかし，検索結果中には同名の人物や組織が混在するため，ユーザはまず各文書が目的のオブジェクトについて記述されているかを判断し，その後に本当に欲しい情報を抽出するという 2 段階の手順を経なければならない．本稿では，このような検索を支援する最初のステップとして，同名が多い人物を検索する場合に検索結果を実世界の人物単位で提示することを考える．この問題に対して，我々は文書中で検索対象の人名とともに言及されている組織や人名等の関連オブジェクトと，検索対象である人物の役割や肩書きを示す属性情報に注目する事で，実世界の人物単位に検索結果の文書を分類できるのではないかと考えた．本稿では，このアイディアに基づき，Web 検索エンジンの検索結果を分類する実験を行い，人物単位での検索結果分類精度について評価を行った結果について報告する．

キーワード Web とインターネット，情報検索，クラスタリング

Classifying Namesakes by Clustering Web Search Results

Rui KIMURA[†]　Hiroyuki TODA[‡]　and　Katsumi TANAKA[†]

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshidahommachi, Sakyo-ku, Kyoto 606-8501, Japan
‡ NTT Cyber Solutions Laboratories, NTT Corporation
1-1Hikari-no-Oka, Yokosuka-shi, Kanagawa 239-0847, Japan
E-mail: †{kimura, tanaka}@dl.kuis.kyoto-u.ac.jp, ‡toda.hiroyuki@lab.ntt.co.jp

Abstract There are various information in the Web and people often use WWW search engines to obtain information about real-world objects, people, organizations, etc. In this case, they input the name of the desired object in a search box of a web search engine as a query, and they can obtain a list of documents, all of which include the query term, as a search result. Because same name people and organizations are mixed in the result, they have to find which documents are written about the desired objects before obtaining desired information. This paper attempts to solve the problem of namesakes as the first step of assistance of such search, and tries to show a search result in which documents about same people are clustered. We have an idea that namesakes in a search result can be clustered by using related objects, such as name of organizations and people, and attributes, such as roles and jobs, mentioned together with the query, desired person's name, on a document. We base on this idea and report experiments to classify namesakes in search results and evaluations of accuracy of the classification.

Keyword Web and Internet, Information Retrieval, Clustering

1. はじめに

多様な情報が WWW 空間上に存在する現在，Web 検索エンジンを利用して人物や組織，店舗等の実世界オブジェクトの情報を調べる事はごく当たり前の事となっている．これらの情報を調べる際，ユーザは Web 検索エンジンに所望のオブジェクトの名称を入力し，その名称を含む文書のリストを検索結果として取得する．現在利用されている一般的な Web 検索エンジンでは，PageRank アルゴリズム[1]や HITS アルゴリズム[2]などのリンク構造に基づいたランキングアルゴリズムを利用する事でドキュメント群をランキング化し，そのランキングを検索結果として提示する．しかし，同名

の人物や組織が存在するようなクエリーが検索に用いられた場合，所望のオブジェクトが検索結果のランキング中に散らばって提示される事になる．このため，ユーザはまず各文書が目的のオブジェクトについて記述されているかを判断し，その後に本当に欲しい情報を抽出するという 2 段階の手順を経なければならず，ユーザを煩わせる原因となっている．

最近では，こうしたユーザの不便を解消し，ユーザが目的の情報の迅速な取得を支援するために，様々な研究が行われている．既に一般に公開され商用利用されている Web 検索エンジンシステムもあり，Clusty[3]や SRC[4]などのいくつかのシステムでは，検索結果をクラスタリングして提示する検索システムが運用されている．しかし，これら商用利用されているシステムでは，ページの内容の類似性に基づきクラスタリングしているため，必ずしもユーザが求めているオブジェクト毎に検索結果が分類されるとは限らないという問題点が残る．

同名のオブジェクトが引き起こす問題は，人物や組織，地名など様々なオブジェクトにおいて起こりうるが，その中でも人名に関しては，同姓同名が存在するケースが少なくないため，こういった問題が頻繁に起こっている．さらに人名に関しては，名前から分野や地域などのある程度の情報を知る事ができる組織名や店舗名と異なり，姓名のみからその人物の情報を類する事が困難である．また，人名をクエリーとした Web 検索は，現在は Web 検索全体の 5-10%と言われているが，Social Networking Service の急激な普及に見られるように，WWW 空間上における人と人との繋がりがこれまで以上に重視されるようになる中で，人名をクエリーとした検索の割合がさらに増えていくのではないかと考えられている．

こうした事から本稿では，対象を人名に限定し，検索を支援する最初のステップとして，検索結果を実世界のオブジェクト単位(人単位)で提示することを考える．この問題に対して，我々は文書中で検索対象の人名とともに言及されている組織や人名等の関連オブジェクトと，検索対象である人物の，役割や肩書きなどを示す職業に関連した表現に注目する事で，実世界の人物単位に検索結果の文書を分類できるのではないかと考えた．本稿では，このアイディアに基づき，Web 検索エンジンから得られる検索結果を分類する実験を行い，人物単位での検索結果分類精度について評価を行った結果について報告する．

以下，2 章で関連研究，3 章で提案手法の概要を，4 章と 5 章で提案手法を二つのフェーズに分けて説明し，6 章で評価実験について,7 章でまとめと今後の課題を述べる．

2. 関連研究

Web 検索エンジンの検索結果中の人名の多義性解消については，現在も様々な手法が試されている段階であり，以下のような研究がある．

Al-Kamha[5]らは, 独立な三種類のアプローチを組み合わせる事で，Web ページが同一人物を示す文書であるかどうかを判定している．この三種類のアプローチとは，一つ目は Web ページ中に出現する電話番号，E-mail アドレス，郵便番号等の ID 情報が同じであるかどうか, 二つ目は Web ページ間のベクトル間類似度, そして三つ目は Web ページのリンク構造の分析である. これら三種類のアプローチを訓練データに適用し，同一人物であるかどうかの確信度を求め，三種類の確信度行列を作成し合成する事で，最終的な確信度行列を求める．この確信度行列と閾値を用いて，Web 検索エンジンから得られた検索結果を人物ごとにグループ化して提示している．

Wan らの WebHawk[6]では，人名による Web 検索結果を同一人物ごとにクラスタリングし，それぞれの人を特徴付けるような単語を提示する事で，ユーザが所望の人物を人物の特徴から選択できるようにしている．この WebHawk では，まず人物情報を含まないノイズとなる Web ページを省き, 人名での検索結果を解析して ID 情報や組織名などを取得し，その情報を用いてクラスタリングを行い，さらにそれぞれのクラスターを特徴付けるような記述を提示するという，四段階からなるシステムを提案している．また, WebHawk では英語人名に多く存在するミドルネームを用いる事で，クラスタリング精度を向上させている．

これら二つの研究は，Web ページ本文を解析する事で，同姓同名の分離・クラスタリングを行っているため，Web ページ本文を解析せずタイトルとスニペットのみを解析対象とする我々の研究と異なる．

検索結果やそこから導かれる Web ページを解析する事以外の手法で同姓同名の分離に取り組んでいる研究もある．以下の二つは，クエリーとする人物の周辺に存在する人名を用いる事で，同姓同名人物を分離している．佐藤[7]らは，人間関係を用いる事で同姓同名の分離を行う手法を提案している．Social Network 内の人間関係を用いた同姓同名の分離手法については Bekkerman[8]らによって提案されている．

Toda[9]らは，ニュース記事検索結果のクラスタリングに，検索結果文書中に存在する固有表現を利用する手法を提案している．この手法自体は，同姓同名を対象とした手法ではないが固有表現として抽出される情報は実世界のオブジェクトを特定するには有益な手法であると言える．

3. 提案手法の概要

2 章で述べた関連研究の多くは検索結果の Web ページに実際にアクセスし、その Web ページを解析する手法である．しかし，Zamir[10]らによって指摘されているように，実際の検索システムを考慮した場合，検索を行った時点で，大量の Web ページにアクセスする手法は現実的ではない．

そこで，我々は，実際の Web ページを取得せずに検索結果中のスニペットのみを利用し，検索結果に含まれる同姓同名人物を分類する事を考える．表 1 は検索結果上位 100 件をクラスタリングする際に現れる，解析対象の違いによる解析するページ数と取得するページ数の違いを表に表したものであるが，この表 1 に示すように，スニペットのみを解析対象とすることで，検索結果のクラスタリングに必要な Web ページのダウンロード数を大幅に低減することができ，また，解析時間も大幅に短縮できることができる．このことから，同程度の分類精度を得る事ができるならば後者の方が有益である．

解析対象	解析ページ数	取得ページ数
Web ページ群	100 ページ	101 ページ
タイトルとスニペット	1 ページ	1 ページ

表 1. 解析対象の違いによる解析ページ数と取得ページ数の違い

しかし，Web ページが解析できないことで，従来研究で利用していた人物に関する ID 情報は必ずしも全て利用する事はできず，また，Web ページ中のリンク情報等は全く利用する事が出来ない．

その一方で，我々人間が検索結果を見る場合には，ある程度の精度で検索結果の人物を分類する事が可能である．人間が検索結果ページ中のどのような言葉を見て，同姓同名の人物の区別をつけているかを考えたところ，人間は，関連研究などで用いられているような人物のメールアドレスや電話番号で区別をつけているわけではなく，人物と同時に出現する人名や組織名（以下，関連オブジェクト），また，その人物自体の職業や肩書き（以下，職業表現）を見ることで，同姓同名人物を見分けている事が分かった．

そこで我々は，検索結果に現れるページタイトルとスニペット内の関連オブジェクトと職業表現を抽出し，その言葉を利用してクラスタリングする手法を提案し，同姓同名の分離を実現する事とした．

この手法は以下のような手順で実現される．

1. 与えられた人名をキーワードとして検索エンジンに問合せし，結果を取得
2. 個々の検索結果を一文書とみなし，それぞれの結果の文書ベクトルを生成
   (ア) 個々の検索結果のタイトルとスニペットからタームを抽出
   (イ) それぞれのタームに重み付けを行い，文書ベクトルを生成
3. 文書ベクトルより求めた文書間の類似度を利用してクラスタリングを実施

次章より，文書ベクトルの生成法，クラスタリング法に分けて提案手法を説明する．

4. 文書ベクトルの生成法

4.1. タームの抽出法

文書ベクトルを生成するためには，ベクトルの要素となるタームを抽出する必要がある．この節では，タームの抽出方法について述べる．

抽出対象のタームとしては，関連オブジェクトと職業表現を抽出しており，これらはそれぞれ別々の手法で抽出している．

関連オブジェクトである人名や組織名は固有表現[11]と呼ばれる物の一つである．固有表現は人名と組織名の他に，地名，時間，日時，割合表現，金額表現，固有物名の計 8 種類からなり，これらを文章から抽出する研究が 1990 年代から盛んに行われている．今回は渡辺らによって開発され公開されている固有表現抽出ツール NExT[12]を用いて関連オブジェクトを抽出した．

また，職業表現に関しては，日本語語彙体系[13]内で職業・地位・役割以下に分類される二文字以上の長さの言葉を取得し，これを形態素解析器”茶筌”[14]の辞書に追加する事で，スニペットとページタイトルから職業表現を抽出した．表 2 に職業表現に分類される言葉の一部を例として示す．表 2 中に現れる“国会議員”という言葉を見ると，職業と地位の両方に存在しているが，この言葉のように複数のカテゴリに分類する言葉も多く存在する．本論文では職業・地位・役割の三つのカテゴリの言葉を区別せず利用しているため，複数のカテゴリに同じ言葉が存在しても特に問題が生じる事はない．

カテゴリ	単語例
職業	医長，アナリスト，通訳，院生，留学生，ゴルファー，ゴールキーパー，神主，国会議員，地方長官など
地位	王様，内閣官房長官，ストアマネージャー，監査役，舞台監督，アシスタント，博士，国会議員など
役割	スタッフ，インストラクター，代表取締役社長，シナリオライター，ストッパー，先発投手，球審など

表 2. 職業表現に分類される言葉の例

我々の提案手法を評価する際に，これら二種類のタームを利用したクラスタリングと比較する対象が必要である．一般的に，ページの類似度を測るためにページ内の名詞を利用する事が多いため，今回は茶筌を用いた単語抽出も行った．茶筌を用いてスニペットとページタイトルをそれぞれ形態素解析し，結果の中から名詞と未知語だけを取り出した．

4.2. タームの重み付け法

ベクトル空間モデルに基づく文書ベクトルを作成した．文書ベクトルには 3.1.1 項で抽出したタームを使い，タームに重み付けをする際に一般的に用いられている TF(term frequency)と IDF(inverse document frequency)を用いてベクトルに重みをつけた．

$$w(d,m_x) = tf(d,m_x)\{1+\log(n/df(d,m_x))\} \quad (1)$$

- $w(d,m_x)$: 文書d中でのタームm_xの重み
- $tf(d,m_x)$: 文書d中のタームm_xの出現回数
- $df(d,m_x)$: タームm_xが出現する文書の数
- n: 文書の総数

ベースラインとしては単純な形態素を使って，以下の文書ベクトル $v(d)$を利用した．

$$v(d) = (w(d,l_1), w(d,l_2), \ldots, w(d,l_n)) \quad (2)$$

- $v(d)$: 文書 d の形態素を用いた文書ベクトル
- $w(d,l_x)$: 文書d中でのタームl_xの重み

一つ目の提案手法として，上記で述べた関連オブジェクトや職業表現のタームを用いて以下の文書ベクトル $u(d)$を作成した．

$$u(d) = (w(d,k_1), w(d,k_2), \ldots, w(d,k_n)) \quad (3)$$

- $u(d)$: 文書 d の関連オブジェクトや職業表現を用いた文書ベクトル
- $w(d,k_x)$: 文書d中でのタームk_xの重み

これを利用した結果は 4.1 節に述べる．

また，我々は，関連オブジェクトや職業表現のタームからなる文書ベクトルと形態素解析によって得られた形態素をもとにした文書ベクトルを組み合わせた文書ベクトルを利用することで，関連オブジェクトや職業表現による効果と単純な形態素による効果の両方を得られと考え，以下の式(4)のベクトル $V(d)$を利用した実験も行った．

$$V(d) = v(d) + \alpha\, u(d) \quad (4)$$

- α: 重み

これを利用した結果は 4.2 節に述べる．

5. クラスタリング手法

文書間の類似度を，文書ベクトル間のコサイン類似度を用いて計算する．文書 a の文書ベクトルを $v(a)$，文書 b の文書ベクトルを $v(b)$とすると，類似度 $sim(a,b)$ は以下の式(5)で表される．

$$sim(a,b) = \cos\theta = \frac{v(a)\cdot v(b)}{|v(a)||v(b)|} \quad (5)$$

こうして計算された類似度を使い，最も似ている文書（最も類似度の大きな文書）から順にクラスターを併合していく最短距離法を用いてクラスタリングを行った．

6. 評価実験

3 章で述べた提案手法を実装し，評価実験を行った．実験対象は，身近な研究者を中心とした同姓同名が存在する 10 人の人名を用い，これらをクエリーとした Google[15]検索上位 100 件の検索結果ページを取得し，これらをクラスタリング対象とした．同じサーバに含まれる Web ページのタイトルとスニペットを一つの文書とみなした．

予め人手でクラスタリングの正解データを作成した．図 1 はクラスターを形成する文書数（クラスターサイズ）とクラスター数の関係を示したものである．クラスターサイズが 1，つまりどの文書とも結びつかないような 1 つの文書だけのクラスターが，全体のクラスターの 68%を占めている．平均クラスターサイズは 4.5，最大クラスターサイズは 68 であった．

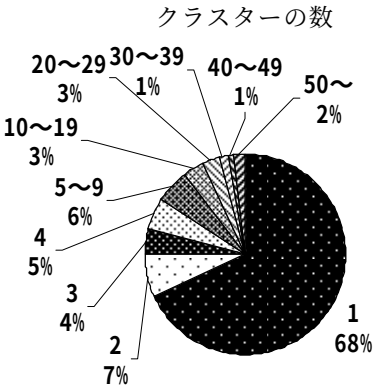

図 1. クラスターを形成する文書数とクラスター数の関係

図 2 はクラスターサイズと文書数の関係である．図 1 を見ると，クラスターサイズが 4 以下のクラスターが 84%ととなり，ほとんどのクラスターのサイズが 4 以下となっているが，文書数との関係で見れば，クラスターサイズが 10 以上のクラスターを形成する文書が 68%と過半数を占める．

なお，一つの人名に対し，文書数の平均個数は 64.7 個であった．

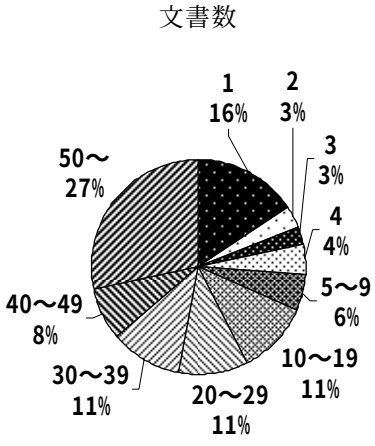

図 2. クラスターサイズと文書数の関係

クラスタリング精度の評価については，個々の正解クラスターと最も一致するクラスターとの類似度の重み付き(重み＝クラスターサイズ)平均である F-score measure[16]を用いて評価した. F-score の算出式は以下に示す通りである.

$$F-score=\sum_{c\in C}\frac{|m_c|}{|m|}\max_{r\in R}\frac{2|m_{r,c}|}{|m_r|+|m_c|} \qquad (6)$$

この式の中で,

- Cは正解セット中の人物集合，cは正解セット中での一人物，m_cは人物cに関する文書の集合
- Rはクラスタリング分類結果中での人物集合，rはクラスタリング分類結果中での一人物，m_rは人物rに関する文書の集合
- $m_{r,c}$は$m_r \cap m_c$
- m は正解セット中に含まれる文書集合

を示している.

6.1. 関連オブジェクト･職業表現による分類

形態素解析で得た文書ベクトルと，提案手法である関連オブジェクトと職業表現を使った文書ベクトルとの，クラスタリング精度を比較した．最短距離法での閾値を 0.05 から 0.25 まで 0.025 間隔で変化させて分類を行った. 結果を図 3 に示す. 縦軸が分類精度の F-score, 横軸がクラスタリングに用いた閾値である．図 3 中の 4 つの系列,「形態素」,「職業表現」,「関連オブジェクト」,「職業表現＆関連オブジェクト」は，それぞれ，クラスタリングのために作成した文書ベクトルにどういった種類のタームを用いているかを表している.

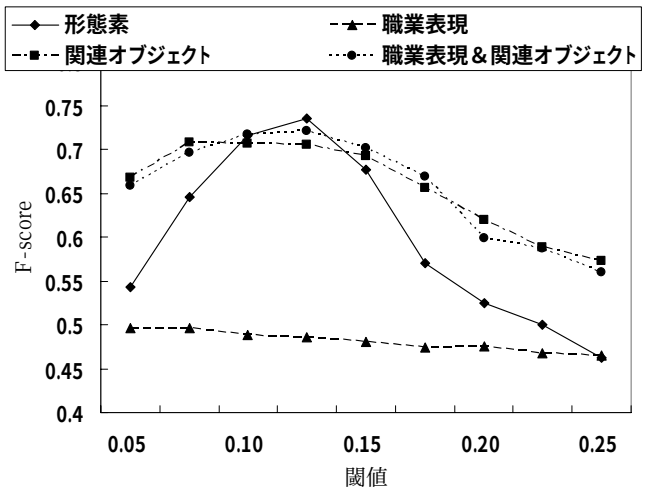

図 3: 閾値と F-score の関係

グラフから見て取れるように，最大値を記録したのは形態素を用いた方法であり，閾値 0.125 の時に F-score=0.735 を記録した. 同じ閾値で職業表現と関連オブジェクトを用いた手法の F-score は 0.721，関連オブジェクトのみを用いた手法の F-score は 0.705 であった.

関連オブジェクトを用いた手法が形態素を用いた手法を上回る事が無かった．抽出された関連オブジェ

クトには,「京都大学大学院情報学研究科」と「京都大学」,「京大」のように，同じ組織名を異なる表記で記述しているものが多く，これが関連オブジェクトが有効に働く範囲を狭めてしまったと考えられる.

職業表現のみの手法は F-score0.5 にさえ達する事は無かったが，これは職業表現を持つ文書が全体の 58%と少なかったため，クラスターを形成しかなかった文書が多かったからであると考えられる.

また，関連オブジェクト，関連オブジェクトと職業表現を用いた両方式は，F-score は閾値 0.075 から 0.15 の間で 0.7 付近の安定した値を記録しており，閾値 0.125 をピークとした山形を取る形態素のグラフとは異なる形となっている．これはつまり，これらの両手法は形態素のみを用いた手法と比較して，精度が閾値の変化に対して強いという事が分かる.

6.2. 形態素と関連オブジェクト･職業表現を合わせた分類

式(4)の係数 α の値を 0.5 から 5 まで 0.5 刻みで変化させることで，関連オブジェクトや職業表現の重みを変えて，クラスタリング精度の変化を確かめた.

図 4 は α=2.5 の際の分類精度のグラフである．図 4 中の 4 つの系列,「形態素のみ」,「形態素＆職業表現」,「形態素＆関連オブジェクト」,「全種類のターム」は，図 3 と同様に，クラスタリングのために作成した文書ベクトルにどういった種類のタームを用いているかを表している.「全種類のターム」については，形態素，職業表現，関連オブジェクトの全てを用いて作成した文書ベクトルである.

茶筌で抽出した形態素のタームに関連オブジェクトを合わせて作った文書ベクトルと，それに加えてさらに職業表現を合わせて作った文書ベクトルは，閾値 0.05 から 0.25 までの全区間で形態素のみで作成した文書ベクトルの分類精度を上回った．同様に α=2.0 の際も，その両方の手法が形態素のみを用いた手法を精度で上回っていた．形態素に関連オブジェクトや職業表現を組み合わせることで，閾値の影響が少なく，高い精度の分類結果を得ることが出来た.

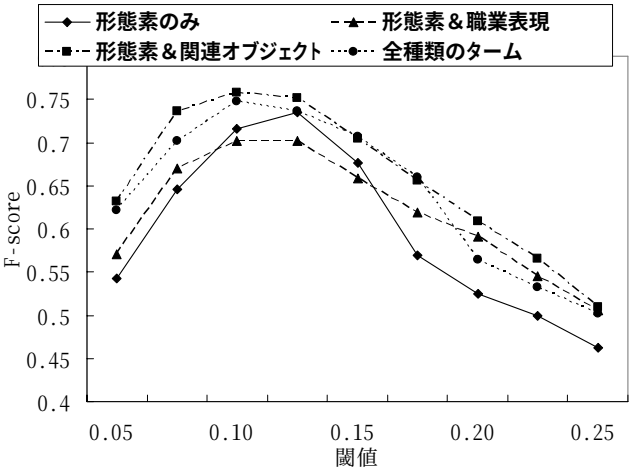

図 4: 閾値と F-score の関係（α=2.5）

クラスタリング精度の最大値に関しては，α=1.5，形態素と関連オブジェクトで作成した単語ベクトル，閾値 0.1 の時に F-score=0.795 を記録して，形態素のみを用いた手法の最大値 0.735 を大幅に上回る結果となった.

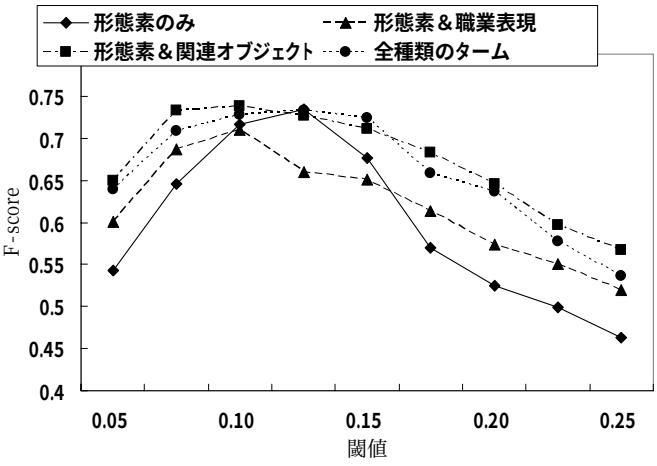

図 5: 閾値と F-score の関係（α=5.0）

図 5 はα=5.0 の際の分類精度のグラフである．係数αがかなり大きくなっているため，職業表現や関連オブジェクトへの重みが大きくなり，閾値の変化に対する影響がより少なくなっている事が分かる．クラスタリング精度の安定性を求めるのであれば，職業や関連オブジェクトへの重みを大きくしてやる事で実現できる.

7. まとめと今後の課題

本稿では，検索対象の人名と共起する組織名，人名，職業を表す表現を用いる事で，Web 検索の検索結果中に出現する同姓同名を分離する手法を提案した．また，

これらのタームを形態素解析によって抽出されたタームと合わせて用いる手法を提案し，評価を通して従来手法と比較して有効性を検証した．

今回の実験で抽出した組織名には，同じ組織を異なる表記で記述しているものが多く見つかったが，これを解消する事でさらなる精度の向上が望まれるため，課題として取り組んでいきたい．

提案手法と関連研究とのクラスタリング精度の比較も行なう必要があるため，これについても取り組んでいきたい．

ユーザがクラスター間の違いを簡単に見つけ，所望の人物の情報だけを簡単に手に入れられるようにする必要がある．今後は検索結果のクラスタリングだけでなく，クラスタリングされたグループを代表する様な単語の抽出も行い，図 6 のようにクラスタリングされた検索結果にラベル付けするという実装を行っていきたい．

また，Web 検索結果をクラスタリングしたものの提示に留まらず，グループ化されたそれぞれの人物の情報を，自動的に抽出という手法についても取り組んでいきたい．

8. 謝辞

本研究の一部は，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」,文部科学省科学技術振興費プロジェクト「異メディア・アーカイブの横断的検索･統合ソフトウェア開発」(代表:田中克己)，および，平成 17 年度科研費特定領域研究(2)「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」(課題番号：16016247, 代表：田中克己)によるものです．ここに記して謝意を表すものとします．

文　献

[1] S. Brin and L. Page.: “The Anatomy of a Large-scale Hypertextual Web Search Engine”, In *Proceedings of WWW7*, pp. 107-117, 1998.

[2] J. M. Kleinberg.: “Authoritative Sources in a Hyperlinked Environment”, *Journal of the ACM*, vol.46, no.5, pp. 604-632, 1999.

[3] Clusty the Clustering Engine, http://clusty.jp/

[4] SRC - Search Result Clustering Toolbar in Microsoft Research Asia, http://rwsm.directtaps.net/

[5] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma and J. Ma.: “Learning to Cluster Web Search Results”, In *Proceedings of SIGIR'04*, pp.210-217, 2004.

[6] R. Al-Kamha and D. W. Embley.: “Grouping search-engine returned citations for person-name queries”, In *Proceedings of WIDM’04, Washington*, DC, USA, pp. 96-103, 2004.

[7] X. Wan, J. Gao, M. Li and B. Ding.: “Person Resolution in Person Search Results: WebHawk”, In *Proceedings of CIKM’05*, pp. 163-170, 2005.

[8] 佐藤進也，風間一洋，福田健介，村上健一郎: “実世界指向 Web マイニングによる同姓同名人物の分離”，情報処理学会論文誌：データベース，Vol. 46, No. SIG 8 (TOD26), pp. 26-36, 2005.

[9] H. Toda and R. Kataoka.: “A search result clustering method using informatively named entities”, In *Proceedings of WIDM’05*, pp. 81–86, 2005.

[10] R. Bekkerman and A. McCallum. "Disambiguating web appearances of people in a social network", In *Proceedings of WWW'05*, pp. 463-470, 2005.

[11] O. Zamir and O. Etzioni.: “Web Document Clustering: A Feasibility Demonstration”, In *Proceedings of the 21st International ACM SIGIR Conferenc*e, pp.46-54, 1998.

[12] 関根聡: “固有表現から専門用語”，言語処理学会第 10 回年次大会(NLP2004) 「固有表現と専門用語」ワークショップ, 2004.

[13] 渡辺一郎，桝井文人，福本淳一: "固有表現抽出ツール NExT の精緻化とユーザビリティの向上"，第 10 回言語処理学会年次大会発表論文集，pp. 413-415, 2004.

[14] NTT コミュニケーション科学研究所監修，“日本語語彙体系”，岩波書店, 1997.

[15] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原 正幸: "日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書", http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1-j.pdf

[16] Google, http://www.google.co.jp/

[17] Y. Zhao and G. Karypis.: “Evaluation of Hierarchical Clustering Algorithms for Document Datasets”, In *Proceedings of CIKM’02*, pp. 515-524, 2002.

田中克己 Google 検索

該当の人物:3人[ピアニスト][研究][詩人・東洋史学者]

一人目:ピアニスト

ピアニスト**田中克己**のホームページ

プロフィール、コンサート情報、新譜情報、雑記帖。

www.oak.dti.ne.jp/~katsumit/ - 5k - キャッシュ - 関連ページ

ピアニスト**田中克己**のホームページ

Christmas Dream for You **田中克己** with 大島紀美江と仲間たち. クリスマスコンサート ということで、第1部を . **田中克己** ピアノトリオ 室内楽への誘い. ～伊藤亮太郎(ヴァイオリン)、平田 ... 金沢力 ワイピアノコンサート **田中克己**ピアノリサイタル ...

www.oak.dti.ne.jp/~katsumit/renew/Recitals.htm - 20k - キャッシュ - 関連ページ

二人目:研究

田中研究室 Tanaka Laboratory

田中研究室関連のインフォメーション. 21 世紀 COE プログラム「知識社会基盤構築のた め の情報学拠点形成」・全学共通科目(B 群)「情報と知財」講義アーカイブ(2005 年)・ 社会情報学フェア 2005(9 月 12 日～14 日開催)・第 3 回情報知財フォーラム ...

www.dl.kuis.kyoto-u.ac.jp/ - 9k - キャッシュ - 関連ページ

Satoshi OYAMA, Dept. Social Informatics, Kyoto Univ.

大島 裕明, 小山 聡, **田中 克己**. ``個人文書から抽出した語彙の意味関係に基づく Web 情報検索,'' 日本データベース学会 Letters ... 小山 聡, **田中 克己**. ``質問の階層 的 構造化を利用した Web 検索手法の提案,'' 日本データベース学会 Letters (DBSJ Letters), ...

www.dl.kuis.kyoto-u.ac.jp/~oyama/j-index.html - 16k - キャッシュ - 関連ページ

三人目:詩人・東洋史学者

田中克己文學館

田中克己 (たなかかつみ)1911-1992. 詩人、東洋史学者。「四季」「コギト」編輯 同人。 昭和初期の抒情詩復興期において、その硬質孤高の抒情が伊東静雄と並び称される。 「詩集西康省」「大陸遠望」等の詩集のほか、東洋史、ドイツ文学、支那古典文学の ...

libwww.gijodai.ac.jp/cogito/tanaka/tanaka.htm - 4k - キャッシュ - 関連ページ

図 6: 実装イメージ