Webの構造情報とプロファイル抽出を用いたオブジェクト識別

白砂　健一†　　小山　聡††　　田島　敬史††　　田中　克己††

† 京都大学工学部情報学科
〒 606-8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科 社会情報学専攻
〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{shirasuna,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　人名を質問として Web 検索を行う場合，検索結果には同姓同名だが実際には異なる人物の情報が含まれる場合がある．この問題は，オブジェクト識別問題の一種である．オブジェクト識別問題はクラスタリングを用いて解決されてきたが，クラスタリングアルゴリズムはデータの特徴ベクトルでの表現方法，データの類似度の定義，クラスタリングの良さを表す指標の定義などの構成要素がある．本論文では，Web 検索結果における人物の識別に対して，これらの技術の異なる組合せの比較と分析を行なう．さらに，オブジェクト識別の精度を向上させるために，Web の構造情報と人物に関するプロファイル抽出を用いる手法を提案し，評価実験を行う．
キーワード　オブジェクト識別, クラスタリング, データマイニング

Object Identification using Web Structure Information and Profile Data Extraction

Kenichi SHIRASUNA†, Satoshi OYAMA††, Keishi TAJIMA††, and Katsumi TANAKA††

† School of Informatics and Mathematical Science, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Infomatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{shirasuna,oyama,tajima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　When we use a person's name as a query for a Web search engine, information on different persons with the same name would be included in the search results. This is an instance of object identification problems, which have been solved by clustering algorithms. Clustering algorithms are composed of several techniques such as feature vector representations of data items, similarity measures, and clustering criteria. In this paper, we compare and analyze different combinations of these techniques for identification of persons in Web search results. We also propose and evaluate new techniques using Web structure information and profile data extraction to improve the accuracy of object identification.
Key words　object identification, clustering, data mining

1. はじめに

近年，情報源としての Web の重要性は増しており，さまざまな情報が Web を通して得られるようになった．Web 上には，従来の公的な情報から，会社情報，個人の情報，趣味の情報，その他娯楽など，さまざまな情報が存在するようになった．また，情報を発信することも容易となり，多くの人々が，自らの情報や興味のある事柄に関して，Web サイトや Blog で公開するといったことが日常的に行われるようになった．

このような背景から，近年では，Web 上から人物の情報を取り出すために，人名を質問として検索することがしばしば行われるようになっている．しかし，このような情報検索を行うときに問題になるのが同姓同名の問題である．例えば，「田中克己」という人名を Google で検索すると，検索結果の中には，大学教授，詩人，ピアニストといった異なる人物の情報が含まれてしまう．膨大な検索結果から，対象としている人物の情報のみを取り出すためには，各 Web ページが同一人物を指しているかどうかを判定し，分類する作業が必要となる．この問題

は，オブジェクト識別問題の一種である．

オブジェクト識別は，一般に人名を含むデータ（文書）を類似度に従ってクラスタリング [1], [2] することで行われる．クラスタリングの結果は，データの特徴ベクトルでの表現方法，データの類似度の定義，クラスタリングの良さを表す指標の定義によって，大きく異なる場合がある．

オブジェクト識別の問題は，文献データベース等のデータベースの分野においては従来から研究されてきたが [3]，Web における人物の同定の研究が報告され始めたのは最近のことであり [4]，異なる手法による実験結果の蓄積が充分であるとはいえない．

本論文の目的は，Web 上から収集した人名を含むデータに対して既存の手法の様々な組合せで実験を行い，手法の比較と分析を行うことである．また，特に Web 上の人名の識別に対して，より精度を改善するための手法について提案を行う．

以下，2. でオブジェクト同定で用いる既存の様々な手法について紹介し，3. でその手法についての実験を行う．その結果を元に，4. で我々の新たな提案を行い，5. でその提案を検証する実験を行う．6. で本論文のまとめと今後の方針について述べる．

2. 既存の手法について

2.1 データモデル

既存の手法では，文書を計算機で扱い易い形にするためのデータモデルには，以下のような，単語で区切り，各単語の出現数を１つの次元とする，特徴ベクトル形式が良く用いられているこのモデルは文書を単純化し，計算量を減らして高速化を図ることができるという点で非常に有効なモデルであるが，有用な情報も失ってしまう可能性があるという欠点もある．

特に Web 上に存在するデータに対象を定める場合，識別の対象となるデータが必ずしも単純なテキストデータであるとは限らない．Web 上において主流となる保存形式である HTML データの場合，段落・タイトル・引用などの各文章・段落の意味づけ，強調などのさまざまな装飾や，また表などのデータを持った文書, それに画像データやフラッシュも存在する．また，Web 上のデータはリンク構造が存在し，データ同士の関係が特に深いものなどが存在する．その他，Web 上のデータには通常のテキストなどとは違い，広告・メニューと言ったコンテンツとは直接関係ない文章群が存在したりするなど，さまざまな特殊な事情も存在する．Web に特化したオブジェクト識別としては，これらも考慮にいれるとより精度の良いオブジェクト識別が期待できる．

2.2 類似度計算

各文書に出てくる人名が同じ人物を指しているか否かは，通常，文書同士の類似度を用いて比較される．文書の類似性が高ければ，同一人物である可能性が高いということになるが，これはある意味でかなり大雑把な近似となっており，工夫の余地がある．

文書同士の類似度の比較において，もっとも単純なものにはコサイン相関値が考えられる．v_i，v_j を文書をベクトルとしたとき，

$$\cos(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i||v_j|}$$

しかし，この形では全ての単語に同等の重みを与えているため，類似度を比較する際，たとえば「ホームページ」や「リンク」といった，出現頻度が高いが識別の役には立たない単語が大きな重みをもってしまうことになる．このような問題を解決するため，いかに各ベクトルにその識別の有用さに応じた重みを付けるかということに焦点が当てられ，一般に以下のような方法がその解決策として用いられている．

- TF-IDF 法
- 教師付き学習を用いた方法

TF-IDF 法は，ドキュメントにおける単語の出現頻度（Term Frequency; TF）に，各単語の重み付けとして，その単語が全ドキュメントのうちいくつに出現しているか，の逆数（Inverse Document Frequency; IDF）を用いたものである [5]．これは文書同士の類似度の他，文書と文とのマッチ率を比較をする際に良く用いられる方法であるが，Web 上の文書を対象としたオブジェクト識別の場合，いくつか特徴的な面があるため，そぐわない部分も存在する．例えば，識別の際，識別対象文書全体を用いて IDF 値を計算した場合，もし，ある人物に深く関係があり良く出てくる単語があっても，その単語がありふれた単語として重みが低く計算されてしまう可能性があるといった不都合が起こるのではないかと考えられる．このため，TF-IDF 法を用いる場合，Web 全体での IDF 値を考えるなど，工夫する必要がある．

教師付き学習を用いた方法は，人手で同一人物に対する正解が与えられたデータから，各特徴の重みを学習する方法である [6], [7]．この方法は多くの数の正解を学習させることで，適切な特徴ベクトルの重み付けを得ることができ，高い精度が期待できる優れた方法である．ただし，事前に解を教師学習させる必要があるため，今回の問題にそのまま利用することはできない。事前に別の手法で確度の高い部分解を得て，それを教師として用いるなど，いくらかの工夫が必要である．

2.3 クラスタリング

類似度を計算した後は，ベクトル間の距離の近いもの同士を１つにまとめる，クラスタリングを行うことで同一人物かの判定を行う [8]．

クラスタリングの手法には，大きく分けて，最短距離法などの階層的手法と，k-means 法などの，分割最適化手法が存在する．前者の階層的手法は，n 個のデータが与えられた時，初期状態を各データ 1 個ずつを構成要素するクラスタ n 個とし，各クラスタ間の距離の近いもの同士を結合していく手法である．この手法ににおいて，クラスタ間の距離を定める方法によって，さまざまなバリエーションが存在する．

- 最短距離法：最も近い点同士の距離をクラスタ間距離とする
  $D(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$

- 最長距離法：最も遠い点同士の距離をクラスタ間距離とする
  $D(C_1, C_2) = \max_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$
- 群平均法；各点間の平均距離をクラスタ間距離とする
  $D(C_1, C_2) = \frac{1}{n_1 \cdot n_2} \Sigma_{x_1 \in C_1} \Sigma x_2 \in C_2 D(x_1, x_2)$
- ウォード法：クラスタ同士を合併したときに増える分散の量をクラスタ間距離とする
  $D(C_1, C_2) = E(C_1 \bigcup C_2) - E(C_1) - E(C_2)$
- セントロイド法（重心法）：各クラスタの重心（平均ベクトル）間のキョリをクラスタ間距離とする

ただし，C_1，C_2 をクラスタとする．

分割最適化手法は，分割の良さ (クラスタリングの優秀さ) を与える評価関数を定め，その評価関数を最適にする分割を探索するものである．代表的な手法には k-means 法があり，これは各クラスタの重心点間の距離の二乗の和を評価関数とし，k 個のクラスタに分割するものである．探索は山登り法で行い，局所的な最適解しか得られないため，ランダムに初期値を変更しながら評価関数を最小にするものを選ぶ．

このようにクラスタリングには基本的な手法から距離を定める関数，優秀さを定める関数など，さまざまなバリエーションが存在する．

2.4 評価手法

分類を行った後は，正解の分類結果との比較を行うことで，分類の精度の評価を行うことができる．評価計算式にはさまざまな種類がある．以下に 2 つを挙げる [9].

- F-Measure
- FScore

FScore はノードに着目した手法である．各正解クラスタに対して，対応する分類結果クラスタを 1 つ決定し，その一致率を取ったあと，正解クラスタの大きさに応じたその一致率の重さ平均を取る．計算式は以下のとおりである．

$$V_{FScore} = \Sigma_{c \in C} \frac{|m_c|}{|m|} \max_{r \in R} \frac{2m_{r,c}}{|m_r| + |m_c|}$$

ただし

C：正解クラスタ集合	m_c：クラスタ c 中の文書集合
c：正解クラスタの 1 つ	m_r：クラスタ r 中の文書集合
R：分類結果クラスタ集合	$m_{r,c}$：m_r と m_c の積集合
r：分類結果クラスタの 1 つ	m：全正解クラスタの全文書集合

エッジ間の正確さに着目するのが F-Measure である．各ベクトルをノード, ベクトル間係をエッジと見たとき，全エッジ数のうち，同じクラスタ同士を正しくつないでいるエッジの割合を計算する．計算式は以下のとおり．

$$V_{FMeasure} = (Precision * Recall * 2/(Precision + Recall))$$

ただし

$$Precsion = e_{r,c}/e_r$$

$$Recall = e_{r,c}/e_c$$

e_r：各分類結果クラスタの文書間のエッジ
e_c：各正解クラスタの文書間のエッジ
$e_{r,c}$：e_r と e_c の積集合

3. 既存手法に関する実験

まず，既存手法による分類の精度を知るために，幾つかの実験を行った．

幾つかの人名を質問とした Google の検索結果上位 100 件を抜き出し，それを手動で分類して正解を作成した．そして，各手法を実装したクラスタリングプログラムによってその 100 件の分類を行い，正解と比較することで精度を求めた．

3.1 サンプルによる違い

まず行ったのは，もっとも単純な手法である．各文書について，文書中の単語の出現頻度に基づく特徴ベクトルを求め，それぞれの間の距離をコサイン相関度によって求めた後，最短距離法を距離尺度としてクラスタ作成を行った．正解と比較する上での評価尺度には，F-Measure を用いた．検索語は「田中克己」「佐藤雅彦」などの 9 種を用意した．各人名の実際の構成例を図 1 に示す．

結果は図 2，図 3 のとおりである．

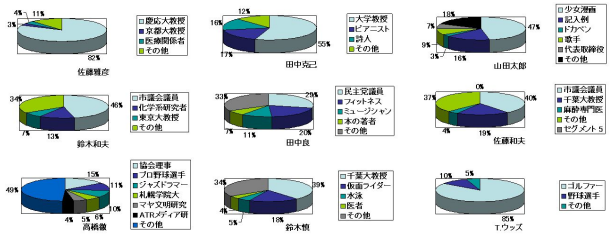

図 1　各検索人名の構成

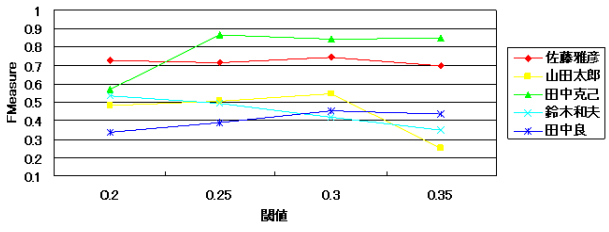

図 2　サンプル 9 種の比較 1

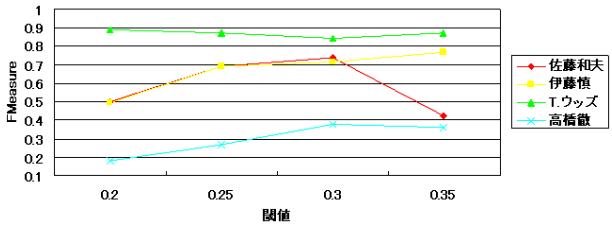

図 3　サンプル 9 種の比較 2

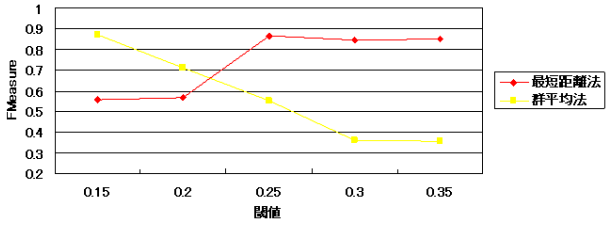

図 4　クラスタリング手法の比較 1(田中克己)

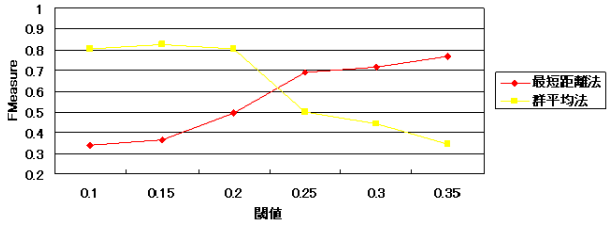

図 5　クラスタリング手法の比較 2(伊藤慎)

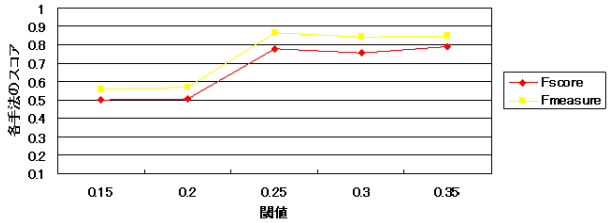

図 6　評価手法の比較 1(田中克己)

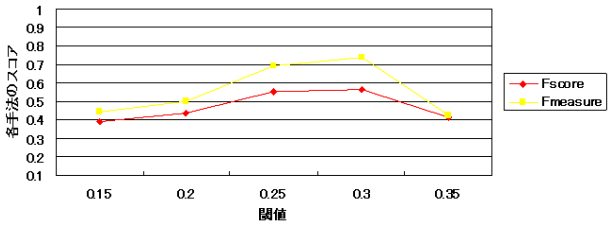

図 7　評価手法の比較 2(佐藤和夫)

サンプルによって，精度は大きく異なる結果となった．

3.2　クラスタリング手法による違い

次に，クラスタリング手法による違いをみるため，クラスタ間距離尺度を，最短距離法と群平均法法の 2 種類に変えて結果を観察した．

図 4 は「田中克己」，図 5 は「伊藤慎」の場合の結果である．もっとも最適な結果となる閾値はアルゴリズムによって異なるものの，最良解はほぼ等しくなるという結果が出た．

3.3　評価手法による違い

最後に，評価手法による数値の違いを観察した．図 6 は「田中克己」，図 7 は「佐藤和夫」の場合の結果である．

値は少し異なるものの，数値の上がり下がりとしてはほぼ同じ傾向を示した．

3.4　IDF に関する実験

IDF を利用することで単語への適切な重み付けができるかどうかを調査するために, 実験を行った．

表 1 は Web から人名で検索した際，上位 100 件の文書群に含まれていた単語のうち, 一部を取り出して (i) 100 件のうち，何件にその単語が入っていたか (ii) Google でその単語を検索した際，何件の文書がヒットするか（≈ Web 全体に何件，その単語を含む文書が存在するか）を示したものである．

表 1　IDF 手法の比較

キーワード	IDF 数	Google ヒット数
赤穂	1	1,130,000 件
静的	6	1,040,000 件
ひろか	1	63,000 件
インストラクター	1	2,430,000 件
権藤	1	131,000 件
蘇	3	2,730,000 件
人情	1	3,100,000 件
データベース	48	13,100,000 件
霊園	1	895,000 件
溝口	3	1,030,000 件
前身	1	2,050,000 件
不易	1	125,000 件
七曜	1	57,100 件
ステップ	11	4,290,000 件

傾向として (i) は多いものはありふれた単語と区別に有用な単語の 2 種が存在した．また単語のほとんどが出現文書数が一桁であった．(ii) はほとんどの単語が数百万件程度であり，たまに人名が数十万件程度になる程度で, 各単語に大きな違いが見られず，利用できるかは疑問であるという結果となった．

3.5　ま　と　め

以上より，次のことが観察された．

- 既存の手法では，サンプルにより，大きく精度が異なってくる．サンプルによっては，あまり精度が良くないものも数多く存在する．
- クラスタリング手法は最良の部分では大差は出ないようである．
- 評価手法を変えても，数字の大小は変わっても，傾向はあまり変わらない．
- IDF 手法を，特徴ベクトルの重み付けにそのまま利用することは難しい．単語の出現数だけでは，とても重要な単語かありふれた単語かの判別はできない．

既存の手法では良い結果が得られなかった具体例を挙げる．サンプル内に，二人の同姓同名の建築家がいて,「登録建築家ホームページ」の紹介ページにおいて，それぞれのプロフィールがそれぞれ別のページに公開されていた．二人は同姓同名であるが，プロフィールによると誕生日，住所，写真等全てが異なるため，人目では容易に別人であると判別することができた．

しかし，これは既存の Web ページ同士の類似度を求める手法では，これが書式・用語が似通っていることから，非常に高い類似度を与えてしまうため，同一人物と判定してしまっていた．このような既存の手法の限界を改善するための手法に関する実験を，次の章で示す．

4. 提 案 手 法

4.1 特徴ベクトルモデル

「文書が似ている＝出現する人物が似ている」というモデルを改めることを考える．単純に文書全体の比較ではなく，人物に関する部分を文書から抜き出し，それを比較することで分類することを目指す．

4.1.1 プロファイル情報

まず考えられるのが，その個人のプロファイルデータである．特に，Web ページに掲載されていやすい，個人を特定できる情報には以下のようなものがある．

- 職業
- 生年月日
- メールアドレス
- 電話番号
- 知人
- 著書・作品など
- 業績など

これらは自然言語解析などを行わなくても，ある程度は抜き出すことが可能である．例えば，職業は人名の直後に現れていることが多いといった規則がある．また，特に著名人などの場合，これらの情報や略歴などのプロファイル情報がまとまって詳しく書いてあることも少なくない．

プロファイルデータは，個人にとって一意である情報がほとんどであることから，識別する上で強力な情報となりやすい．しかし，誤った部分をプロファイルとして抜き出すと，大きく識別を誤ってしまうことから，慎重な取り扱いが必要であると考えられる．

4.1.2 背景となる単語

文書中に出てくる単語の中には，生年月日・職業・メールアドレスといった，完全に個人を特定できる「プロファイル」とは言えないまでも，ある程度人物の対象を絞ることができる単語が存在する．本人の仕事や趣味とかかわりの深い単語などで，例えば本人の職業が野球選手の場合，(本人が所属しているかにかかわらず）阪神タイガースといった球団名，また，アウト・盗塁・グランドスラムといった野球用語である．同一人物かどうかの識別の場合においては，同姓同名の他者との区別ができれば事足りるため，これらのような単語も識別に十分有用であるといえる．これらに特別な重みを与えることができると，精度が上げることが可能である．ただし，これらを取得するには，オントロジなど，やや大規模な辞書が必要である．

4.1.3 文章と人名との相関の深さ

ある文書の中に人名が出てくる場合において，文章全体が本人に関係のある内容であるとは限らない．その本人に関する記述は全体のうちの 1 段落のみで，他は関係のない内容かもしれないし，全てが本人に関係がある内容かもしれない．また，Web 上の文書には独特の事情があり，閲覧のユーザーインターフェース部（メニューなど），商業広告などといった，通常の文書には存在しない，本文そのものとなんら関係のない内容も存在する．本人に関係のない内容を識別の際に組み込むと，関係のある内容の重みが薄れ，また関係のない内容が識別を惑わす恐れがある．そこで，これら，“ 文章と本人との相関の深さ ”を考え，重み付けを適宜行うことで，精度向上があげることができると考えられる．

4.2 Web のリンク構造を用いた手法

Web 独自の構造を用いることで，識別の補助とし，精度を上げる手法を考える．リンク構造を用いたアプローチがある．

- 同一 Web サイト内にある文書
- リンク関係でつながっている文書

このような文書は内容的に相関関係が高く，同一人物を指している可能性が高いと考えられる．ただし，どこまでが同一 Web サイトを判定するのはやや難しく，また，Web サイト単位でのリンクを検出するのには非常に時間がかかる．

通常，同一ドメイン内であれば，高い確率で同一 Web サイトであると考えることができる．しかし，これには少なからず例外も存在する．以下のような例がある．

- 無料ホスティングサービス
- ニュースサイト
- オンラインショップ（特に人名が出てくるケースは，書籍販売など.）

また，単に同一 Web サイトか，同一でない Web サイトか，では表現しにくいものも存在する．１つの Web サイト内に，内容の関連のある，内容が完結している Web ページ群が存在する場合などである．

その他,「.ac.jp ドメインが共通しているので双方は類似した人物（大学関係者)」といった判定も考えられる．

これらを踏まえると，単純に同じ・異なるの判定を行うのではなく，近い・遠いといった距離の概念を用いて URL 同士の関係を表すのが適切であると考えられる．これにより，URL 同士の関連の深さを数値化し，類似度を求め，識別に用いることができる．

4.3 分 類 手 法

上記のような手法には，さまざまな傾向がある．例えば，全てのデータ間距離について結果を示すかわりに，ベクトル間距離についてゆるい改善を行うものもあれば，ごく一部のデータ間についてしか示せない代わりに，一致性について確度の高い結果を与えるものも存在する．よって，これらの手法を複合的に用いる場合，その特徴を生かす必要がある．例を示す．

例えばプロファイル情報を用いた手法は，一致した場合，高い確度で同一人物であると考えられる．ただし，100%ではない．そこで，各手法によって得られた結果を１つにまとめ，それを新たな“ 特徴ベクトル ”とする．このとき，各手法の確度によって，その要素に重み付けを行う．これにより，この新たな特徴ベクトル間に距離を定義することが可能となり，各手法の良さを生かして複合的に用いることができる．

また，確度の高い手法により部分解を得た後，それを教師と

した教師付き学習による分類を用いることも考えられる．教師付き学習は部分的な解を与えることで，特徴ベクトルの重み付けを高い精度で行うことができ，手法の特長を組み合わせていることができていると言える．

5. 提案手法に関する実験

5.1 プロファイル抽出に関する実験

4.1.1 で述べたとおり，プロファイル情報を抽出し，それを比較することで，より識別の精度を上げることができるということを示すために，次のような実験を行った．

まず，どの程度，文書からプロファイルを抽出できるかを調べた．文書からプロファイル情報を抽出する手法として，ヒューリスティックな評価をいくつか考案し，それを行うことで抽出した．

- 人名の直後に，助詞などをはさまずに載っている単語は，プロファイル情報である可能性が高い．(特に（）にはさまれている場合.)
- 生まれ，略歴などのプロファイルがまとまって掲載されているページが存在するので，それを検出する．そのようなページの特徴としては，タイトルが検索語であることが多い,「名前」,「所属」,「生年月日」,「略歴」,「メールアドレス」,「住所」,「連絡先」などプロファイルを示すキーワードが入っていることが多い，などが挙げられる．

1 つ目のヒューリスティックを用いて抽出した結果の例は次のとおりである．プロファイルだと思われる単語などをかなり取ることができた．

しかし，現時点ではこれらの単語はプロファイルと呼べないものも存在し，また，例えば“ 筑波大教授 ”と“ 教授 ”を同一かどうかの判定を行うなどといったことは難しい．そこで，とりあえずこの単語が現時点でどの程度利用できるかを調べるために，次のような実験を行った．プロファイル単語を単純に文字列比較し，全く同一のものが他の文書のプロファイル単語に存在するかどうか調べる．そして，存在した場合，その 2 つの文書が同一かどうか調べる．(ただし，最低限「氏」「さん」「夫妻」など一般的な人名に付く単語は取り除く.)

この実験を行った結果，次のようなことが観察された．

- 同一のプロファイルを持つ文書のペアは，ある程度存在するがそれほど多くはない．
- 同一のプロファイルを持つ文書のペアは，ほぼ全て同一人物であった．

検索語数	検出された単語
4 個	法 (前)
1 個	校長 (後)
4 個	著 (後)/著 (後)/著 (後)/著 (後)
4 個	さん (後)/子ども (後)
17 個	作品 (後)
18 個	筑波大 (後)/筑波大 (後)/筑波大 (後)/東邦大 (後)
2 個	大阪府議会議員 (前)
2 個	著 (後)/著 (後)
3 個	大阪府議会議員 (前)

表 2　プロファイル単語の例

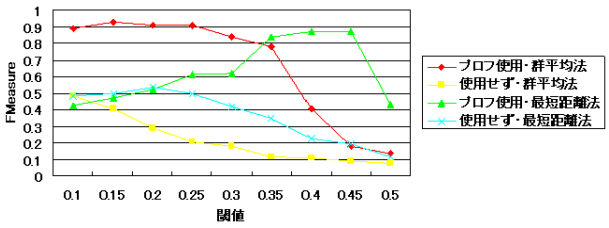

図 8　プロファイル情報を用いて同定を行った場合の精度の変化 (鈴木和夫)

そこで，この情報を用いることで，現時点でどの程度オブジェクト識別の精度が上げられるかという実験を行った．プロファイルを示す単語が一致したものに対し，以下のような計算式を適用し，類似度を底上げした．

$$Sim = Sim_{original} * (1 - \alpha) + \alpha$$

ただし，α はパラメータで，今回は 0.3 とした．

図 8 は，プロフ単語を用いたときと用いなかったときの精度の比較の例である．図は検索人名「鈴木和夫」，評価手法には，F-Measure を用いた例である．

全体の傾向として，やや精度を改善する程度のものがほとんどであり，大きく精度を改善するものは少し，また逆に精度をわずかながら下げてしまう例も存在した．プロファイル単語が一致するケースは数が少なく影響が限定的であることなどが理由として考えられる．

5.2 文章と人名の相関の深さ

次に，提案 4.1.3 を確かめるために，実験を行った．文章と人名の相関の深さを，HTML タグの段落的深さによって求め，それによって各単語の重みを定める．

- 質問として入力した人名の近く（同じ段落）に現れる単語は，特に人名に関係のある重要な単語である．
- 質問として入力した人名が現れる段落と同じ深さの段落に現れる単語は，人名に関連する単語である．
- 質問として入力した人名が現れる段落より浅い段落に現れる単語は，人名に関係のない単語である．

これらのヒューリスティクスを用いることで，どの程度精度が改善されるか調査した．具体的に，各単語について，通常の単語の重みを 1 としたとき，関係の深い単語の重みを α，関係のない単語の重みを $\frac{1}{\alpha}$ とした．

図 5.2 に，精度の変化を表す例のグラフを示す．図は検索人名が「田中良」の例である．提案手法が精度を改善していることがわかる．

全体の傾向として，精度をやや改善していた他，閾値の変化に対し精度がやや安定的であった．全体的な各ベクトル同士の位置関係を改善させているためと考えられる．

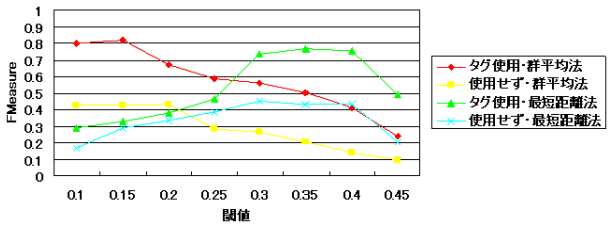

図 9　段落情報の使用とその精度の変化 (田中良)

5.3　Web のリンク構造を用いた手法

4.2 で提案した手法について，実験を行った．

まず，URL の類似性が，判定に利用できるかどうかを調べた．各 URL について，ドメインが同じものの類似度を 1 と定義し，以後ディレクトリについて，浅いものから一致するたびに 2，3，4 … と増やしていくとする（例：図 5.3）．このとき，各類似度について，どの程度の割合で同一であるかどうか調べた．

http://www.hogehoge.com/abc/aaa.html
http://www.hogehoge.com/bbb.html　類似度1
http://www.hogehoge.com/abc/bbb.html 類似度2

図 10　類似度の例

実験の結果の例を以下に 2 つ示す．(表 3，表 4）図の一致率は，分母はその類似度の URL の組み合わせがいくつあったかを示し，分子はそのうちいくつが同じクラスタに所属していたかを示している．

結果について，URL の類似度が 1 以上存在するものは，高い確率で同一人物である，という結果が得られた．

次に，この情報を用いることで，オブジェクト識別の精度が上げられるということを示す．URL の類似度が 1 以上存在した式に対し，以下のような計算式を適用し，類似度の底上げを行った．

$$Sim = Sim_{original} * (1 - \alpha) + \alpha$$

ただし α はパラメータであり，今回は 0.3 を用いた．

結果の例を図 11 に示す．この例では検索語に「鈴木和夫」を用いている．URL 情報を用いることによって，精度が大幅に改善されていることがわかる．

傾向として，最短距離法より群平均法の方が優れた結果が出ることが多かった．最短距離法では，ごくまれに存在する，URL が似ているが同一人物ではない例に対して大きく反応してしまい，精度がそれに影響されてしまいがちであった．群平均法はそのような誤情報に対し安定的であるからと考えられる．

また，全体の傾向として，大きく精度を改善する例が 3 割，

表 3　URL の類似度とクラスタの一致性 1(田中克己)

類似度	一致率
4	2/2
3	97/97
2	211/211
1	44/44
0	1432/4399

表 4　URL の類似度とクラスタの一致性 2(佐藤雅彦)

類似度	一致率
7	1/1
4	1/1
3	141/141
2	7/7
1	58/59
0	3217/4447

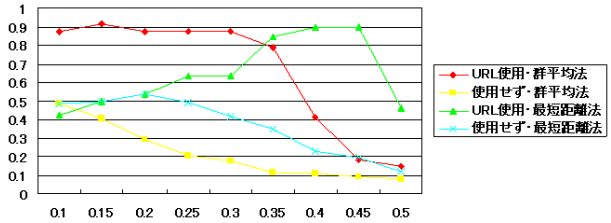

図 11　URL 類似度を用いた場合の精度の変化 (鈴木和夫)

やや精度を改善する例が 5 割，逆に精度をわずかながら下げてしまう例が 2 割程度存在した．もともと高い精度で識別されているものは，それほど精度向上につながらなかったほか，URL 情報がマッチしていない情報がいくつか存在する場合には，かえって混乱してしまうことになると考えられることから，類似度の計算式を適切に設定し，その影響をできるだけ限定的にする必要があると考えられる．

6.　今後の課題

現在の課題とともに，今後の方針を示す．

6.1　プロファイル抽出による手法について

今回，プロファイル単語を抽出するために用いた手法は 1 つだけであり，文章中のプロファイル単語のごく一部を用いた比較しか行うことができなかった．他のヒューリスティックも用いて，より多くのプロファイル単語を抽出することを，まず第一の課題としたい．

また，今回は，取得したプロファイル単語同士の照合について，単純な文字列比較しか行うことができなかった．しかし，より発展的には，例えば「京都大学教授」と「大学教授」について比較・照合するといったことが考えられる．

今回，実際に取得できた単語には，以下のような分類が考えられた．まず，この分類を行うなど，一歩一歩確実にこなしていきたい．

- プロファイルデータ.「教授」「議員」など.
- 一般的に人名に付く修飾語.「氏」「さん」「夫妻」など.
- 本人の名前を冠した，別の固有名詞.「○○作品集」「○○詩集」など.
- 並列に並べられた，人名.「○○ ×× △△ 共著」における ×× など.
- 著者・作者・サブタイトル などの理由で，人名が単独で出ている.

最後に，3 つの手法に共通していえる課題として，手法の効

果を最大限に引き出すための適切なパラメータ設定が挙げられる．今回行った実験では，時間の都合上あまりパラメータを調整することができなかったが，適切なパラメータを選ぶことで，より手法の効果を引き出すことができる可能性がある．試行錯誤を重ねることで，より手法の性能を引き出すことを課題としたい．

6.2 人名と段落の関係の深さについて

この実験については，技術的な課題が多かったので，それを改善し，より厳密な実験を期すことを今後の課題と考えている．また，この実験では，非常にヒューリスティックな重み付けを行ったが，単語と検索人名との距離関係などを用いることで，より細やかな重み付けを行うことが考えられる．重み付けの式をしっかりと定式化し，パラメータを変化させ実験を行うことを目指したい．

6.3 Web の構造を用いた手法について

本実験では，対象のファイルのみを扱い，またその URL のみしか扱わなかったが，さらなる課題として，Anchor タグを用いる，またリンク先のファイルといった別のファイルにまで踏み込んだ情報を利用する，などといったことが挙げられる．

また，本実験では，二つの文書が同一 Web サイト内にあるかの簡単な判定手法として，URL の類似度を用いたが，より詳しく解析することで，より精度の高い判定を行う研究も存在する [10]．URL を用いる上での，今後の参考としたい．

6.4 手法の組み合わせ

本実験では 3 つの手法を提案したが，そのそれぞれ手法単体での性能実験は行ったものの，それらを組み合わせた実験はまだ行っていない．組み合わせたときの性能評価とともに，最適なパラメータ組み合わせの発見に力を入れたい．

6.5 システムの実装

本論文では，3 つの手法の提案と，その有効性を検証する実験を行ったが，その手法を実装した，オンラインから利用可能な人物の同定を行った後結果を表示するまでのシステムを構築するまでにはいたっていない．今後の方針としては，実際に第三者が利用可能な状態にして，より多くのユーザから意見を聞くことができる状態にすることを目指したい．

7. おわりに

本論文では，Web における人物の識別を対象として，既存の手法および新たな提案手法の比較実験を行った．

既存手法に関する実験の結果では，最良の結果を生じるパラメータ領域ではクラスタリング手法による違いは少ないが，パラメータの変化に対する安定度が異なった．また，特徴ベクトルおよび類似度の定義の違いによる精度への影響が大きかった．クラスタリング実験の評価尺度である FMeasure と FScore は，ほぼ同じ傾向の評価を示した．

さらに，それらの実験を元に，3 つの提案とその検証実験を行った．プロファイル情報を用いる手法では，取り出せることができたプロファイル単語は不完全で，また内容も多様なものがあったが，精度を一定量向上させることができた．人名と文章の関係度を利用した手法では，非常にヒューリスティックな手法ではあったが，緩やかな精度向上が見られた．URL による Web 構造を用いた手法では，URL が類似したデータは特に同一人物である傾向があることが示され，3 つの手法のうちでも特に精度を上げることができた．また，特にサンプル間で提案手法の効果が異なり，精度が向上しやすいサンプルやしにくいサンプルがあるなど，興味深いさまざまな傾向が観察された．

このように，本論文では精度の向上に一定の結果が得られたが，現時点ではより詳しく行うべき課題や不完全な点が数多く存在し，まだまだ満足できる状態とはいえない．6. で述べた通り，現在抱えている多くの問題を検討するとともに，課題と今後の方針に基づき，より発展させていきたい．

謝辞

本研究の一部は，平成 17 年度科研費若手研究（B）「参照の同一性判定に基づく複数 Web ページの検索閲覧方式の研究」（課題番号：16700097，代表：小山聡）および，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「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）および，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」によるものです．ここに記して謝意を表します．

文　献

[1] David J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. MIT Press, 2001.

[2] Boris Mirkin. *Clustering for Data Mining: A Data Recovery Approach*. Chapman & Hall, 2005.

[3] Mikhail Bilenko, William W. Cohen, Stephen Fienberg, Raymond J. Mooney, and Pradeep Ravikumar. Adaptive name-matching in information integration. *IEEE Intelligent Systems*, Vol. 18, No. 5, pp. 16–23, 2003.

[4] Ron Bekkerman and Andrew McCallum. Disambiguating web appearances of people in a social network. In *Proceedings of the 14th international conference on World Wide Web (WWW2005)*, pp. 463–470, 2005.

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[6] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pp. 39–48, 2003.

[7] Satoshi Oyama and Christopher D. Manning. Using feature conjunctions across examples for learning pairwise classifiers. In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004)*, pp. 322–333, 2004.

[8] 神嶌敏弘. データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！－. 人工知能学会誌 vol.18, no.1, pp. 59–65, 2003.

[9] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clusting. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22, 1999.

[10] Wen-Syan Li, Okan Kolak, and Hajime Takano Quoc Vu. Defining logical domains in a web site. In *ACM Hypertext 2000*, pp. 123–132, 2000.