実世界指向 Web マイニングによる同姓同名人物の分離

佐藤進也[†] 風間一洋[†]
福田健介[†] 村上健一郎[††]

巨大なデータベースである Web から知識を抽出する一手法として実世界指向 Web マイニングを提案する．従来のマイニングでは主に統計的な処理によりデータの特徴が抽出されていた．これに対し，実世界指向マイニングでは，実世界を意識したデータの解釈，具体的には，実世界のエンティティがデータの中にどのように現れ，相互にどういう関係を形成しているかを調べる．この考え方を Web における人物の識別に適用し，同姓同名人物の分離を行った．これは，与えられた人名が出現する Web ページを同一人物ごとにグループ分けするタスクで，本手法を用いた場合，平均 9 割以上の高い率で正しく処理できることを確認した．

Distinguishing between People on the Web with the Same First and Last Name by Real-world Oriented Web Mining

Shin-ya Sato,[†] Kazuhiro Kazama,[†] Kensuke Fukuda[†]
and Ken-ichiro Murakami[††]

This paper proposes a technique called “real-world oriented Web mining” for extracting knowledge from the Web regarded as a huge database. While conventional mining techniques search for characteristics of data mostly by statistical analysis, the proposed technique interprets data from real-world oriented point of view. In more concrete terms, it locates real-world entities in the data and analyzes relationships among them. This idea has been applied for performing a task to distinguish between people on the Web with the same first and last name. The task is to classify Web pages with a given person’s name into groups each of which corresponds to a person in the real world. With the proposed technique, people have been identified with accuracy more than 90% on average.

1. はじめに

Web ページの数は 2004 年 4 月の時点で 92 億を超えると推測されている[1)]．この巨大さは Web の 1 つの特徴であるが，それ以上に特筆すべき性質として“社会性” — つまり，実社会の状況を反映していること — があげられる[2),4)]．これは，Web が普及し日常生活に浸透してきたことの自然な結果としてとらえることができる．Web は，この社会性ゆえに，実社会に関する多種多様な「活きた」情報を保持している知識ベースとして期待されている[4)]．

実世界に関する知識を Web ページに記述されている語句やハイパーリンクといったデータの集まりから取り出す方法は次の 2 つに大別される：

- データに潜んでいる特徴的パターンを発掘し，それを実世界に照らし合わせて解釈するボトムアップ的アプローチ．データマイニングの手法を Web に適用した Web マイニング[5)] がその典型例．
- あらかじめ知識（の表現）の枠組みを用意し，データをその枠組みにあてはめて整理・解釈するセマンティック・ウェブ[3)] のようなトップダウン的アプローチ．

これらを，得られる知識の質や知識抽出に要するコストなどの観点で評価すると，まず前者は，マイニングという手法の特徴として，隠れた知識の発見が期待できるという点で優れている．しかし，抽出されたデータそのものの特徴が果たして実世界を説明しうるものなのか，実世界をどれだけ正確にとらえているのかは不明であり，妥当性の点で検討の余地がある．

一方，後者では，あらかじめ設計された知識の枠組みの中で得られる結果なので妥当性は高い．しかし，得られる知識も与えられた枠組みの制約を受けるので，

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories
†† 法政大学ビジネススクールイノベーション・マネジメント研究科
Hosei Business School of Innovation Management

限定的である．たとえば，タグで意味付けをする場合，タグの種類の少ない/多いがそのまま概念分類の粗さ/細かさにつながる．さらに，枠組みの構成（たとえば，オントロジーの構築）にコストがかかるという問題もある．

このように，2 つのアプローチは相補的関係にある．本論文では，これらをお互い欠点を補い合うように融合させた新しいアプローチを提案する．具体的には，解析対象となるデータの中でも実世界を構成する要素（人，組織など）に焦点を当て，それらの関係などをマイニング的手法により明らかにする．マイニングをベースとしながら，実世界に関する知識の枠組みを適用するという意味で，本手法を実世界指向マイニングと呼ぶ．

実世界指向マイニングを具体的に説明するとともにその有効性を示すため，以下，Web における人物の識別という問題を考える．Web に限らず，一般に文書中に人物を登場させるにはその名前を表す文字列を記述すればよい．しかし，逆の対応は一意的でなく，文書中の人名を実世界の人物に対応させるためには同姓同名☆人物を分離するという問題を解かなければならない．本論文では，この問題を解決する手段として実世界指向マイニングが有効であり，Web 上の実データを 9 割を超える高い精度で処理できることを示す．

以降，本論文での議論は次のようにすすめる．まず，2 章で Web における同姓同名分離問題を定式化する．次に，実世界指向マイニングの考え方を適用して得られるこの問題の解法を 3 章で示す．4 章では，実データを用いて実際に同姓同名分離処理を行い，提案手法の有効性を確認する．この結果をふまえ，5 章で関連研究との比較を行い，実世界指向マイニングというアプローチの意義を明確にする．

2. 同姓同名分離問題の定式化

人物とその名前，そして名前が記述された文書（Web ページ）の関係は**図 1** のようになっている．複数の人物により 1 つの名前（図では X）が共有され，また，その 1 つの名前は複数の文書に出現している．

同姓同名人物は，当然のことながら，名前という文字列だけでは分離不能である．しかし，文書中においては，文脈によりその文字列がどの人物を指し示しているかを判別することが可能になる．

1 つの文書に同じ人名が複数回出現する場合，それらが異なる人物を指し示す可能性は否定できない．し

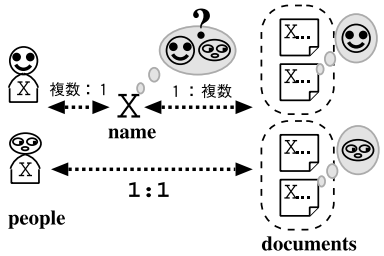

図 1 人物，名前，文書の相互関係
Fig. 1 Relationships among people, their names and relevant documents.

かし，実際には，基本的に 1 つの文書は一貫した文脈を持っており，その中で人物が言及されるので，複数回出現する同じ名前が異なる人物を指し示す可能性は非常に低いことが実験を通して経験的に分かっている．名前のリストのような明確な文脈が認められない場合でも，1 文書にすべての名前が載せられているのではなく，何らかの属性に基づいて複数のページに分割されている．分野別の書籍一覧（の一部としての著者リスト）がその例である．

そこで，本論文では，1 つの文書内に複数回現れる名前が同 1 人物を指している状況を仮定する．このとき，各文書を個々の人物に対応付けることができる．そして，同じ人物に対応する文書をまとめることで，文書集合と人物を 1 対 1 に対応させることができる．同姓同名人物の分離はこの文書集合の構成に帰着される．この対応関係は，形式的には次のように書くことができる．

まず，人名 x を含む Web ページの集合を $D(x)$，Web ページ d に出現する x が指し示す（実世界の）人物を $p(d,x)$ とする．そして，名前が x であるすべての人物の集合を $P(x)$ とし，その要素に番号を付与する：

$$P(x) = \bigcup_{d \in D(x)} \{p(d,x)\} = \{p_1,\ p_2,\ \ldots\}$$

このとき，文書集合

$$C_i = \{\ d \mid d \in D(x),\ p(d,x) = p_i\}$$

を人物 p_i に対応させる．ここで得られた文章集合群 $\{C_i\}$ を，以下，Ω と呼ぶことにする．

なお，厳密性を期して，1 つの文書に複数回出現する名前が異なる人物を指し示す可能性を考慮する場合には，Web ページ d において x で指し示される人物の集合を $\tilde{p}(d,x)$ とすると，以下のように人物と文書集合の対応を与えることができる：

$$C_i = \{\ d \mid d \in D(x),\ \tilde{p}(d,x) \ni p_i\}$$

☆ ここでは，名前の読みではなく表記が同一の場合だけを考える．

3. 実世界指向マイニングによる解法

3.1 基本方針

文書と人物の対応付け $p(d,x)$ を得ることは d において x が言及されているコンテキストを理解することであるが，この処理を計算機上で実現するのは困難である．そこで，$p(d,x)$ の値を利用するのではなく，$D(x)$ の分類として Ω を構成する方法を考える．

文書分類の方法としては，文書ごとに特徴的な語（特徴語）を抽出し，クラスタリングや機械学習を適用するものが一般的である．語の特徴語としての妥当性は，主にその文書における出現の統計的特徴で評価される．たとえば，特徴語の評価尺度としてよく使用される tf・idf は，語の文書内出現頻度などに基づいて計算される．

本論文においても基本的には，文書を特徴付けし，その特徴に基づいて分類する，というアプローチをとる．ただし，これらの具体的処理は従来手法に従わず，データ（Web ページ）と実世界の対応関係を考慮して行う．

3.2 実世界を意識した文書の特徴付け

3.2.1 実世界と Web の対応付け

従来手法が文書を統計的に処理してその特徴を抽出するのに対し，本論文では実世界を意識した特徴抽出を行う．そのために，まず Web による情報提供・利用の状況を整理しておく．

Web は情報流通の場であると同時に情報利用実験の場としての側面も有しており，いままでに様々な情報利用法☆が試みられてきた．その影響で（典型的な）利用形態も少しずつ変化してきたが，Web 上で提供される情報は基本的に個人や組織などの活動に付随して生産されるものであることは一貫して変わりはない．そして，大雑把なとらえ方ではあるが，ある人物が Web ページに登場するということは，その人物の当該活動への（間接的なものも含めた）関与を示していると解釈できる．これは，実世界の活動（あるいは活動のための組織や場）に対応して Web サイトが用意され，それに関与する人物の名前が Web サイト中のページに出現しているというよく知られている事実の言い直しにすぎないが，ここに，実世界における活動の場と Web サイト，活動の主体である人とページ上の人名という対応関係を見出すことができる（**図 2**）．

3.2.2 人と活動の場の関係を利用した同姓同名分離

実世界では，活動の場を介して人間関係が形成され，

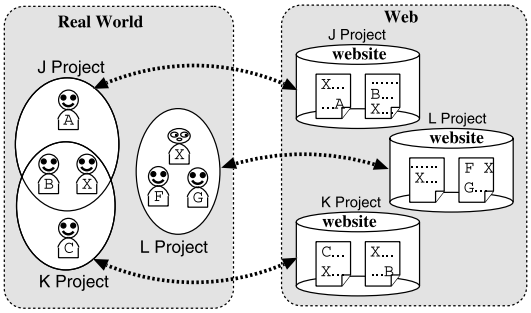

図 2 実世界と Web の対応付け
Fig. 2 Correspondence between the real world and the Web.

逆に，その人間関係によって活動の場の間につながりが生まれる．つまり，2 つの活動の場の関連性はそれらに共通して登場する人物の存在により示されると考えられる．たとえば，図 2 では，プロジェクト J とプロジェクト K（以降,「プロジェクト」を略す）という 2 つの活動の場は X と B という人物を共有しているが，J と L，K と L の間には共通する人物はいない．よって J と K には関連性があり，J と L，K と L には関連性がないと推測される．

図 2 はさらに，この実世界の状況が Web にも反映されている様子を示している．J のサイトと K のサイトには X と B という人名がともに現れているが，J と L，K と L の間には X 以外に共起する人名がない．

この Web における人名の出現状況を利用すると，X という名前を持つ人物の同一性は次のように推定できる．まず，J と K における B の共起は J と K の関連性を示しており，関連のある活動の場に出現する X は同一であると推測される．一方，共起する人名が存在しない J と L，K と L には関連性が認められず，よって，J および K の X と L の X とは別人であると推測できる☆☆．

この例を一般化することにより，同姓同名人物を分離するアルゴリズムが得られる．ここでのポイントは，(1) 個々のページ単位で特徴付けをして類似性の高いものどうしをグループ化するのではなく（活動の場に対応する）文書群を単位として特徴抽出し類似性を判定することと，(2) 文書群の特徴として，そこに出現する人名を用いることである．実世界の活動の場と区別するため，対応する文書群をワークスペースと呼ぶことにする．このとき，名前 x を持つ同姓同名人物を分離のアルゴリズムは次のとおりである：

☆ たとえば，近年では Weblog を例にあげることができる．

☆☆ 人名 X の出現は，それが同一性判定の対象であるので，プロジェクト間の関連性を与える根拠から外す．

(i) $D(x)$ の要素をワークスペース $w_i(i = 1, \ldots)$ ごとにまとめる．
(ii) 各 w_i から人名を抽出する．
(iii) 人名の共起に基づき，関連のある w_i どうしをまとめ Ω を構成する．

3.2.3 ワークスペースの構成方法

(i) では，ワークスペースを構成する方法として，URL の表記が類似している Web ページをグループ化するアルゴリズム[8] を用いる．その具体的な手順は以下のとおりである．

(1) 与えられた人名を含むすべての Web ページ（を指し示す URL）の集合を U とする．
(2) URL の集合を要素とする集合 W を以下のように初期化する：
$$W = \{\{u\} | u \in U\}$$
(3) W の 2 つの要素 w_1，w_2 に対して，ある $u_1 \in w_1$ と $u_2 \in w_2$ が存在して，u_1，u_2 が置かれているディレクトリ階層の隔たりがたかだか 1 である場合には，w_1 と w_2 を統合する．たとえば，
$$w_1 = \{\texttt{http://a.jp/x/v.html}\}$$
$$w_2 = \{\texttt{http://a.jp/x/y/z.html}\}$$
であったとき，W から w_1 と w_2 を削除し，新しい要素 w_{new} を追加する：
$$w_{new} = \{\texttt{http://a.jp/x/v.html}, \texttt{http://a.jp/x/y/z.html}\}$$
(4) 統合されるものがなくなるまで (3) を繰り返す．

ここで最終的に得られた W の要素をワークスペースとする．

この方法は処理が単純で計算量が少ないという点で優れている．しかし，その一方で，本来なら 1 つのワークスペースとして統合されるべきページ群が複数のワークスペースに分解されたり（細分化），当該人物に関する他者による記述がワークスペースとして抽出されたりすること（二次的な情報☆の混入）がある[8]．それぞれの例としては，あるプロジェクトが提供する活動内容などの情報が年度単位で分割される場合や，新聞や個人の日記などで当該人物が取り上げられている場合があげられる．

よって，ある人物が主体的に活動している場を正確に抽出するという目的に対してこの方法を適用するのは妥当ではない．しかし，同姓同名人物分離というタスクにおけるワークスペースの役割は，与えられた人物と関連人物を結び付けることであり，本来の意味での活動の場を正確に抽出するという厳密性は必ずしも求められていない．むしろ，以下に述べるように，上記の特性がタスクの目的を達成するうえで効果的に働く場合がある．

まず，一般論として，ワークスペースが多くのページを含めば，そこには多くの人名が出現することが期待できる．しかし，同時に，同姓同名の複数の人物が混在する可能性も高くなる．一方，本アルゴリズムでは，関連性のあるワークスペースどうしは最終的に (iii) において統合されうるので，(i) の時点でワークスペースを細分化して抽出することは複数同姓同名人物の混在を抑えるための有効な戦略であるといえる．

また，二次的情報も本来のワークスペースと同様に関連性のある人物どうしを結び付ける手がかりとして有用である．特に，ドラマの主人公などの実在しない人物は二次情報により原作者や出演者などとの結び付きを与えられる．これにより，その主人公が 1 人の（仮想的な）人物として識別され，同名の実在人物との分離が可能となる．

3.2.4 ワークスペースの特徴抽出

(ii) では，まず，各ページの内容を形態素解析し，連続して出現した姓と名をつなげて人名とする．得られた人名 y を，その特徴語としての妥当性を示す関数 $f(y)$ でランキングし，1 ワークスペースあたりたかだか n_{max} 個の人名を選ぶ．本論文では，$f(y)$ として次の 3 種類を用いる：

- $f_{tfidf}(y)$
  w_i を 1 つの文書と見なしたときの語 y の tf・idf 値．
- $f_{psr}(y)$
  y の出現に関するページとサーバの比率（Page-to-Server Ratio）．語 y を含むページを持つ Web サーバの集合を $S(y)$ とすると
  $$f_{psr}(y) = \log|D(y)| / \log|S(y)|$$
  で与えられる．なお，$D(y)$ は 2 章で定義した，人名 y を含む Web ページの集合である．**図 3** は $|D(y)|$ と $|S(y)|$ の関係を示したグラフだが，普通名詞（黒丸）では $\log|D(y)|$ と $\log|S(y)|$ はほぼ比例関係にあることが分かる．一方，人名（白抜き四角）はその比例関係からはずれている．$f_{psr}(y)$ はそのはずれの度合いを示す数量で，語 y の偏在性を表していると考えられる[8]．
- $f_{tfpsr}(y)$
  f_{psr} に w_i を 1 つの文書と見なしたときの y の出現頻度（tf）を掛け合わせたもの．

☆ 当該人物の活動そのものではなく，それを参照・利用しているという意味で “二次的” という表現を用いている．

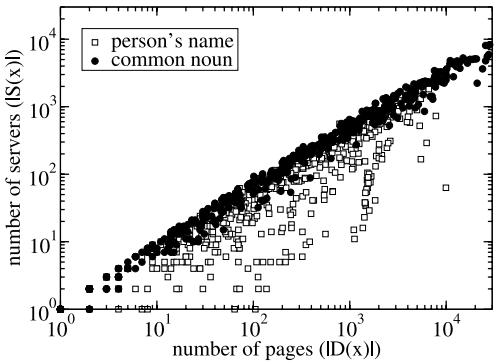

図 3 語の出現ページ数とサーバ数の関係
Fig. 3 Number of relevant servers versus number relevant of pages for a term.

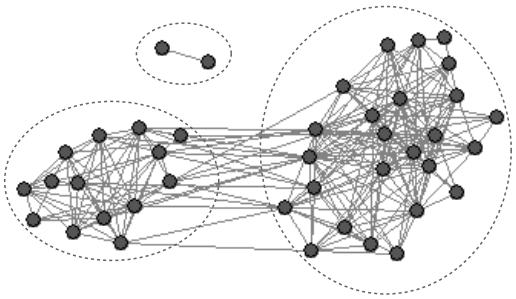

図 4 ワークスペースの相互関係を示す $G(x)$ の例
Fig. 4 Example of $G(x)$ showing relationships among workspaces.

アルゴリズムの最後のステップである (iii) の具体的な方法については，次節で述べる．

3.3 グラフ構造に基づくワークスペースの分類

3.3.1 ワークスペースの相互関係

ワークスペースをノード，ワークスペース間での人名共起の関係を無向リンクで表すと，ワークスペースの相互関係を表すグラフが得られる．なお，リンクには共起する人名の数で重み付けする．以下，人名 x から得られるこのグラフを $G(x)$ と書くことにする．$G(x)$ の例として，$x =$“江川卓” の場合を**図 4** に示す．グラフの描画は Fruchterman-Reingold のアルゴリズム[9] を用いて自動的に行ったもので，破線の楕円は後から人手で追加した．

異なる人物が属するワークスペースにまたがって同じ（名前を持つ）人物が現れることがなければ，グラフの連結成分をそのまま Ω の要素とすることができる．しかし，複数の話題を含むページなどでワークスペースの主テーマと直接関係のない人物が言及されている場合もあり，この条件は必ずしも満たされない．実際，図 4 のグラフは 2 つの連結成分から構成されているが，“江川卓” が出現する個々のページに実際にアクセスして調べた結果，3 人の異なる人物が確認された．

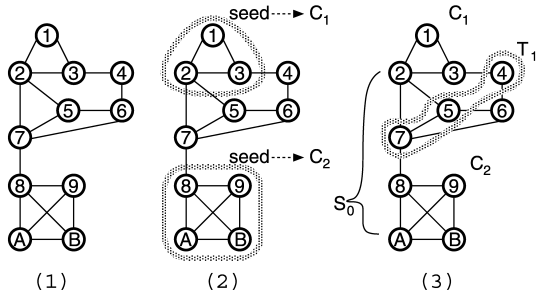

図 5 グラフ分割アルゴリズムの適用例
Fig. 5 Applying the graph decomposition algorithm to a graph.

この例では，3 人の人物はそれぞれ図 4 中の破線で囲った部分，すなわち，上方にある 2 ノードからなる連結成分と，下方の連結成分の左右にある相互に密につながっている部分に対応している．このことから，異なる人物に対応するワークスペースのグループ（Ω の要素）はグラフ中の稠密な部分に対応していると推測される．この仮説が正しければ，グラフを稠密な部分に分解することで同姓同名人物を分離できる．

3.3.2 稠密な部分グラフへの分解

グラフを稠密な部分に分解する既存の方法としては betweenness に基づくクラスタリング[6] をあげることができる．betweenness とは，簡単にいえば，各リンク☆の “ノードどうしを取り持つ” という役割に注目したときの重要さを測る尺度であり，グラフを構成する全ノードの相互接続状況から導かれる．その尺度を利用したクラスタリングは，やはり，グラフの大域的性質（たとえば，対称性の有無など）の影響を受ける．

一方，ワークスペースの相互関係は人と人とのつながりに由来しており，基本的に，大域的な構造ではなく局所的なつながりが意味を持つ．そこで，本論文では，既存の方法を利用するのでなく，ワークスペースの相互関係解析に適した新しい方法を提案する．

本方法の基本的な考え方は次のとおりである．まず，グラフ $G(x)$ の（各連結成分の）中で特に稠密な部分をシードとして選び出す．そして，その他の部分は，複数あるシードのうち最もつながりの強いものを選んでグループ化する．この具体的な手順（**I**〜**IV**）を，**図 5** の (1) のグラフを例にとり説明する．

☆ ノードの betweenness というものもリンク同様に考えることができる．

I. まずはじめに，シードとしてクラスタ係数[7] が 1 であるノードとそのリンク先を選ぶ．シードが複数ない場合（0 個を含む）には，その連結成分全体を 1 つのシードとする．シードが複数存在する場合には，まず，互いに近接しているシード，すなわち，リンクでつながれたシードどうしをまとめて 1 つのシードとする．このようにして得られたシード群に適当に番号を振り，$\{C_i\}(i = 1, \ldots M)$ としておく．

図 5 では，(2) に示したのがシードに対応する部分である．上部にはノード①，②，③からなるシードが存在し，下部にはノード⑧，⑨，Ⓐ，Ⓑからなるシードが存在する．下部のシード中には，実際には，クラスタ係数が 1 であるノードが 3 つ（⑨, Ⓐ, Ⓑ）存在している．いずれも，それらノードとそのリンク先からなるノードの集合は {⑧, ⑨, Ⓐ, Ⓑ } で，3 つのシードが完全に重なり合っており，結果的に，これらは 1 つのシードとして統合される．上部，下部のシードに属するノードの集合を順に C_1，C_2 とする．

II. 次に，ワークスペースの集合 S_i，T_i を，

$$S_i = S_{i-1} \cup T_i$$
$$T_i = \{w | w \text{ は } G(x) \text{ のノードで } S_{i-1} \text{ からの距離が } 1\}$$

として順次構成する．ここで，w と S_i の距離とは，これらの $G(x)$ 上の最短経路長のことである．S_0 はシードを構成する全ワークスペースの集合とする．

図 5 の例では，(3) に示したように，C_1 と C_2 を合わせたものが S_0 である．そして，S_0 より距離が 1 だけ離れているノードの集合 $\{④, ⑤, ⑦\}$ が T_1 である．

III. II で得られた T_i のそれぞれの要素 w を $\{C_j\}$ のいずれかに追加していく．C_j の選択には，以下の数量を用いる．

$$q(w, C_j) = \sum_{c \in C_j} l(w, c)$$

ここで，$l(n_1, n_2)$ は 2 つのノード n_1，n_2 を結ぶリンクの重みであり，リンクが存在しないときは 0 とする．w は $q(w, C_j)$ が最も大きい C_j を選んで，そこに追加する．これは，最も関連性の高い C_j を選ぶことに相当する．

図 5 の場合，T_1 の要素である④，⑤，⑦それぞれを，C_1 と C_2 のうちより関連性が高い方に帰属させる．④，⑤については，C_1 にのみつながっているため，C_1 に帰属させる．一方，⑦については，C_1 と C_2 双方につながりがある（②-⑦, ⑦-⑧）ので，C_1 と C_2 のうちいずれかつながり（リンクの重み）の強い方を選ぶ．

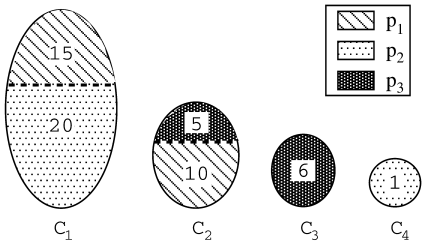

図 6 分離された状態の例
Fig. 6 Schematic example of decomposition.

IV. II，III を T_i が空になるまで繰り返す．その結果得られた $\{C_i\}$ が，各人物に対応するワークスペースの集合である．

4. 評　　価

本章では，前章で示した同姓同名分離法の有効性を検証する．

4.1 評価尺度

まず，分離の正確さを測るために，認識率 r，不純度 ι，分散度 δ という 3 種類の尺度を導入する．

いま，同姓同名人物分離の処理を行った結果，$\Omega = \{C_i\}$，$(i = 1, \ldots, M)$ と M 人の異なる人物が認識され，一方，実際には p_j，$(j = 1, \ldots N)$ という N 人の人物が存在しているとする．また，a_{ij} を，C_i に帰属するワークスペースで人物 p_j が登場しているものの数とする．

図 6 は分離された状態の例を図式的に示したものであり，各数字はワークスペースの数を示している．たとえば，C_1 の上方の数字は，C_1 に属するワークスペースで人物 p_1 が登場するものの数，すなわち a_{11} は 15 であることを示している．

4.1.1 認　識　率

さて，

$$a_{i*} = \sum_{j=1}^{N} a_{ij}$$
$$a_{*j} = \sum_{i=1}^{M} a_{ij}$$

と定義すると，a_{i*}，a_{*j} はそれぞれ，C_i に帰属するワークスペースの数，p_j が登場するワークスペースの数である．

p_j に対しある i があって，

$$\frac{a_{ij}}{a_{i*}} > 0.5 \text{ かつ } \frac{a_{ij}}{a_{*j}} > 0.5$$

が満たされるとき，p_j は C_i によって認識されているといい，$p_j \leftarrow C_i$ と書くことにする．定義から明らか

なように，各 p_j を認識できる C_i はたかだか 1 つしか存在しない．p_j のうち，認識されうるものの割合，すなわち，

$$r = \frac{1}{N}|\{p_j|^{\exists}C_i, p_j \leftarrow C_i\}|$$

を認識率とする．図 6 の例では，$p_2 \leftarrow C_1$，$p_3 \leftarrow C_3$ であるが，p_1 を認識する C_i は存在しないため，$r = 2/3$ である．

4.1.2 不純度と分散度

認識率は同姓同名分離性能をマクロにとらえる指標である．本項では，性能の詳細を示すミクロな指標として不純度 ι と分散度 δ を導入する．

図 6 の分離例では，p_2 は C_1 により認識されているものの，(1) C_1 内部の p_1 の混入，(2) p_2 の C_1 外部（C_4）への分散という 2 点で改善の余地がある．不純度と分散度は，それぞれ (1)，(2) の状況を数量的に示すものである．

数量化にあたってはエントロピーの考え方を適用する．(1) の C_i 内に複数人物が混入する問題では，混入の量（全体に対する割合）と混在種別の多さを，混入割合の分布 $\{a_{ij}/a_{i*}\}$ のエントロピー

$$H_{i*} = -\sum_{j=1, a_{ij}\neq 0}^{N} \frac{a_{ij}}{a_{i*}} \log \frac{a_{ij}}{a_{i*}}$$

によって把握する．

いま仮に，C_i には k 人の人物が混在しており，混入の割合がすべて等しいとすると，$H_{i*} = \log k$ となる．そこで，H_{i*} の代わりに

$$e^{H_{i*}} - 1$$

という数量を考えると，混入のないときは 0 になり，k 人が等しく混在すると $k-1$ となる．これは，余分に混入している人数を表していると考えられる．この値の i についての平均

$$\iota = \frac{1}{M}\sum_{i=1}^{M} e^{H_{i*}} - 1$$

を不純度とする．

分散度 δ についても同様に考え，以下のように定義する．

$$H_{*j} = -\sum_{i=1, a_{ij}\neq 0}^{M} \frac{a_{ij}}{a_{i*}} \log \frac{a_{ij}}{a_{i*}}$$

$$\delta = \frac{1}{N}\sum_{j=1}^{N} e^{H_{*j}} - 1$$

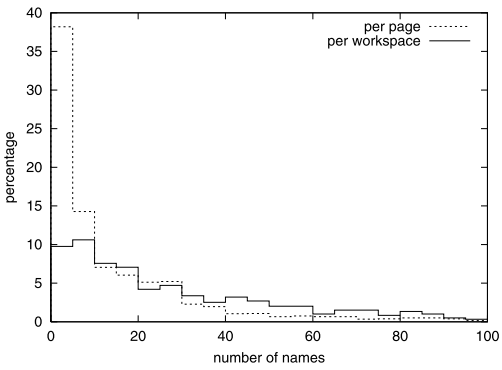

図 7 ワークスペース化による効果
Fig. 7 Effects of workspaces.

4.2 結果と考察

本節では，3 章で提案した同姓同名分離法を評価する．評価には 20 の人名を用い，人名 x が出現する Web ページの集合 $D(x)$ は，あらかじめ Web ロボットにより収集した約 5 千万ページ☆から x を含むものを抜き出して構成した．

C_i と p_j の対応付け，すなわち，各ページに出現する名前が実世界のどの人物に対応するかは，ページを実際に見て判定した．この判定作業は手間がかかるものなので，人名の選択においては非常に有名な人物を避け，作業者が有する知識で判定しやすいものであることを考慮した．

4.2.1 ワークスペースの抽出

20 の人名それぞれに対して 3.2.3 項で述べたアルゴリズムを適用した結果，全部で 621，1 つの人名あたり 31 のワークスペースが得られた．

図 7 は，これらのワークスペースにおける人名の出現数の分布を，ページあたりの出現数の分布と比較したものである．ページあるいはワークスペースあたりの出現人名数を横軸にとり，当該ページ/ワークスペースの全体に占める割合（百分率）をヒストグラムで示した．ここで，出現数とは異なる人名の数のことであり，延べ数ではない．

ページあたりの人名出現数分布は冪分布に近く低頻度が支配的である．出現数の平均は 48.3 であるが，この平均に満たないページが 8 割以上を占め，10 以下のものが半数以上にのぼる．一方，ワークスペースあたりの人名出現数の平均は 166.31 であり，その分布にはページの場合のような極度な偏りがない．つまり，ページをワークスペースにまとめあげることには，出現人名数の平均値を向上させるだけでなく，分布の低

☆ 収集は，2003 年 7 月に主に jp ドメインを対象に行った．

頻度への集中を緩和し，人名共有関係の形成を促す効果があることが分かる．

3.2.3 項で議論したように，複数のページを 1 つにまとめると，そこにはより多くの人名が期待できるが，その一方で，同姓同名の異なる人物が混在してしまう可能性も高くなる．今回の評価実験におけるこの問題の発生状況を調べたところ，混在が認められたのは 621 のワークスペースのうち 3 つ（0.48%）であり，本論文で採用したワークスペースの構成法では混在の発生率が低く抑えられていることを確認した．なお，これらはすべて人名のリスト（書籍リストに現れる著者）であった．

異なる人物が混在した場合の評価尺度の計算については，4.1 節における定義「a_{ij} を，C_i に帰属するワークスペースで人物 p_j が登場しているものの数とする」をそのまま適用する．たとえば，C_i に属するワークスペース w に p_j，p_k が混在しているものがあった場合には，a_{ij}，a_{ik} それぞれに w を計上する．

4.2.2 分離性能

人名それぞれに対して，分離性能を 4.1 節の尺度で評価した結果を**表 1** に示す．表中の WS にはワークスペース数を，r には認識率を「認識された人物数/認識すべき人物数」のまま約分せずに記してある．さらに，ワークスペースの統合状況を示す数量として，$|\Omega|/|P(x)|$ の値を表中の Ω/P の欄に示した．なお，ワークスペースを特徴付ける人名の最大数 n_{max} は 100，ランク付けのための関数には f_{tfpsr} を用いた（3.2.2 項）．ワークスペースによっては人名の出現数が少ないものもあり，特徴語としての人名の数は平均すると 1 ワークスペースあたり 48.6 であった．

この結果は，提案手法が全般的に優れた同姓同名分離の性能を持っていることを示している．認識率は 8 割の人名で 1 であり，1 に満たなかった残りの 2 割についても，(l) を例外として，0.66 を上回っている．全体の認識率の平均は 0.91 であった．

(l) は完全に認識を失敗した例である．野間佐和子氏は講談社の社長であるが，同時に，多くの組織の重要な役職☆に就いている．これらの組織はワークスペースとして抽出されているが，野間氏の名前が出現する文脈☆☆ではこれらの組織間に人名共起に由来する関連性を見出すことができなかった．

一方，(b)，(o)，(s) では，ドラマや漫画の登場人物などの仮想的人物と実在の人物とがほぼ正しく分離されている．仮想的人物の同一性を識別するためには，登場人物どうしの関係だけでなく，それを演じる俳優や原作者との関係が利用されていた．

表 1 分離性能
Table 1 Decomposition performance.

人名	WS	r	ι	δ	Ω/P
(a) 伊庭幸人	5	1/1	0.0	0.65	2.0
(b) 上田次郎	32	2/3	0.09	0.0	0.67
(c) 江川卓	46	3/3	0.17	0.17	1.0
(d) 木下和彦	16	7/7	0.0	0.0	1.0
(e) 栗原はるみ	34	2/2	0.0	0.35	2.5
(f) 五斗進	5	1/1	0.0	0.0	1.0
(g) 五嶋みどり	57	1/1	0.0	0.55	5.0
(h) 新垣紀子	13	2/2	0.0	0.0	1.0
(i) 竹内郁雄	58	3/3	0.0	0.07	1.67
(j) 田中克己	70	12/16	0.23	0.34	1.0
(k) 中村紘子	25	4/4	0.0	0.07	1.25
(l) 野間佐和子	17	0/1	0.0	8.67	11.0
(m) 野村紀子	6	5/5	0.0	0.0	1.0
(n) 畑村洋太郎	68	1/1	0.0	0.61	7.0
(o) 菱沼聖子	6	2/2	0.0	0.38	1.5
(p) 福原愛	67	1/1	0.0	1.30	10.0
(q) 三浦麻子	16	4/5	0.22	0.0	0.8
(r) 水野晴郎	31	1/1	0.0	2.79	7.0
(s) 山岡士郎	20	2/2	0.0	1.56	3.5
(t) 和田英一	29	6/6	0.0	0.0	1.0

不純度についてはいずれの人名においても低く抑えられている．分散度もおおむね低くなっている．ただし，なかには (i) のように分散度が比較的大きな値を示しているものもある．これは,「水野晴郎」という同一の人物が映画評論だけでなく歌舞伎評論など複数の領域で活動しており，それらが別人によるものとして認識されてしまうためである．

その一方で，(a) の「江川卓」の場合には野球選手時代の活動と引退後の芸能活動が同一人物によるものであると正しく認識されている．(i) と (a) の違いは，活動間の関係の一般性にある．(i) では，複数領域にまたがる活動が個人固有のものであるのに対し，(a) では野球界引退後に芸能界で活動することは「江川卓」という人物に限らず一般的に行われている．つまり，(a) では 2 種類の活動の場にまたがって「江川卓」とともに登場する人物が少なからず存在しているのである．その結果，これらの活動の場は相互関係を表すグラフ $G(x)$ 上で密に結び付けられ，そこに出現する「江川卓」は同一人物であると判断される．

このように，本手法は，人物や活動の場の相互関係に関する実世界の普遍的な知識（ルール）をデータから汲み取り自然なかたちで人物の認識に利用することができる．これは，高い認識率とともに特筆すべき本手法の特徴である．

☆ 国際文化フォーラム会長，共用品推進機構理事，日航財団評議委員，読書推進運動協議会会長，日本写真家協会理事，など．
☆☆ 役員の紹介など．

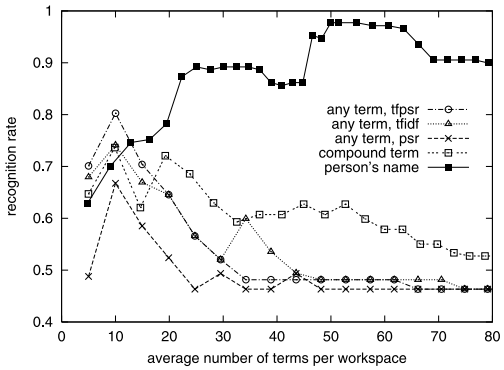

図 8 特徴語数と認識率
Fig. 8 Recognition rates versus number of characteristic terms.

4.2.3 人名による特徴付けの効果

さて，文書の集まりであるワークスペースを特徴付ける方法として，提案手法では人名という特別なカテゴリに属する語のみを選んで使用したが，次にこの効果を確認する．

いま，ワークスペースの特徴語を選択する基準として

(i) 任意の語で f_{tfpsr} の値が大きいもの（any term, tfpsr）
(ii) 任意の語で f_{tfidf} の値が大きいもの（any term, tfidf）
(iii) 任意の語で f_{psr} の値が大きいもの（any term, psr）
(iv) 任意の複合語で f_{tfpsr} の値が大きいもの（compound term）
(v) 人名で f_{tfpsr} の値が大きいもの（person's name）

の 5 つを考える．そして，選択基準とワークスペースあたりの特徴語の最大数 n_{max} をそれぞれ選び，上記の 20 の人名について同姓同名分離処理を行い結果の違いを比較する．具体的には，選択基準ごとに，n_{max} を変化させたときのワークスペースあたりの特徴語数の平均と認識率の平均の関係を調べる．その結果をまとめたものが**図 8** である．

どの選択基準のグラフにも，ある特徴語数（の範囲）で認識率がピークに達し，語数がそこから離れるにつれて認識率が低下するという共通した傾向がある．これは，語数が少なければワークスペースを十分に特徴付けることができないが，逆に語数を増やしすぎると一般的な語が混ざり込みやはり適切に特徴付けされないためである．前者は分散度の大きい状態であり，後者は不純度の大きい状態である．

選択基準を認識率のピーク値とその位置について比較すると，提案手法に対応する基準 (v) とその他の基準との間に大きな違いが認められる．基準 (v) のピーク値は 5 基準中最大であり，これは人名を用いることが同姓同名分離の性能を向上させるために効果的であることを示している．

人名の特徴としては高い偏在性（図 3）や，姓と名を組み合わせた複合語であることがあげられる．そこで，人名の代わりにこれらの性質を持つ語を用いることも考えられる．しかし，選択基準の (i)，(iii) および (iv) の結果はこの可能性を否定している．

5. 関 連 研 究

1 章では，Web というデータ集合からの知識抽出法をボトムアップとトップダウンに大別してその簡単な比較を行い，実世界指向マイニングをその融合として位置付けた．提案手法の詳しい内容とその効果が明らかになったところで，もう一度関連研究を整理し本手法の意義を明確にしておく．

5.1 トップダウン的アプローチ

トップダウンなアプローチとは知識の枠組みをあらかじめ用意しデータをそこにあてはめて解釈するものであるが，意味的な枠組みを文書に対して直接あてはめるものに大域的修飾（GDA）[10] がある．GDA ではテキストのタグ付けにより文書の意味構造を明らかにする．これにより，意味的検索などの正確さが大幅に向上すると期待されるが，タグ付けにコストがかかるという問題がある．その対策としては，コンテンツ作成者のタグ付けのインセンティブを高めることが試みられている[10]．

知識体系（カテゴリの集合）をあらかじめ用意し，文書をそのカテゴリに分類するテキスト分類[11] もトップダウン的アプローチの 1 つである．分類の方法としては現在機械学習を利用するものが主流になっている．同姓同名分離問題は文書分類に帰着可能なので（2 章），この問題を解く手段として機械学習によるテキスト分類の利用が考えられる．ただし，当然，人物を同定する以前にその人物に関する知識（学習のためのデータ）を得ることはできないので，分類先のカテゴリとして個々人を選ぶことはできない．代わりに，職業のような人物の属性に関するものの分類☆や（Web）ディレクトリのような一般的な情報の分類を利用することも考えられるが，その分類が人物を分離するためのカテゴリとしてつねに妥当である保証はない．たとえば，4.2.2 項の「江川卓」の例では野球と芸能に関

☆ 日本標準職業分類など．

する文書が同じカテゴリに分類されるべきだが，一方で，野球と芸能を分離したいケースも当然ありうる．これは，固定的な枠組みで実世界の複雑な状況を把握し解釈することは困難であることを示している．

5.2 ボトムアップ的アプローチ

そこで，実世界の状況をデータのありようをもって語らしめる，データ（あるいは Web）マイニングの手法が有望視される．従来のマイニングでは主に統計的な処理によりデータの特徴が抽出されていた．これに対し，実世界を意識したデータの解釈，具体的には，実世界のエンティティがデータの中にどのように現れ，相互にどういう関係を形成しているかを調べるのが実世界指向マイニングのアプローチである．これは，デジタル情報を実世界に結び付け意味や価値を与える方法を問うシンボルグラウンディング（記号接地）問題[12] に近似解を与えるものとしてとらえることができる．

このアプローチが単純な統計処理に比べより効果的であることは，4 章の評価結果が示している．

2 章で述べたように，同姓同名人物分離問題は文書の分類に帰着させることができる．そして，文書の分類は，(1) 文書を特徴付けし，(2) その特徴に基づいて相互関係を判定するという 2 つの処理かならなる．(1) の処理において従来広く用いられていた方法が，統計的処理により文書から特徴的な語を抽出するというものである．4 章の評価実験では，f_{tfpsr}，f_{tfidf}，f_{psr} を使って特徴語を抽出するもの（4.2.3 項の (i), (ii), (iii)）がそれに該当する．一方，実世界指向マイニングに対応するのが f_{name} を用いた場合（4.2.3 項の (v)）である．実験結果は，(2) の処理に同じアルゴリズムを用いたとき，実世界指向マイニングのアプローチが従来の統計的処理よりも優れていることを示している．

(2) のための処理方法，すなわち文書群（一般にはデータ集合）を相互の関連性に基づいてグループ分けする方法はクラスタリングと呼ばれている[13]．一般に，クラスタリングは主観的かつ探索的なデータ解析手法であり[13]，データ集合の特徴を大雑把に把握するといった目的に利用されることが多い．最短距離法，k-means 法など複数存在するクラスタリングの標準的な手法は，特徴探索のためのツール群として位置付けられる．

一方，同姓同名人物分離のタスクでは，個々の文書のグループが人物に対応としているという客観性が要求される．この客観性を得るためには，既存の手法をそのまま適用するのでは不十分であり，3.3 節に示したような解析対象である人名（が指し示す人）どうしの関係の特徴を考慮した手法が必要となる．

6. むすび

以上，Web から信頼性の高い知識を効率良く取り出す方法である実世界指向マイニングの考え方を示し，Web 上における同姓同名人物の分離への適用によりその妥当性を確認した．本手法の特徴は，高い精度とともに，データ自体が示す実世界のルールを意識的に掘り起こすことなくそのままデータの解釈に適用できるという柔軟な解析能力にある．この柔軟性ゆえ，本論文で示した同姓同名分離以外にも，人物のように多面性があり形式的にとらえにくいものの解析に対して特に効果を発揮すると思われる．

我々の日常生活と Web との関係は今後さらに密になっていくと思われる．その結果として，Web はいま以上に詳細に社会を映す鏡となり，そこからの知識抽出に対する需要も高まると思われる．ここで期待されるのは，形式化，固定化された知識だけでなく，“活きた” 知識の獲得である．そのためには，本論文で提案したような，データのありように実世界の状況を “語らせる” 工夫が今後重要になるだろう．

参考文献

1) 山名早人：Web データの新しい利用法の開拓を目指して，情報処理学会研究報告，2004-FI-75, pp.107–110 (2004).
2) 野島久雄：データベースとしての WWW，データベースとしての社会,（CMC 研究ノート第 8 回），Computer Today, No.84, pp.60–67 (1998).
3) Berners-Lee, T., Hendler, J. and Lassila, O.: The Semantic Web, *Scientific American* (May 2001).
4) 武田英明：知性のネットワークとしての WWW — Web インテリジェンスに関する一考察，人工知能学会誌，Vol.17, No.3, pp.346–351 (2002).
5) Kosala, R. and Blockeel, H.: Web Mining Research: A Survey, *SIGKDD Explorations*, Vol.2, No.1, pp.1–15 (2000).
6) Girvan, M. and Newman, M.E.: Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA 99*, pp.7821–7826 (2002).
7) Watts, D.J. and Strogatz, S.H.: Collective dynamics of small-world networks, *Nature*, No.393, pp.440–442 (1998).
8) 佐藤進也，原田昌紀，風間一洋：Web 上の「活動の場」に着目した人物の特徴付け，情報処理学会研究会報告 2004-DBS-133-9/2004-FI-71-9, pp.75–82 (2004).

9) Fruchterman, T.M.J. and Reingold, E.M.: Graph Drawing by Force-directed Placement, *Software — Practice and Experience*, Vol.21, No.11, pp.1129–1164 (1991).
10) 橋田浩一：GDA：言語データの意味的構造化とインテリジェントコンテンツ，電子情報通信学会技術研究報告，TL2000-5, pp.33–39 (2000).
11) 永田昌明，平博　順：テキスト分類—学習理論の「見本市」，情報処理，Vol.42, No.1, pp.32–37 (2001).
12) Harnad, S.: The symbol grounding problem, *Physica D*, Vol.42, pp.335–346 (1990).
13) 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).

(平成 16 年 9 月 14 日受付)
(平成 17 年 1 月 29 日採録)

(担当編集委員　石川 博，原 隆浩，片山 薫，佐藤 聡，土田 正士)

佐藤　進也（正会員）

1988 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在 NTT 未来ねっと研究所主任研究員．ACM，Internet Society，電子情報通信学会各会員．

風間　一洋（正会員）

1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT 未来ねっと研究所主任研究員．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．

福田　健介

1999 年慶應義塾大学大学院理工学研究科計算機科学専攻後期博士課程修了．同年日本電信電話（株）入社．現在未来ねっと研究所研究主任．この間，2002 年ボストン大学訪問研究員．インターネットトラフィックのダイナミクス，ネットワーク構造の統計的解析等の研究に従事．博士（工学）．

村上健一郎（正会員）

1955 年生．1979 年九州大学工学部情報工学科卒業．1981 年同大学院修士課程修了．同年日本電信電話公社入社．以来，超大型計算機用 OS，記号処理計算機，インターネットパラダイム，超高速インターネットプロトコルの研究に従事．2005 年 4 月より，法政大学ビジネススクールイノベーション・マネジメント研究科教授．博士（情報科学）．電子情報通信学会，ACM，ソフトウェア科学会各会員．主な著書『はやわかり TCP/IP』（共立出版，共著），『インターネット縦横無尽』（共立出版，共著），『インターネット』（岩波書店）等．