実世界指向 Web マイニングの提案とその同姓同名人物分離問題への適用

The Concept of Real-World Oriented Web Mining and Its Application to the Task of Distinguishing between People on the Web with the Same First and Last Name

佐藤 進也♣　風間 一洋*
福田 健介*　村上 健一郎*

Shin-ya SATO　Kazuhiro KAZAMA
Kensuke FUKUDA　Ken-ichiro MURAKAMI

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 97%以上の高い率で正しく処理できることを確認した．

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 97% on average.

1. はじめに

Webページの数は2004年4月の時点で92億を越えると推測されている[1]．この巨大さはWebの一つの特徴であるが，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― が挙げられる[2, 3]．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捉えることができる．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[3]．

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[4]がその典型例．
- 予め知識（の表現）の枠組みを用意し，データをその枠組みに当てはめて整理・解釈するセマンティック・ウェブ[5]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得るものなのか，実世界をどれだけ正確に捉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，予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，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（例えば，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．

2. 実世界指向Ｗｅｂマイニング

ここで提案するアプローチは，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ものである．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．

実世界の構成要素に焦点を当てるということには主観を排するという意味もある．この考え方は，日常生活にかかわる“モノ”(生活財)を悉く調べ上げ，“モノをして世相や文化を語らしめる” 生活財生態学[6]に共通するものである．

3. 同姓同名人物分離問題への適用

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ることを示す．

3.1 基本方針

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つの文書内に複数回現れる同じ名前は同一人物を指していると仮定すると，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，文書集合と人物を一対一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

♣ 正会員 NTT未来ねっと研究所
sato@ingrid.core.ntt.co.jp
* 非会員 NTT未来ねっと研究所

* 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．

着される．

いま，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)とすると，求めるべき文書集合群Ω ={C_i} (i=1,...)は以下の条件を満足するD(x)の部分集合の族である：

$$D(x)=\bigcup_i C_i, {}^{\forall}d\in C_i, {}^{\forall}d'\in C_j \text{に対して} p(d,x)=p(d',x)\Leftrightarrow i=j$$

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もとづいて計算される．

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

近年，実世界における活動（あるいは活動のための組織や場）に関する情報提供のため Web サイトを用意することが一般的に行われている．そして，多くの場合，それらの活動に関与する人物の名前は Web サイト中のページに出現している．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い出すことができる（図１）．

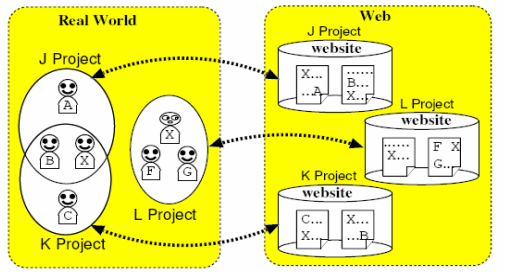

図１ 実世界と Web の対応付け

Fig. 1 Correspondence between the real world and the Web

実世界では，活動の場を介して人間関係が形成され，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，二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１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）という二つの活動の場は 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１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 という人名が共に現れているが，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 をもつ同姓同名人物を分離のアルゴリズムは次の通りである：

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

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[7]を用いる． 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高々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 つの文書とみなしたときの語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‡．具体的には，語yを含むページをもつWebサーバの集合を$S(y)$とすると$f_{psr}(y)=\log|D(y)|/\log|S(y)|$．なお，$D(y)$は **3.1** で定義した，人名$y$を含むWebページの集合である．普通名詞では$\log|D(y)|$と$\log|S(y)|$はほぼ比例関係にあるが，人名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[7]．
- $f_{tfpsr}(y)$: f_{psr}にw_iを 1 つの文書とみなしたときのyの出現頻度（tf）を掛け合わせたもの．

iii.の具体的な方法については，次項で述べる．

3.3 グラフ構造に基づくワークスペースの分類

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２に示す．グラフの描画は Fruchterman-Reingold のアルゴリズム[8]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２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

†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．

‡ psrはPage-to-Server Ratioの略．

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.

この例では，3 人の人物はそれぞれ図２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．

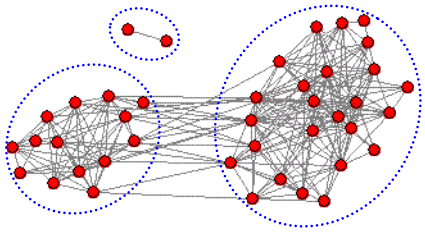

図２　ワークスペースの相互関係を示すG(x)の例

Fig.2 Example of G(x) showing relationships among workspaces

グラフを稠密な部分に分解する一般的な方法としてはbetweennessに基づくクラスタリング[9]などがあるが,ここでは，同姓同名分離という目的に特化した方法を提案する．

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）なか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３の(1)のグラフを例にとり説明する．

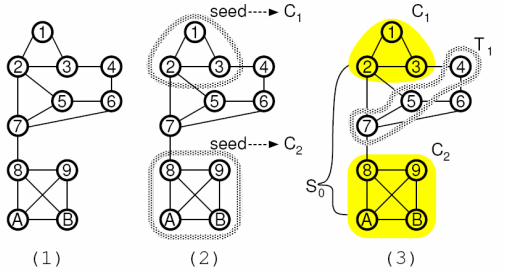

図３　グラフ分割アルゴリズムの適用例

Fig.3 Applying the graph decomposition algorithm to a graph

I.　まずはじめに，シードとしてクラスタ係数[10]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一つのシードとする．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, ..., M)としておく．図３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はG(x)のノードでS_{i-1}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３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Ti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，$l(n_1, n_2)$は二つのノードn_1，n_2を結ぶリンクの重みで，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．

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

以上に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

いま，同姓同名人物分離の処理を行った結果，Ω = $\{C_i\}$, (i=1,.., 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，..,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４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．

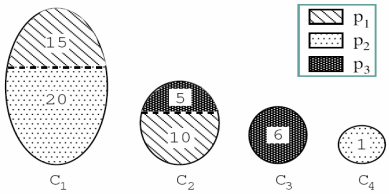

図４　分離された状態の例

Fig.4 Schematic example of decomposition

さらに，

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

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，$a_{ij} / a_{i*} > 0.5$ かつ $a_{ij} / a_{*j} > 0.5$ 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は高々1 つしか存在しない．p_jのうち，認識され得るものの割合を認識率とする。図４の例では，p_2，p_3はそれぞれ C_1，C_3により認識されているが，p_1を認識するC_iは存在しないため，認識率は 2/3 である．

4.2　評価の方法と結果

評価には 10 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．このデータに提案手法を適用して同姓同名人物の分離を行い，従来手法による結果と比較した．その具体的手順は以下の通りである．

まず，3.2 で示したアルゴリズムの手順(i)でワークスペースを構成した後，(ii)において，人名に限らず， 以下の基準に基づいてワークスペースの特徴を表す語を選び出す：

(a)任意の語でf_{tfpsr}の値が大きいもの（any term, tfpsr）
(b)任意の語でf_{tfidf}の値が大きいもの（any term, tfidf）
(c)任意の語でf_{psr}の値が大きいもの（any term, psr）
(d)任意の複合語でf_{tfpsr}の値が大きいもの（compound term）
(e)人名でf_{tfpsr}の値が大きいもの（person's name）

(a)～(c)は従来の統計量に基づく特徴語抽出，(e)は本論文の提案手法にそれぞれ対応する．

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}を選び， 10 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．認識率を計算するうえで必要なC_iとp_jの対応つけ，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．

この結果をまとめたものが図５である

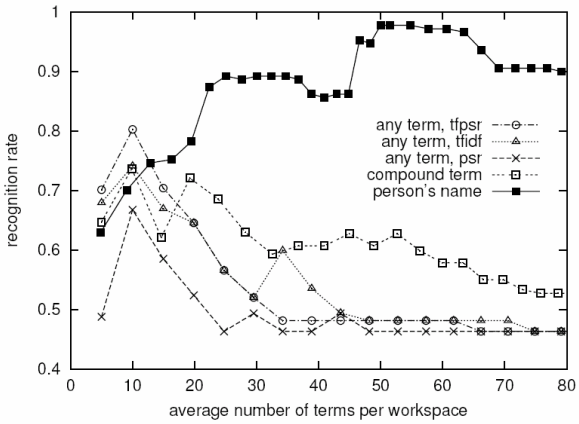

図５　特徴語数と認識率

Fig.5 Recognition rates versus number of characteristic terms

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．

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(e)とその他の基準との間に大きな違いが認められる．基準(e)のピーク値(0.9773)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や，姓と名を組み合わせた複合語であることが挙げられる．そこで，人名の代りにこれらの性質をもつ語を用いることも考えられる．しかし，選択基準の(a),(c)および(d)の結果はこの可能性を否定している．

5. むすび

以上，Web から信頼性の高い知識を効率よ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．今後は，Web をより詳細に社会を映す鏡とするため，人物以外にも解析対象を拡げていく予定である．

［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] 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).
[4] R. Kosala, H. Blockeel: Web Mining Research: A Survey, SIGKDD Explorations, Vol. 2, No. 1, pp. 1-15 (2000).
[5] T. Berners-Lee, J. Hendler, O. Lassila: The Semantic Web, Scientific American, May 2001 (2001).
[6] 疋田正博: 生活財生態学の方法，川添登・佐藤健二(編)，生活学の方法，光生館, pp.69-97, (1997).
[7] 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).
[8] T. M. J. Fruchterman, E. M. Reingold: Graph Drawing by Force-directed Placement, Software - Practice and Experience, Vol. 21, No. 11, pp. 1129-1164 (1991).
[9] M. Girvan, M. E. Newman: Community structure in social and biological network, Proc. Natl. Acad. Sci. USA 99, pp. 7821-7826 (2002).
[10] D. J. Watts, S. H. Strogatz: Collective dynamics of small-world networks, Nature, No. 393, pp. 440-442 (1998).

佐藤 進也　Shin-ya SATO
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，情報処理学会，日本データベース学会等各会員．

風間 一洋　Kazuhiro KAZAMA
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．情報処理学会，ソフトウェア科学会，ACM 各会員．

福田 健介　Kensuke FUKUDA
1999 年慶応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終了. 同年日本電信電話(株)入社．現在未来ねっと研究所に所属．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)．ACM 会員．

村上 健一郎　Ken-ichiro MURAKAMI
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．現在，NTT 先端総合研究所および未来ねっと研究所主幹研究員. 博士(情報科学)．情報処理学会，電子情報通信学会，ACM，ソフトウェア科学会各会員．

§ 収集は，2003 年 7 月に主にjpドメインを対象に行った．