# Web 情報からの人物知名度の推定

## —篠田麻里子は長嶋茂雄より有名って本当？—

小紫　弘貴†　　田島　敬史††

† 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　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †komurasaki@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

**あらまし**　近年，Web を現実社会のセンサーと捉え，Web 上の情報から現実社会に関する情報を推定する研究が行われている．しかし，個人による Web への情報発信量は年齢や嗜好によって異なり，また，ある情報が Web へ発信される確率もトピックによって異なるため，Web は現実社会の偏った反映であると推測される．本研究では，有名人の知名度を題材に，Web 情報から偏りを補正して現実の知名度を推定する手法を提案する．提案手法では，上述の偏りを考慮した，Web への人名の発信過程を表すモデルを設計し，Web 上で観測可能な値からモデル中の変数を推定することで，知名度を推定する．提案手法による推定値と，現実社会でのアンケートによる知名度との比較により，提案手法の妥当性を示す．

**キーワード**　情報抽出，情報要約，情報集約，情報発信，Web データマイニング，Web 情報の生成モデル，人物情報の発信モデル

## 1. はじめに

近年，Web はより身近なものとなり，その中でも特にブログや SNS などのソーシャルメディアの普及，またスマートフォンといった携帯情報機器が発展・普及したことにより，Web が個人による情報発信の場として利用されることが多くなっている．情報発信の方法は多様なものとなっており，ブログに記事を投稿する，Twitter で発言する，掲示板で議論を交わす，など各個人が好きなように情報を Web へ発信することが可能となった．様々な人が様々な情報を発信することで Web に蓄積された情報は膨大なものとなっているが，ある情報について Web 上に存在する情報の量は，情報発信を行う個人の年齢や嗜好，また発信される情報の性質などの違いによって現実社会で一般に得られる情報量と異なることがあり，Web は現実社会に存在する情報量を正確に反映したものではないといえる．

例えば，ある有名人についての情報を Web 検索を用いて収集しようとする時，調べようとする有名人によっては，現実社会と比べて Web 上ではよく知られているために情報発信が活発に行われており，想像以上の情報を取得できることがある．また反対に，多くの人に知られている有名人であるはずなのに，Web 上ではあまりその様子が伺えずあまり情報が得られないといったこともある．これは現実社会と比べて Web 上で人気がある，その人物を知っている人が Web をよく使う人であることが多い，またとても有名な人物であるが最近は特に話題がない等といったことが原因で Web 上の情報量に偏りが存在しているためと考えられる．

ここで，実際に情報によってその情報量がどの程度偏っているかについて例を挙げる．図 1 に示すように，「AKB48」のメンバーである「篠田麻里子」について調べてみると Web 検索結果


図 1　Web 検索結果数の比較—篠田麻里子と長嶋茂雄—

数は「15,500,000 件」となっており，「ミスタージャイアンツ」として知られる「長嶋茂雄」の Web 検索結果数は「1,530,000 件」となっている．Web 検索結果数が多いと直感的に有名な人物であるように推測できる．しかし，「篠田麻里子」の検索結果数が「長嶋茂雄」の 10 倍程度で得られるが現実では「長嶋茂雄」の方が知名度は 11.1 ポイント高いという調査結果が出ている．(注1) これは，「篠田麻里子」を知っている人の傾向と「長嶋茂雄」について知っている人の傾向が異なっているため，つまり「篠田麻里子」を知っている人は Web を利用し，情報発信を積極的に行う人である可能性が比較的高く，「長嶋茂雄」を知っている人は Web をあまり利用せず，利用したとしても情報発信には消極的な人である可能性が比較的高い，と考えられる．また最近では「篠田麻里子」の方が「長嶋茂雄」に比べて活発に活動しており話題があるため，このような結果になっていると考えられる．

Web 情報からのデータマイニングによって現実社会に関する有用な情報や知識を抽出する研究は盛んに行われているが，現実社会と Web との情報量の偏りにより，Web 上に情報が多く得られて必要以上に重要視してしまう，あるいはあまり得られないことで軽視してしまうといった問題があり，現実社会にお

(注1)：本研究で用いた知名度データは株式会社ビデオリサーチから提供して頂きました

けるその情報の重要性を正確に反映したデータマイニングを行うことは困難となっている．

そこでわれわれは，有名人の知名度に焦点を当て，Web と現実社会との情報量の偏りを補正する手法について研究を行った．本論文では，有名人に関する情報の Web 上における情報量は Web 検索結果数が表していると仮定し，現実社会でのその有名人の知名度と比べて妥当でない程度を Web と現実社会との情報量の偏りとして，その偏りを補正することで Web 上の情報から知名度を推定する手法を提案する．

まず，Web 上において有名人に関連する情報が発信される過程を提案し，情報の発信過程に影響を与える有名人それぞれの性質が，Web 上の情報のどのような要素，例えば発信媒体や言葉遣いなどに影響を与えるかを推測する．そしてそれらの要素を Web 上の情報から観測可能な値によって表現し，それらの値を特徴量として SVM によって学習を行い，作成した SVM を用いた回帰分析によって知名度の推定を行う．また，よりよい結果を得るためにより詳細な情報の生成モデルを提案し，その有効性の検証を行う．

本章で研究の背景，およびその目的について述べた．以下，まず，関連研究を紹介し，続いて，本研究の特徴を明確にした後に，有名人に関する情報の発信過程を表すモデルを提案し，その後，まず SVM を用いた回帰分析による知名度推定の手法について述べる．さらに，より詳細な情報の生成モデルを提案し，実装および実験と評価を行う．最後に，得られた結果を元に，問題解決の発展的な手法への展望を述べる．

## 2. 関 連 研 究

### 2.1 Web 情報からの情報抽出

本研究ではある情報の Web 検索結果数，すなわち Web 上の情報量が重要な要素であると考えており，他に Web 検索結果数など情報量に価値を置いた研究として山本ら，金ら，平野ら，Lan らが行った研究が挙げられる．

Yamamoto ら [6] は入力された知識が Web 上でどの程度言及されているかを調べ，類似する，あるいは対立する情報を Web から取得し，それらと比較を行うことによって入力知識の信憑性を計算する研究を行なった．金ら [8] は Web 検索エンジンの検索結果数を用いて特定の企業間の関係を表す単語を抽出する手法を提案した．金らは複数の企業と，関係を表す語をクエリとして検索を行い，得られた検索結果数が多い関係を表す語が，それらの企業間の関係を適切に表す語であるとした．平野ら [10] は Web 検索結果数を用いて英語の冠詞誤りを検出する研究を行った．平野らはクエリとして，英文の構文解析によって得られた名詞句の冠詞を変化させたものを 3 パターン用意し，それぞれについて検索を行い，検索結果数が最も多いものが正解であるとしている．

また，Lan ら [7] は Web 掲示板中に存在する情報の中から話題のトピックを抽出する研究を行い，トピック抽出にはそのトピックに関する投稿の数を用いる方法がシンプルで効果的であるとし，それに加えて投稿の様々な特徴から話題のトピックを抽出している．

しかし，山本らは信憑性を測る際に，対立情報のうち Web 検索結果数が多いほうが信憑性が高いと考えており，金らは関係を表す語の中で Web 検索結果数が多いものが企業間の関係を表しているとし，平野らは英語の名詞句の冠詞を変化させ，Web 検索結果数が最も多くなる冠詞が正解であるとした．また，Lan らはトピックに関する投稿数が多ければそれが話題のトピックであるとしている．これらに対し，われわれは直感的には Web 検索結果数が知名度を表しているとしているが，Web 上の情報に偏りが存在するため，単純に Web 検索結果数が多ければ知名度が高いとは言えないと考えている点で異なる．

また，本研究では Web 検索結果数以外の Web 上の情報も用いて有名人の知名度を推定するための手法を提案しており，Web 上の情報を用いて有用な知識を抽出する研究として立石ら [9] の研究が挙げられる．立石らの研究では，Web を人々が情報発信出来る場であると捉え，発信された主観的な意見を効率的に収集，分類・分析する手法を提案しており，その意見の内容に着目しているが，本研究では Web 上に発信されたそういった意見などを含む情報の量に着目している点で異なる．

### 2.2 情報の伝播過程に関する研究

本研究では情報の発信過程について述べており，情報発信過程に関する研究として佐藤ら [11] の研究が挙げられる．佐藤らは個人の情報空間をもとに流通範囲や関連情報などの情報のメタ情報を動的に設定することで，柔軟な情報の利用と流通を可能にした個人の情報発信モデルを提案した．佐藤らは情報発信をする個人に焦点を当て，その過程についてモデルを提案したが，本研究では発信される情報の性質によってその発信方法などが変化すると考えている点で異なる．

発信過程と関連する情報の拡散過程の研究として，Gruhl ら [2] や Matsubara [4] の研究が挙げられる．Gruhl らはブログ空間における情報の伝播過程をマクロとミクロ，つまりあるトピック単位と個人単位という二つのレベルについてそれぞれ特徴抽出とモデル化を行った．松原らは，ブログなどのインターネットメディア上におけるニュース情報の拡散過程を時系列モデルとして表現する手法を提案しており，提案するモデルについて最小二乗法を用いて平均二乗誤差を最小とするようなパラメータを発見し，提案モデル *SpikeM* の実装を行った．

## 3. 有名人情報の発信モデル

本章では Web 上の一般的な情報がどのような過程で発信されるのかを提案し，それを用いて情報の中でも有名人に関する情報に限定した場合の発信モデルについて述べる．

### 3.1 Web 上の情報の生成過程

まずはじめに，人物名に絞らず，ある情報がどのように発信されているかを説明する．

ある個人がある情報について発信する際の発信モデルは図 2 のように書ける．まず，ある個人が情報を発信するためには，その情報を知っていなければならないため，まずその情報を知っている人かそうでないかによって分類される．なお，この情報が人名の場合，知っているかどうかを集約したものが知名度となる．そしてその人物を知っている個人の場合，その人物

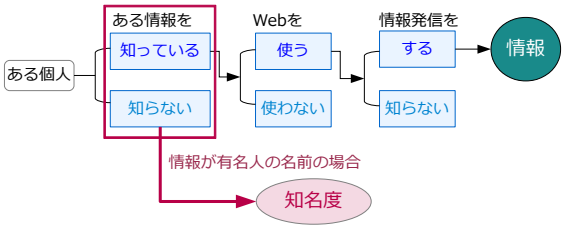


図 2　単純な Web 情報の発信モデル

に関する情報を知っている人が Web を使わなければ Web を用いて情報発信することはない．従って，その人が Web を使うかどうかによっても分類がなされる．そしてその情報を Web を用いて発信することによって，ある個人が持っている情報が Web 上の情報として存在し得るため，情報発信を行うかどうかによっても分類が行われる．以上より，ある情報について知っており，Web を利用し，情報発信を行う人によって，その情報は Web 上の情報として存在することになる．

ある情報が発信される確率について述べるにあたり，各事象を以下のように定義する．

**A:** ある個人がある情報を知っている
**B:** ある個人が Web を利用する
**C:** ある個人が情報発信を行う

まず，ある個人がある情報を知っている確率を $P(A)$ と表す．ある情報がどのようなものであるかによって，その情報を知っている人の性質が異なると考えられるため，全人口について見た場合のある人が Web を使う確率とその情報を知っている人が Web を使う確率も異なるものであるといえ，後者の確率は $P(B|A)$ と表される．さらに，その情報を知っており，かつ Web を使う人が情報発信を行うかどうかも同様に，全人口について見た場合のある人が情報発信を行う確率とは異なるといえる．従ってある個人がある情報を知っており，Web を利用して情報発信を行う確率は $P(C|A \cap B)$ と表すことが出来る．以上より，ある人がある情報について Web を用いて情報発信を行う確率は，$P(A) \cdot P(B|A) \cdot P(C|A \cap B)$ となる．

### 3.2　有名人の情報の性質

ここでは有名人の性質について述べる．本研究では，Web 上の情報量に影響を与える有名人の性質を

（1） Web との親和性
（2） 話題性

という 2 つであると考える．

ある有名人の Web との親和性とは，その有名人の情報がどの程度 Web 上に発信されやすいかを表す指標であり，有名人の Web 上での人気やその人についての Web 上における議論の活発さを表すものである．Web との親和性が高いほど Web 上で頻繁に言及されている，あるいは活発に議論がなされており，それに伴い気軽に情報発信出来る状況となっていることを表す．その人物が，Web を日常的に利用して情報発信を行う人に好まれるような人であれば Web との親和性が高くなると考えられる．

そして有名人の話題性とは，最近その人物について発信されるべき情報が発生したかどうか，その情報が持つニュースとしての強さはどれほどか，そのニュースの中でその人物はどれほど重要であるかということを表す．話題性が高いほど，その人に関する最近の情報が，その人を中心とする社会的に大きなニュースであったということを表す．

### 3.3　有名人に関する情報の生成過程

3.1 節で説明した情報の発信モデルを，有名人に関する情報に特化させることで．有名人に関する情報の生成過程を得る．

(1) ある有名人を知っている人が Web を使うかどうか，(2) ある有名人を知っており，かつ Web を使う人が情報発信を行うかどうか，(3) 情報発信の手段および記述方法という 3 点について，発信される情報となっている有名人の性質，つまり Web との親和性および話題性がどのように影響を与えるかを説明する．

まず，(1) ある有名人を知っている人が Web を使うかどうかについてである．これについて条件付き確率を用いて表すと，$P$(Web を使う | その人を知っている) となり，その人を知っている人が Web を使うような人であるかどうかは Web との親和性が影響を与えると考えられる．Web との親和性が高い人物であれば，その人物についてある人が Web 以外で知ることがなくても Web を使っている内に目にして知ることがあり，また知っているならば，Web を使っているから知っている確率が高いとも考えられるためである．

例えば「篠田麻里子」は「AKB48」というアイドルグループに所属しているアイドルであり,「AKB48」は特に Web 上での人気があるグループであると思われる．また，篠田麻里子は，自身のブログで記事を書き，また，Twitter を用いて発言をするなど積極的に Web を利用する人々と関わっており,「篠田麻里子」の Web との親和性は非常に高いと考えられ，Web を使う人であれば目にすることも多いと考えられる．

それに対し,「長嶋茂雄」は元プロ野球選手で，読売ジャイアンツの終身名誉監督であるが，公式ブログも持たず，最近では活発に Web 上で議論が行われているわけでもないため，Web との親和性は比較的低く，Web 以外，例えばテレビなどで知る機会が多い人物であると考えられる．

これらのことから,「篠田麻里子」を知っている人は Web を利用している確率，$P$(Web を使う | 篠田麻里子を知っている) が高く，それに比べて「長嶋茂雄」を知っている人は Web を利用している確率，$P$(Web を使う | 長嶋茂雄を知っている) が低いといえる．

次に (2) ある有名人を知っており，かつ Web を使う人が情報発信を行うかどうかの確率は，$P$(情報発信を行う | その人を知っている∩ Web を使う) と表すことが出来る．Web との親和性が高ければ Web を日常的に Web 上での議論が活発であり，それによって気軽に情報発信出来る状況となっているため，Web との親和性が影響を与えると考えられる．また話題性が高ければ，その有名人を中心とした新しい重大ニュースがあるということなを表し，ニュースメディア等では大きく取り上げられ，そして普段

は情報発信をしない個人でもごく最近の情報であれば古くなった情報に比べて発信する確率が高くなり，また重要なニュースである，或いは知っているその人物にとって重要な情報である，と考えて発信する確率も高くなると考えられる．これらのことから，$P$(情報発信を行う | その人を知っている ∩ Web を使う) に対して Web との親和性，話題性の両方が影響を与えるといえる．

「篠田麻里子」はメディアに露出する機会が多く，それに伴って関連するニュースも多いため話題性も高くなると考えられ，「長嶋茂雄」は以前に比べてそれほど頻繁にはニュースに登場しておらず，話題性は「篠田麻里子」と比べると低く．先ほど述べた Web との親和性と話題性の両方とも「長嶋茂雄」より「篠田麻里子」の方が高いと考えられる．従って Web を日常的に利用して情報発信を行う人に好まれており，かつよりホットな話題のある「篠田麻里子」についてある個人が情報発信する確率，$P$(情報発信を行う | 篠田麻里子を知っている ∩ Web を使う) の方が，$P$(情報発信を行う | 長嶋茂雄を知っている ∩ Web を使う) よりも高くなると考えられる．

最後に (3) 情報発信の手段および記述方法についてである．

発信される有名人の性質によって，その情報発信の方法に差が出てくるものであり，有名人の性質は (1) 記事内での扱い，(2) 言葉遣い，(3) 発信媒体，(4) ニュース性，(5) 情報量という 5 個の要素に影響を与える．

まず，記事内での扱いとは，個人が Web 上で有名人について記述する際に．その人物にどれだけ重きを置いた情報として記述するかである．これは有名人の性質の内，話題性が影響を与えるもので，あるニュースについてその人物が重要な存在であれば，その人物名をタイトルに冠した記事が作成されることや，人名を列挙するときに先頭に記述されることが多くなると考えられる．

次に言葉遣いとは，その有名人に関する情報を記述する際に用いられる言葉の選択のことであり，どのような人が情報発信を行うかによって変化するものである．発信する人の Web の利用頻度および情報発信頻度によって，発信する際の言葉遣いに変化，例えば Web を頻繁に利用し掲示板などで情報発信を行う人であればネット用語を利用することが多くなる，といった変化があると考えられる．従って，言葉遣いには Web との親和性が影響を与えると推測出来る．

次に，発信媒体とはその有名人について情報発信を行う人が，情報を発信する先のことであり，個人が発信する場としてブログを用いることが多いと考えられる．ここで，Web との親和性が高い有名人，つまり Web 上で人気があれば新聞社などのニュースサイトといった公式なメディアによって発信される情報だけでなく，個人がブログを用いて発信する可能性が高くなると考えられる．このことから，発信媒体としてブログが用いられるかどうかは Web との親和性によって決定されると考えられる．

ニュース性とは，発信されるべき情報がどれほど最近に存在するかということである．これは有名人の話題性が影響を与えるもので，話題性が高い，つまり新しい重大な情報が多く存在

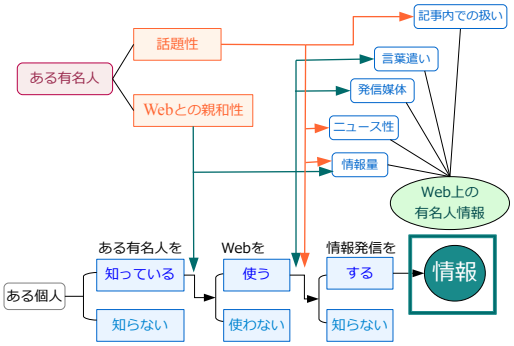


図 3　有名人情報の発信モデル

する場合やその有名人が大物と呼ばれるような人物であった場合，小さな話題であっても大きく取り上げられることとなるといえる．

最後に情報量についてである．Web 上の情報とは誰かが情報発信を行うことで生成されるものであり，頻繁に情報発信を行う人に多く知られていれば，必然的に情報量は多くなる．また，発信される情報が多く存在しており，特にその人物が中心となるようなニュースが多ければ関連する情報量は増える．従って情報量には，Web との親和性および話題性の両方が影響を与えると考えられる．

図 3 は，前節で説明した Web 上の情報の生成過程に，本節で説明した (1) ある有名人を知っている人が Web を使うかどうか，(2) ある有名人を知っており，かつ Web を使う人が情報発信を行うかどうか，(3) 情報発信をどのようにするかという 3 点について，発信される情報となっている有名人の Web との親和性および話題性がどのように影響を与えるかを表したものである．

## 4. SVM による知名度推定

本章では第 3. 章で述べた有名人情報の発信モデルに基づき，SVM による回帰分析を行い，知名度を推定する手法について説明する．

SVM を用いた知名度推定手法の概要を以下に述べる．まず，ある有名人の名前をクエリとし，関連する Web ページ集合をクローリングによって取得する．そしてそのページ集合中から，有名人の性質が影響を与える 5 個の要素をそれぞれ表す特徴量を抽出し，学習によって作成した SVM を用いて回帰分析を行い，知名度を推定する．

### 4.1 有名人情報の特徴量

本節では，有名人の性質が影響を与える 5 個の要素，つまり (1) 記事内での扱い，(2) 言葉遣い，(3) 発信媒体，(4) ニュース性，(5) 情報量を SVM で用いる特徴量として使用するため，それらの要素を表す特徴を，関連する Web ページ集合から抽出する．

以下では特徴量として使用した 9 個の特徴について説明する．

(1) から (3) が記事内での扱い，(4) が言葉遣い，(5) が発信媒体，(6) と (7) がニュース性，(8) と (9) が情報量を表す特徴

となっている．

(1) 出現位置

これは記事内での扱いを表す特徴であり，関連するページ中のどの程度上部に有名人が登場するかによって得られるものである．有名人が登場するのがページの上部であるほど，そのページ中の情報にとってその有名人が重要な役割を果たしているといえる．そのようなページが，つまりそのような情報が多く存在するならば，その人物を中心とするニュースが多く存在することとなり，話題の人物であると推測できる．

(2) タイトルに登場する割合

ある有名人に関する Web ページが，タイトルにその名前を含む場合，出現位置が高い場合と同様にページ内の情報がその有名人を中心とする情報であるといえる．また，タイトルはページ内の情報を端的に表現したものであるため，そのページはその有名人のために作成されたページであると考えられる．従って，そのようなページが関連 Web ページ集合内に多く存在すればその有名人に重要な情報が多く存在することを表す．

(3) 人名の並び順

例えば映画の出演者を紹介する記事のように情報内に複数の有名人が並んで登場する場合，先頭に登場する人物ほどその記事内の情報において重要な立場であると考えられる．平均して先頭の方に登場する人物であれば，社会的に重要な人物であるといえ，そのような人物の話題であれば大きく取り上げられると予想され，話題性に大きく寄与するものとなる．

(4) 共起語の特徴

ある有名人について情報発信を行う人々がどのような人物であるかによって，情報が記述される際の言葉遣いが変化するものといえる．その言葉遣いとは，有名人の名前と同時に出現する語，つまり共起語の特徴によって表される．共起語の特徴はベクトル空間モデル [5] を利用して求める．まず，ある有名人に関する Web ページ集合の文書ベクトルを考え，関連ページ内に登場する語の特徴をそれらの文書ベクトルの重心とする．そして，様々な有名人それぞれに関するページ集合全てを文書ベクトルで表現し，その重心ベクトルと先程求めたある有名人に関する文書ベクトルの重心とのコサイン類似度を共起語の特徴とする．

(5) ブログ検索結果数

個人が情報発信を行う場所，つまり発信媒体としてブログが挙げられ，Web を用いて日常的に情報発信を行う人々に多く知られていれば，ある有名人に関する Web ページ集合中に，ブログ記事が多く出現すると考えられる．また，そのような人々が多い，かつその人物の話題が多ければブログ記事の数が増えるものと予想される．従って，有名人の性質のうち Web との親和性と話題性の両方に強く影響されるものと考えられる．なお，他の特徴量に比べて重要視してしまうことを防ぐために，ブログ検索結果数の対数をとったものを特徴量として用いる．

(6) Wikipedia の更新頻度

有名人についてある情報が発生すれば，その人物についての Wikipedia 記事が更新される可能性が高いと推測できる．そこで，有名人についての Wikipedia 更新頻度を，その人物についての Wikipedia 記事が 100 回更新されるまでの一日あたりの平均記事更新数を特徴量とした．

(7) ニュース検索結果数

有名人に関連するニュースがどの程度存在しているかはその人物の話題がどれほど多くあるかを表すもので，また，Wikipedia と異なりその人のニュースが全てニュース検索によって得られるものではないため，最近の話題がどれほどあるかをよく表すものであり，有名人の性質のうち話題性に強く影響されるものと考えられる．ブログ検索結果数と同様にニュース検索結果数の対数をとったものを特徴量として用いる．

(8) Web 検索結果数

Web 上に関連する情報がどの程度存在しているか，つまり情報量は Web 検索結果数が直感的に表しているものであり，一般に Web 検索結果数が少ない人よりも多い人の方が有名である，或いは人気があると予想されるため，重要な指標であると考えられる．なお特徴量として，Web 検索結果数の対数をとったものを用いる．

(9) Wikipedia の記事サイズ

Web 上の情報量として Web 検索結果数だけでは不十分であると考え，Wikipedia 記事のバイト数を，情報量を表すもう一つの特徴として用いる．Web 検索結果数が検索エンジンのインデックスによって変化してしまうという問題があるのに対し，Wikipedia には記事が作成された時点からの情報が掲載されており，安定した情報量を表していると考えられる．Wikipedia の記事サイズの対数をとったものを特徴量として用いる．

以上で述べた，(1) から (9) の 9 個の特徴量を用いて SVM で学習を行う．

### 4.2 SVM による検証

本研究では，株式会社ビデオリサーチから知名度データを提供された 1000 名の有名人について Yahoo! JAPAN(注2)の提供する検索 Web API(注3)を利用して得られた検索結果のうち，上位 100 件の Web ページをスクレイピングによって収集し，また，形態素解析器として MeCab(注4)を用いた．SVM は LIBSVM(注5)の回帰分析用の SVR(Support Vector Regression) [1] の 1 つ，epsilon-SVR を利用し，カーネル関数としては LIBSVM でデフォルトとして指定されている動経基底関数 (Radical Basis Function) を用いた．

また，抽出した全ての特徴量を，学習に用いる前に全て 0 から 1 の値となるようにスケーリングを行った．

本実験では 3 種類の SVM を作成した．(1)Web 検索結果数のみを用いたもの，(2)Web 検索結果数，ブログ検索結果数，ニュース検索結果数の 3 つを用いたもの，(3)9 個全ての特徴量を用いたものをそれぞれパターン A，パターン B，パターン C として学習および交差検定による精度の検証を行う．直感的に Web 検索結果数が多い有名人であれば知名度が高いと考え

---

(注2)：Yahoo! JAPAN：http://www.yahoo.co.jp/

(注3)：Yahoo!デベロッパーネットワーク
:http://developer.yahoo.co.jp/webapi/search/

(注4)：MeCab：http://mecab.sourceforge.net/

(注5)：LIBSVM：http://www.csie.ntu.edu.tw/ cjlin/libsvm/

表 1　SVM による回帰分析結果

| 特徴量パターン | A | B | C |
|---|---|---|---|
| 平均二乗誤差 | $2.975 \times 10^{-2}$ | $2.827 \times 10^{-2}$ | $2.721 \times 10^{-2}$ |
| 重相関係数 | $3.769 \times 10^{-2}$ | $8.749 \times 10^{-2}$ | $1.251 \times 10^{-1}$ |

表 2　SVM による「篠田麻里子」と「長嶋茂雄」の知名度推定例

| 特徴量パターン | A | B | C |
|---|---|---|---|
| 篠田麻里子 | 83.66% | 86.52% | 88.79% |
| 長嶋茂雄 | 82.34% | 85.36% | 88.99% |

られるため，パターン A をベースラインとし，パターン B はニュース検索結果数およびブログ検索結果数という 2 個の検索結果数は有名人の性質の Web との親和性と話題性に強く影響される特徴量である，という仮定を検証するために用意した．パターン A とパターン B の 2 つに対して 9 個の特徴量全てを用いたパターン C がどの程度結果が向上するかを検証した．

表 1 にそれぞれの 5-交差検定を行った際の，平均二乗誤差と重相関係数を，さらに表 2 に例で用いている「篠田麻里子」と「長嶋茂雄」について知名度を推定した結果を示す．

### 4.3　SVM による結果の考察

表 1 の結果から，特徴量として Web 検索結果数のみを用いたパターン A よりもブログ検索結果数とニュース検索結果数をあわせて用いたパターン B，さらに 9 個の特徴量全てを用いたパターン C の方が平均二乗誤差および重相関係数ともに結果が向上していることが分かり，4.1 節で挙げた特徴量は妥当なものであったといえる．

さらに，強く関係があると推測したニュース検索結果数とブログ検索結果数を用いたパターン B とパターン A では重相関係数が 2 倍以上となり，パターン B とパターン C の間の差よりも大きくなっていることから，ニュース検索結果数とブログ検索結果数が知名度に強く影響していることがわかった．

また,「篠田麻里子」と「長嶋茂雄」では，第 1. 章で述べたように正解となる知名度は「長嶋茂雄」の方が 11.1 ポイント高いが，Web 検索結果数では「篠田麻里子」は 15,500,000 件で，「長嶋茂雄」は 1,530,000 件の 10 倍ほどの検索結果数となっていた．そして表 2 の結果について見ると，パターン A では Web 検索結果数が多い「篠田麻里子」の方が 1.32 ポイント高いが，ニュース検索結果数とブログ検索結果数を考慮したパターン B ではその差が 1.16 ポイントとなっており，わずかながら向上が見られた．さらにパターン C では「長嶋茂雄」の方が 0.2 ポイントと極わずかではあるが結果が上回ることとなり，Web 検索結果数に顕在する情報量の偏りを補正する手法として効果があったといえる．

また，全てのパターンについて表 1 にあるように，平均二乗誤差については 3 種類の SVR においておよそ 0.028 前後であり，誤差は小さいといえる．

しかし，重相関係数の値が非常に低く，4.1 節で説明した特徴量では思うような精度が得られなかった．この理由は，図 4 に示すように正解データ内の不均衡性があったためであると考えられる．

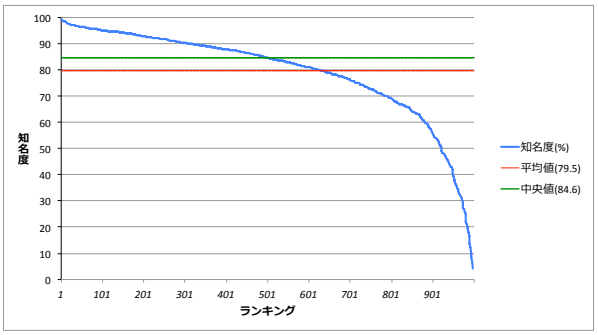


図 4　知名度の正解データ

有名人の現実社会における知名度 1000 人分が正解データとして利用し，知名度は 0 から 1 の値をとるが，図 4 中では 100 をかけることで%表示とした．正解データ中の知名度の平均値は 0.795 であるが中央値は 0.846 となるようなデータとなっている．このようにデータ内に比較的高い値の知名度データが多く存在しており，低い値の知名度データが少なくなっている．

従って，出力をある程度高い数値にすることで平均二乗誤差を最小化する結果となってしまっているため，不均衡を解消する処理を行ってから SVM による学習を行うことでよりよい精度が得られると考えられる．

他の原因として，有名人の名前が一般的な名前である場合，その人物と関係の無い情報を多く取得してしまうことがあり，また，特にブログなどでは通称や愛称によって記述されていることが多く，正確な特徴量を取得することが困難であったことが考えられる．そのため，有名人の別称を考慮した情報抽出 [3] を行う必要があるといえる．

今後は，正解データ内の不均衡性を解消する処理を行い，さらに有名人に関連する情報を正確に取得することで精度の向上を図りたい．また，今回の結果から各特徴量がどれだけ影響を与えたかを詳しく分析し，特徴量に適切な重みを与える方法，および知名度に影響するより多様な特徴量の発見によって，SVM を用いた回帰分析によってよりよい精度を出す方法について検討する．

## 5.　知名度推定のモデル化

本章では，ブログ検索結果数とニュース検索結果数を用いた Web との親和性および話題性の導出モデルを，さらに Web との親和性および話題性と Web 検索結果数を利用した知名度推定モデルを提案する．

### 5.1　有名人の性質の導出モデル

以下では Web との親和性，話題性をそれぞれ導出するためのモデルを提案する．これらは 4 ページの図 3 で説明したとおり，$P$(Web を使う | その人を知っている) にはその人物の Web との親和性が，$P$(情報発信を行う | その人を知っている ∩ Web を使う) には Web との親和性および話題性が影響を及ぼしていると考えられる．

第 4.1 節で説明したように，ある有名人 $Name$ の Web との親和性はブログ検索結果数 $Blog_{Name}$ に，話題性はニュース検索結果数 $News_{Name}$ と $Blog_{Name}$ の 2 つに対して強く影響して

いると推測される．

ここで，Web との親和性を $W_{Name}$，話題性を $H_{Name}$ とし，上記のことより以下に示すモデルを提案する．

$$News_{Name} = p_0 + p_1 \cdot H_{Name} \quad (1)$$

$$Blog_{Name} = p_2 + p_3 \cdot H_{Name} \cdot W_{Name} \quad (2)$$

知名度と Web 検索結果数 $Web_{Name}$ の間には偏りが存在するが，直感的には知名度が高ければ $Web_{Name}$ が多くなる可能性が大きくなると考えられるのと同様に，$News_{Name}$ は $H_{Name}$ が高ければ多くなり，$Blog_{Name}$ は $H_{Name}$ と $W_{Name}$ が高ければ多くなると考えられる．

そこで，第 1 式に示すように，$News_{Name}$ は $H_{Name}$ の線形関数で，また，第 2 式に示すように，$Blog_{Name}$ は $H_{Name} \cdot W_{Name}$ の線形関数によって表されると仮定した．なお，$p_i (i = 0, \cdots, 3)$ をパラメータとする．

上記の 2 式を変形し，下の式を得る．

$$H_{Name} = \frac{News_{Name} - p_0}{p_1} \quad (3)$$

$$W_{Name} = \frac{p_1 \cdot (Blog_{Name} - p_2)}{p_3 \cdot (News_{Name} - p_0)} \quad (4)$$

これら第 3，4 式より有名人の Web との親和性と話題性を求める．

### 5.2 知名度推定モデル

前節では Web との親和性および話題性の導出式を立てた．本節では，知名度推定の定式化を行う．

まず，$N$ を人口とし，$N$ 人の中に何人その人物を知っている人がいるかが現実の知名度となる．知っている人のうち，確率 $P$(Web を使う | その人を知っている) で Web を使い，さらに確率 $P$(情報発信を行う | その人を知っている ∩ Web を使う) で情報発信を行うことによって Web ページが作成される．

また，確率 $P$(Web を使う | その人を知っている) と確率 $P$(情報発信を行う |
その人を知っている ∩ Web を使う) について，$W_{Name}$ は両者に対して，$H_{Name}$ は後者に対して影響を与えると考えられる．ここで，確率 $P$(Web を使う | その人を知っている) $= W_{Name}$，確率 $P$(情報発信を行う |
その人を知っている ∩ Web を使う) $= W_{Name} \cdot H_{Name}$ で得られると仮定すると，確率 $P$($Name$についてある人が情報発信をする) は，

$$P(Name\text{についてある人が情報発信をする}) = W_{Name} \cdot (W_{Name} \cdot H_{Name}) \quad (5)$$

によって求められる．

また，情報発信を 1 回することで検索結果が 1 ページ増えるものとすると，この確率は，言い換えれば $Name$ を知っている人が 1 人当たり何回情報発信を行うかということになる．従って，$Name$ を知っている人が 1 人当たり複数回情報発信を行う可能性もあるため，第 5 式は 1.0 を越える結果となりうる．

以上より，第 5 式から以下に示す，有名人の推定知名度 $F_{Name}$ を計算する式が得られる．

$$F_{Name} = \frac{Web_{Name}}{N \cdot W_{Name} \cdot (W_{Name} \cdot H_{Name})} + p_4 \quad (6)$$

先ほどと同様に，$p_4$ はパラメータとする．

前節の第 3，4 式および 6 式から以下の式が導かれる．

$$(F_{Name} - p_4) \cdot N \cdot p_1 \cdot (Blog_{Name} - p_2)^2 - Web_{Name} \cdot {p_3}^2 \cdot (News_{Name} - p_0) = 0 \quad (7)$$

このように得られた第 7 式に正解データとなる入力，つまり知名度とブログ，ニュース，Web の 3 個の検索結果数を与え，最小二乗法を用いて誤差が最小となるようなパラメータを求める．そして得られたパラメータを用いて未知の入力に対して推定を行う．

## 6. 知名度推定モデルの実験

本章では，提案した知名度推定モデルの実装および評価実験を行い，得られた結果を基に考察を行った．

### 6.1 実験の概要

本実験では 5.2 節で提案した第 7 式の知名度推定モデルに対して，実際に 1000 件の $Name$，$F_{Name}$，$Web_{Name}$，$Blog_{Name}$，$News_{Name}$ を入力シーケンスとして与え，誤差が最小となるようなパラメータ $p_i (i = 0, \cdots, 4)$ を最小二乗法によって求める．また，最小二乗法は与える初期パラメータによって出力が変化するため，初期パラメータとして $p_0$ から $p_4$ にそれぞれランダムな値を与えて交差検定を複数回実験し，その平均を精度として求めた．

また，$Web_{Name}$，$Blog_{Name}$，$News_{Name}$ については各 $Name$ による Yahoo! JAPAN の検索 Web API を用いて検索を行い，検索結果数を取得することでデータを収集した．最小二乗法の計算にはプログラミング言語 Python の科学技術計算ライブラリである Scipy(注6)を利用した．

### 6.2 結果と考察

5-交差検定を 100 回行った際の，平均平均二乗誤差と最小平均二乗誤差を表 3 に，平均二乗誤差が最小となった時のパラメータを表 4，平均二乗誤差が最大となった時のパラメータを表 5 に,「篠田麻里子」と「長嶋茂雄」について最良パラメータを用いて知名度推定を行った結果を表 6 に示す．なお，表 6 中では，得られた 0 から 1 の値をとる知名度に 100 をかけて%表示をしている．

表 3 を見ると，平均二乗誤差の最小値は非常に低い値になっており，6 ページ中の表 1 に示した SVM による回帰分析結果の全てのパターンよりも小さな値となっている．よって，より正確な情報生成について考慮した提案モデルによって効果が得られることが示された

しかし，平均二乗誤差の平均値が非常に悪いものとなっており，これは表 5 にあるようにパラメータ $p_1$ が影響を与えているものと思われる．第 7 式中の $N$ の値は日本の人口である 127,800,000 としており，$p_1$ が大きければ二乗誤差が非常に大きな値となってしまう．これは初期パラメータが正解から非常に離れた値であったためと予想される．

(注6)：http://www.scipy.org/

表 3　平均二乗誤差の平均値および最小値

| | |
|---|---|
| 平均二乗誤差の平均値 | $4.10\times 10^{18}$ |
| 平均二乗誤差の最小値 | $2.51\times 10^{-2}$ |

表 4　最良パラメータ

| | |
|---|---|
| $p_0$ | $5.40\times 10^{-1}$ |
| $p_1$ | -2.39 |
| $p_2$ | $4.94\times 10^{-1}$ |
| $p_3$ | $-8.55\times 10^{-2}$ |
| $p_4$ | $\mathbf{7.95\times 10^{-1}}$ |

表 5　最悪パラメータ

| | |
|---|---|
| $p_0$ | $8.50\times 10^{-1}$ |
| $p_1$ | $\mathbf{6.23\times 10^{5}}$ |
| $p_2$ | $-1.25\times 10^{-3}$ |
| $p_3$ | $2.14\times 10$ |
| $p_4$ | $8.15\times 10^{-1}$ |

表 6　最良パラメータによる「篠田麻里子」と「長嶋茂雄」の知名度推定結果

| | |
|---|---|
| 篠田麻里子 | 79.54% |
| 長嶋茂雄 | 79.54% |

また，表 4 に示す平均二乗誤差が最小となった時のパラメータを見ると，$p_4$ の値が約 0.795 となっている．これは 4.3 節で述べた，正解データの平均値である 0.795 と非常に近い値となっている．従って，$p_0, p_1, p_2, p_3$ によって第 6 式の右辺第一項を 0 に近付けるものになったと考えられ，表 6 の結果からもその影響が強く出ていることが分かる．

これは図 4 に示したように正解データ内の不均衡性の影響を受けており，また，$Name$ それぞれの 3 個の検索結果数，$Web_{Name}$，$Blog_{Name}$，$News_{Name}$ の値にばらつきが大きすぎたためであると推測される．

以上より，本稿で提案した有名人知名度推定モデルが妥当なものではなく，言い換えれば 3 個の検索結果数だけを用いた線形モデルでは有名人の知名度は説明できないということが明らかになった．

今回の結果を踏まえ，今後は 4.1 節で説明した SVM に用いた特徴量など，より多くの要素を考慮し，ニュース，ブログの検索結果数が Web との親和性や話題性による線形関数ではなく，例えば指数関数，対数関数のようなより高度な関数によって表されている可能性について検討していきたい．

また，本研究では有名人の知名度を題材として現実社会と Web との間に存在する情報量の偏りを補正する手法を提案した．今後は有名人の知名度に絞らず，多種多様な性質を持つあらゆる情報に対応できるような，手法についての研究を行うことを検討している．

## 7. おわりに

本研究では，現実社会と Web 上の情報の量の間に存在する偏りを問題として，有名人の知名度を題材に取り組んだ．

偏りの補正を行うための手法として SVM による回帰分析によって知名度を推定する方法と，知名度推定モデルを提案し，それを用いて知名度を計算することで推定する方法を提案した．また，これら 2 種類の手法について実装し，実験によって効果を検証した．

今後の課題・展望として，本研究で行った SVM を用いた回帰分析および知名度推定モデルによる実験の結果を受け，それぞれについて精度を改善する方法について検討する．また，今回提案した手法ではある時点における Web 情報からの知名度推定であるが，今後は情報が伝播し，人々の知識となるまでの時間経過を含めたモデルについて考察を進めることを検討している．さらに，本研究では有名人の知名度を題材に，現実と Web 上の情報量の偏りを補正する研究を行ったが，より幅広い種類の情報について同様の問題を解決する本研究を応用した手法を考案したい．

## 文　献


[1] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Chris J. C, Burges* Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. 1996.

[2] D. Gruhl, David Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explor. Newsl.*, Vol. 6, No. 2, pp. 43–52, December 2004.

[3] Tomoko Hokama and Hiroyuki Kitagawa. Extracting mnemonic names of people from the web. In *Proceedings of the 9th international conference on Asian Digital Libraries: achievements, Challenges and Opportunities*, ICADL'06, pp. 121–130, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pp. 6–14, New York, NY, USA, 2012. ACM.

[5] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[6] Yusuke Yamamoto, Taro Tezuka, Adam Jatowt, and Katsumi Tanaka. Honto? search: estimating trustworthiness of web information by search results aggregation and temporal analysis. In *Proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management conference on Advances in data and web management*, APWeb/WAIM'07, pp. 253–264, Berlin, Heidelberg, 2007. Springer-Verlag.

[7] Lan You, Xuanjing Huang, Lide Wu, Hao Yu, Jun Wang, and Fumihito Nishino. Exploring various features to optimize hot topic retrieval on web. In Fu-Liang Yin, Jun Wang, and Chengan Guo, editors, *Advances in Neural Networks UTF2013 ISNN 2004*, Vol. 3173 of *Lecture Notes in Computer Science*, pp. 1025–1031. Springer Berlin Heidelberg, 2004.

[8] 金英子, 松尾豊, 石塚満. Web 上の情報を用いた企業間関係の抽出. 人工知能学会論文誌 = Transactions of the Japanese Society for Artificial Intelligence : AI, Vol. 22, pp. 48–57, nov 2007.

[9] 立石健二, 石黒義英, 福島俊一. インターネットからの評判情報検索. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2001, No. 69, pp. 75–82, jul 2001.

[10] 平野孝佳, 平手勇宇, 山名早人. 検索エンジンを用いた英文冠詞誤りの検出. 日本データベース学会 letters, Vol. 6, No. 3, pp. 1–4, dec 2007.

[11] 佐藤彩子, 河合栄治, 藤川和利, 砂原秀樹. 個人知識情報を対象とした情報発信・流通モデルの提案. 第 14 回データ工学ワークショップ, 2003.