DEIM Forum 2012 F3-5

人物の呼称を用いたマイクロブログ記事検索に関する一検討

山口裕太郎† 島田 諭†† 佐藤 哲司††

† 筑波大学情報学群知識情報・図書館学類 〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学大学院図書館情報メディア研究科 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{yamaguchi,satoh}@ce.slis.tsukuba.ac.jp, ††sat@slis.tsukuba.ac.jp

あらまし 現在生じている事柄に対して，ユーザが直感的な意見を手軽に投稿可能なマイクロブログでは，人物は人名以外にも様々な呼称で参照されている．呼称は，文脈や投稿者の感情などを反映している記事の中で使用されていることから，単なる人物の参照手段にとどまらない．本研究では，検索エンジン及び Wikipedia から呼称を抽出する手法を提案し，呼称が出現する記事の評価極性やトピックを分析する．分析の結果を踏まえ，人名を入力することで，その人物の呼称が使用される文脈や評価極性などを検索できる，マイクロブログ記事検索システムを実装する．
キーワード マイクロブログ，呼称，SVM

A study of Retrieval in Microblogging based on Person's Aliases

Yutaro YAMAGUCHI†, Satoshi SHIMADA††, and Tetsuji SATOH††

† College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
1–2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
†† Graduate School of Library, Information and Media Studies, University of Tsukuba
1–2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
E-mail: †{yamaguchi,satoh}@ce.slis.tsukuba.ac.jp, ††sat@slis.tsukuba.ac.jp

Abstract In microblogging which the user can easily post comments intuitive, People are referenced in a variety of aliases other than personal names. Aliases is used in the tweets which reflect the context and user's feelings, it's not limited to mere means of referring to the person. In this paper, we propose the method to extract person's aliases using search engine and Wikipedia ,and analyze topic and polarity of the article . Based on the result, we created the system which can retrieve context and polarity of the article in which the person's alias appear when user input a personal name.
Key words Microblogging, alias, SVM

1. はじめに

人物について言及する際に，私たちは，その人の人名以外に様々な呼称を使用することがある．例えば斎藤佑樹投手の場合，ファンは斉藤投手を「ハンカチ王子」や「佑ちゃん」といった呼称で呼ぶことが多い.「ハンカチ王子」は斎藤投手の振る舞いに由来するものであり,「佑ちゃん」は斉藤投手の名前「佑樹」に由来する．また，前田敦子には「あっちゃん」,「顔面センター」といった呼称が存在する.「あっちゃん」と「顔面センター」では，使用される文脈や話題は異なると考えられる．このように，呼称は，単なる人物の参照手段にとどまらず，記事の投稿者の人物に対する印象や，記事の文脈を反映している．

近年，Twitter(注1)に代表されるマイクロブログが注目を集めている．マイクロブログは従来のブログに比べて，ユーザは直感的な意見を短文で手軽に投稿可能であり，記事中には様々な呼称が使用されている．Twitter では，身近な出来事や人物など多くの話題を含む記事が時期を逸せずに投稿されている．人物に関する話題は出来事に関する話題に比べて継続性があると考えられる．ある人物に関する評価や印象はその人物が起こすイベントによって時系列に従い変化する．

現在，Web 上には人に関する情報が溢れている．人物の呼称を収集し呼称の特徴を明らかにできれば，情報検索や，テキストマイニングにおいて呼称は大きな情報源となりうる．

評価極性と投稿数の分析を行い，呼称の潜在的な特徴を推定

(注1)：http://twitter.com/

する手法を確立することを目的とする．本論文では呼称が使用されているマイクロブログ記事の，ポジティブ（肯定的），ネガティブ（否定的）といった評価極性及び，記事の投稿数に着目する．

本論文では人名以外の参照のされ方を呼称と定義し，フリー百科事典 Wikipedia(注2)の人物に関するページに記述されている，多くの人名に対して幅広い呼称を収集する．だが，Wikipedia には使用頻度が小さいものや蔑称に近いものは記述されておらず Web から収集する必要がある．先行研究では呼称を収集するために検索エンジンを利用した方法が提案されている．本研究では呼称を Wikipedia と検索エンジンを併用して収集し，ある期間内の記事の呼称ごとの評価極性と投稿数を分析する．

2. 先行研究

Web から人物の呼称を収集する研究では，呼称候補の抽出法と呼称候補の評価法が課題となっている．外間ら [6] は，「○○こと（人名）」というルールを使用し，「こと」の前の文字列を呼称候補として抽出している．本間ら [11] は，形態素解析に頼らずに，「こと」の直前の文字列の統計情報を元に，呼称候補を抽出している．加えて，「○○こと（人名）」や「（人名）通称○○」などの呼称抽出のルールについても評価し，「○○こと（人名）」が最も有効であるとしている．若木ら [8] は，呼称には名前由来のものと，名前由来でないものがあるとし，2 種類の方法で呼称候補を取得している．名前由来の呼称は，生成ルールを学習させた SVM に，人名と読みを入力することで生成し，名前由来でない呼称は，本間らの手法を踏まえ，文字列の統計情報を用いて抽出している．生成ルールによる呼称生成は，呼称が取得しにくい人物に対して有効であるとしている．Bollegala ら [1] は，呼称候補を，検索エンジンの検索結果とアンカーテキストから取得している．前者は，呼称抽出パターンにマッチする部分以降の，単語 5-gram を取得し，それらの中から，ストップワードのみからなる文字列を除外したものを呼称候補としている．後者では，ある URL へのアンカーテキストで人名と同じ URL を指している頻度の多い文字列の，2-gram を呼称候補としている．

外間ら [6] は，人名と共起しやすい文字列を重み付けし，その表現と，呼称候補をクエリとした際の検索エンジンのヒット数を利用して呼称候補を評価している．本間ら [11] は，ランキング SVM (注3)を使用し呼称候補を評価している．SVM による評価は，呼称候補の出現頻度や人名との共起度を単独で使用するよりも精度が高いと報告している．若木ら [8] は，検索エンジンの検索結果数で呼称候補を評価している．Bollegala ら [1] は，呼称候補の評価に，検索エンジンの検索結果と，アンカーテキストの共起頻度を素性とした SVM を使用している．

呼称を使用して，ブログから人物のトピックを抽出する研究には外間ら [7]，大根ら [9] がある．外間らは「こと」を使用した呼称収集法 [6] で呼称を収集し，人名が使用されている記事と呼称が使用されている記事に，投稿時刻を考慮したクラスタリングを行い，得られるトピックの違いを分析している．投稿時刻を考慮することで，時間軸を持つ現実の出来事に対応したトピックが抽出できたと報告している．さらに、同じトピックに対する記事を，人手で賛成・反対・中立といった観点で分類したところ，呼称と人名が使用される記事では，それぞれの割合が異なったとしている．大根らは，タレントの評判やイメージを抽出する際に，Wikipedia の人物に関するページに記されている呼称を使用している．

マイクロブログに関する研究は数多く行われている．Brendan ら [3] は，マイクロブログ上の評判の変化と，消費者信頼感指数(注4)や，大統領支持率などの社会調査の結果の相関関係を明らかにしている．記事の評判の算出は，記事中のポジティブ・ネガティブな単語の比率を用いる簡便な手法を使用しているが，大統領支持率では 0.725 と高い相関があると報告している．David ら [5] は，一定期間の単語の出現頻度を用いて，マイクロブログストリームからトピック抽出を行なっている．Aniket ら [4] は，マイクロブログ記事を対象に K-means，特異値分解を利用したクラスタリング，グラフベースの Affinity Propagation [2] といった 3 つのクラスタリング手法を比較している．その中では，素性に単語の idf ベクトルのコサイン類似度を用いた Affinity Propagation が有効であると報告している．

以上述べたように，Web 上から人物の呼称を抽出する研究は，数多く行われているが，それらの研究では，呼称抽出の精度向上が主眼となっており，抽出した呼称の印象や特徴の分析は行なわれていない．人名と呼称の特徴を分析している点は，外間ら [7] から着想を得ているが，記事の極性を計算しそれを元に評価している点，とマイクロブログを対象とし，より直感的な意見を考慮している点で既存研究とは異なる．

本研究では検索エンジンに加えて Wikipedia を使用することで，呼称を網羅的に収集する．人名が使われている記事と呼称が使われている記事の評価極性や投稿数から，呼称の特徴を明らかにする手法も，本研究の特徴となっている．

3. 提案法

3.1 呼称抽出

3.1.1 呼称抽出の概要

人物の呼称を網羅的に収集するために，検索エンジンと Wikipedia を組み合わせて呼称を収集する．先行研究では，Wikipedia から収集した呼称は正解データとして用いられている [8] が，本研究では，Wikipedia から収集した呼称と検索エンジンで収集した呼称を組み合わせることで，ある人物の呼称を網羅的に収集する．

提案手法の流れを図 1 に示す．本研究では，マイクロブログ上で現在話題に挙がっている人物の呼称を収集する．以下で呼称を収集する人物の選択手法を説明した後に，提案する検索エ

(注2)：http://ja.wikipedia.org/

(注3)：複数の呼称候補を入力し，呼称候補のランキングを出力する．

(注4)：全米産業審議員会（Conference-Board）が毎月発表する，消費者マインドを指数化した経済指標．現状と 6 カ月後の景況感に関するアンケート調査の結果から算出される．

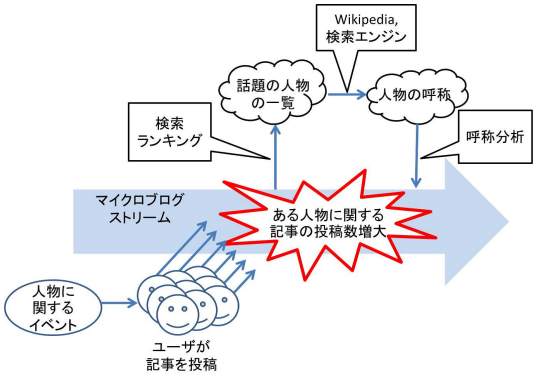

図 1　提案手法の流れ

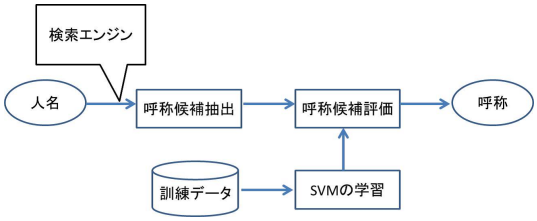

図 2　検索エンジンを用いた呼称抽出の流れ

ンジン及び Wikipedia を使用した呼称の収集法について説明する．

3.1.2　対象人物の選択

マイクロブログ記事を収集する人物の一覧を取得するには，2 つの方法が考えられる．まず，Wikipedia に存在する，日本のタレントや存命人物カテゴリなどを利用して，カテゴリ以下に属する人名を，網羅的に取得する方法である．この方法では，収集した人物が全て現在話題に挙がっている訳ではない，という問題がある．次に，検索ランキングを使用する方法がある．網羅性は前者の方法に劣るが，現在話題になっている人物の一覧を取得可能である．本研究では，マイクロブログという今を反映するメディアを対象とすることから，検索エンジンのランキングを利用する．

対象人物の呼称を，Wikipedia と検索エンジンを用いて抽出し，呼称と人名をクエリとしてマイクロブログストリームから，現在話題に挙がっている人物の記事を収集する．

3.1.3　Wikipedia からの呼称抽出

人物の呼称で広く知られているものは，人物に関する Wikipedia の文章中に記載されている．それらの呼称を，Wikipedia の当該人物のページから正規表現を用いたパターン照合で抽出する．

3.1.4　検索エンジンによる呼称候補抽出

Wikipedia に掲載されていない蔑称や知名度の低い呼称は，外間ら [6]，本間ら [11] の手法を踏まえ，検索エンジンを利用して抽出する．図 2 に呼称抽出の流れを示す．呼称抽出は，呼称候補の抽出部分と呼称候補の評価部分からなる．以下で呼称候補の抽出法について説明し，3.2 節で呼称候補の判定法について説明する．

文字列 $alias$ が，人名 $fullname$ の人物の呼称であることを述べる際,「こと」で呼称と人名を連結し,「$alias$ こと $fullname$」

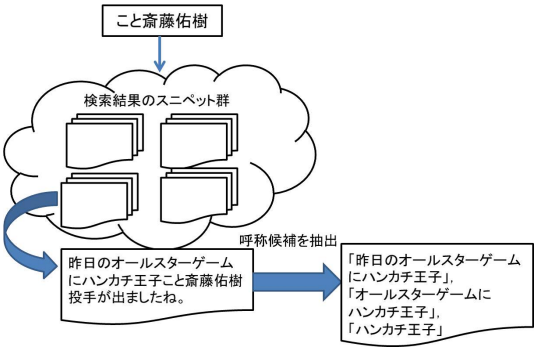

図 3　呼称候補抽出の例

と表現するというルール [6] を使用して呼称候補を収集する．例えば，文字列「ハンカチ王子」が人名「斎藤佑樹」の呼称である場合は,「ハンカチ王子こと斎藤佑樹」と表現される．呼称と人名を紐付ける表現は,「$alias$ こと $fullname$」以外にも「$fullname$ 通称 $alias$」や「$fullname$ 愛称 $alias$」などが考えられる．本間ら [11] による，複数の表現の呼称候補獲得能力の評価結果において，他の表現に比べて呼称と人名を紐付ける性能が良いことから，本研究では「$alias$ こと $fullname$」のみを使用する．以下で，詳細な手順について述べる．

(1)　「こと $fullname$」をクエリとして Web ページを検索し，スニペット（N 件）を取得する．

(2)　スニペットを形態素解析し,「こと $fullname$」の直前文字列（5 形態素）を取得する．取得した文字列の部分文字列のうち，スニペット全体での出現回数が 3 回以上のものを，呼称候補 $candidate$ とする．

$fullname$ を斎藤佑樹投手として呼称候補を抽出した例を図 3 に示す.「こと斎藤佑樹」をクエリとして，スニペットを取得する．スニペット内に,「昨日のオールスターゲームにハンカチ王子こと斎藤佑樹投手が出ましたね。」という文字列が存在する場合,「こと斎藤佑樹」の直前の 5 形態素（「昨日」「の」「オールスターゲーム」「に」「ハンカチ王子」）を取得する．取得した文字列の部分文字列のうち名詞で始まる「昨日のオールスターゲームにハンカチ王子」,「オールスターゲームにハンカチ王子」,「ハンカチ王子」を呼称候補とする．

3.2　SVM を用いた呼称候補の判定

本研究では SVM を用いて，呼称候補の判定を行う．Support Vector Machine（SVM）は，データを分類する際に，マージン最大化によって最適な分離平面を得る学習手法である．SVM で分類を行うには，訓練データを用いて学習を行う必要がある．ここでは，適切な呼称候補を正例，不適切な呼称候補を負例とする．訓練データは，複数の人物の呼称候補を 3.1.4 節の方法で取得し，人手で判定し作成する．

本間ら [11] は，入力を順位付けて出力するランキング SVM で呼称候補を評価している．「$alias$ こと $fullname$」を含めて 6 つのパターンで呼称候補を抽出し，素性には $Dice(fullname, candidate)$，$OverlapC(fullname, candidate)$，$OverlapN(fullname,$

$candidate$)，呼称候補の抽出時に用いた 6 パターンでの候補が出現した回数，の 9 つを使用している．

$$Dice(fullname, candidate) = \frac{Hits(fullname, candidate)}{Hits(name) + Hits(candidate)} \quad (1)$$

$$OverlapC(fullname, candidate) = \frac{Hits(fullname, candidate)}{Hits(candidate)} \quad (2)$$

$$OverlapN(fullname, candidate) = \frac{Hits(fullname, candidate)}{Hits(name)} \quad (3)$$

$Hits(name, candidate)$ は「$name$ AND $candidate$」で検索した際のヒット数であり，$Hits(name)$，$Hits(candidate)$ はそれぞれ「$name$」,「$candidate$」で検索した際のヒット数である．

本研究では，本間らの素性を参考に以下の 8 つの素性で特徴ベクトルを作成し，分類器は SVM を使用し，呼称候補が呼称として適切かどうかを判定する．

(1) $Dice(fullname, candidate)$

(2) $OverlapC(fullname, candidate)$

(3) $OverlapN(fullname, candidate)$

(4) 文字列「こと $candidate$」での検索結果のスニペット内の呼称候補 $candidate$ の頻度の対数 $\log(cf(candidate))$

(5) 文字列「$candidate$」での検索結果のスニペット内の呼称候補の前の名詞の出現頻度 $fn(candidate)$

(6) 文字列「$candidate$」での検索結果のスニペット内の呼称候補の前の助詞の出現頻度 $fp(candidate)$

(7) 文字列「$candidate$」での検索結果のスニペット内の呼称候補の後の名詞の出現頻度 $bn(candidate)$

(8) 文字列「$candidate$」での検索結果のスニペット内の呼称候補の後の助詞の出現頻度 $bp(candidate)$

表 3.2 に示すように先行研究とは，分類器や呼称候補抽出に使用するパターンが異なるため，単純に精度を比較することはできないため，検索結果数を利用した指標に加えて，呼称候補前後の品詞情報を考慮することが呼称候補の評価の精度の向上に寄与することを示す．

3.3 呼称分析

本節では，3.1 節で提案した方法で収集した呼称の分析手法について述べる．本研究では，呼称が使用される記事の文脈を明らかにすることで呼称の分析を行う．具体的には，呼称が出現する記事の評価極性と，投稿数を分析する．

3.3.1 記事の評価極性の算出

東山 [10] らの作成した，日本語評価極性辞書（名詞編）を使用して記事の評価極性を算出する．辞書には評価極性を持つ名詞約 8,500 表現に対してポジティブ・ネガティブ・中立の評価極性情報が付与されている．

ある記事 $tweet$ の評価極性は次の式で算出する．

$$pn(tweet) = \frac{|posi| - |nega|}{|posi| + |nega|} \quad (4)$$

$|posi|$ は，一つの $tweet$ に出現するポジティブな名詞の延べ数，$|nega|$ は一つの $tweet$ に出現するネガティブな名詞の延べ数であり，$pn(tweet)$ は $(-1.0 \leqq pn(tweet) \leqq 1.0)$ の範囲の値をとる．例えばある $tweet$ にポジティブな名詞「可愛さ」が 3 回出現し，ネガティブな名詞「可哀想」が 1 回出現した場合，式 4 の値は 0.5 になる．

3.3.2 ある期間内の呼称の評価極性の算出

呼称によってはネガティブな記事でのみ使用されるものや，反対にポジティブな記事でのみ使用されるものがあると考えられる．呼称と評価極性の関連を明らかにするために，ある期間内での呼称が使用されている記事全体の評価極性を算出する．ある人物の呼称 $alias$ の，ある日 $date$ の評価極性 $pn\_alias(alias, day)$ は，以下の式で算出する．

$$pn\_alias(alias, date) = \frac{1}{n} \sum_{tweet \in T} pn(tweet) \quad (5)$$

ここで T は，ある人物の呼称を用いて $date$ に投稿された記事集合，n は T の要素数，すなわち期間 $date$ 内で投稿された記事数である．$pn\_alias(alias, date)$ は呼称 $candidate$ を使用して，ある期間内に投稿された記事の $pn(tweet)$ の平均を表している．$pn\_alias(alias, date)$ も式 4 同様に $(-1.0 \leqq pn\_alias(alias, day) \leqq 1.0)$ の範囲の値を取る．

4. 評価実験

4.1 評価の概要

3. 章で提案した呼称収集・記事分析手法を実装し，評価実験を行い，手法の有効性を確認する．呼称抽出について述べた後に，呼称の分析について述べる．

4.2 呼称抽出

Wikipedia の存命人物カテゴリ(注5)の人名を対象に 3.1 節で述べた手法で各人物について，呼称候補を抽出する．呼称候補に人手でラベルを付与し正例と負例各 500 件ずつからなる実験データを作成した．

本研究の課題の 1 つとして，検索エンジンを用いた呼称候補の収集が挙げられる．そこで，適切な呼称をどれだけ収集できたかを評価する．SVM による呼称候補抽出の評価尺度には，情報検索で用いられる適合率，再現率，F 値を用いる．評価値は正例・負例ごとに算出する．各尺度の算出式を以下に示す．

(注5)：http://ja.wikipedia.org/wiki/Category:存命人物

表 1　先行研究との比較

	本間ら	提案手法
分類器	ランキング SVM	SVM
分類器への入力	呼称候補	呼称候補
分類器の出力	呼称候補のランキング	呼称候補が適切か（2 値分類）
呼称候補抽出のパターン	「こと $fullname$」を含め 6 種類	「こと $fullname$」のみ

$$Precision = \frac{R}{N}$$
$$Recall = \frac{R}{C} \qquad (6)$$
$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

ここで，R は SVM が正しく分類した正例（負例）の数，N は SVM が出力した正例（負例）の数，C は評価用データの正例（負例）の数である．

評価値は，10 分割の交差検定を行い算出する．

4.3　呼 称 分 析

大場美奈，加護亜依の 2 名の呼称について分析を行った．表 2 に使用した呼称の一覧を示す．分析には，図 4 に示す方法で収集した記事を用いる．Twitter の Search API から表 4 に示す条件で収集した記事集合に対して，人名と各呼称，それぞれをクエリとして検索した結果のうちで投稿日時が 2011 年 8 月 27 日 0:00:00 から 9 月 20 日 23:59:59 のものを投稿数の分析対象とした．

投稿数の分析対象の記事集合から本文に RT，URL を含むものを除いたものを評価極性の分析対象に用いた．なお，呼称をクエリにして得られた記事のうちで明らかに対象人物以外を参照している記事は除外した．

表 2　分析に用いた呼称

人名	呼称
大場美奈	みなるん，コロちゃん
加護亜依	加護ちゃん，あいぼん

表 3　各呼称の記事数

人名（呼称）	記事数	RT・URL 除外後
大場美奈	224	70
みなるん	346	313
コロちゃん	3	3
加護亜依	1127	259
加護ちゃん	1389	985
あいぼん	53	34

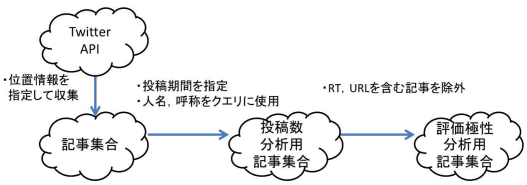

図 4　記事収集の流れ

表 4　収 集 条 件

収集期間	2011 年 6 月 27 日 0:00:00 から 2011 年 9 月 26 日 0:00:00 まで
位置情報	geocode = ’35.67012719,139.8094368,100km’

5.　結果と考察

5.1　SVM による呼称の判定

3.1.4 節で述べた，呼称候補抽出の際のパラメータ N は 300 とした．SVM は機械学習の分野で広く使用されているツールの LibSVM(注6)を使用する．カーネル関数と SVM のタイプはデフォルトの RBF カーネルと C-SVC を使用した．カーネル関数のパラメータ γ と，コストパラメータ C は，任意の範囲を格子点検索をする gridsearch で求めた．パラーメータの値を表 9 に示す．

excite トレンドトラッカー(注7)の人物カテゴリのランキング及び，NAVER 人物検索の今、話題の人物(注8) に 9 月 1 日から 9 月 15 日の間に掲載された人物 867 名に対して提案手法で呼称抽出を行った．SVM で使用する訓練データは 4.2 節で作成したデータを用いた．Wikipedia から抽出した呼称は，正規表現でを用いて取得した呼称のうち，明らかに誤りであるものを人手で取り除いたものを用いた．抽出できた呼称数を表 5 に，呼称を抽出できた人物数を表 6 に示す．表 5 に示すように，検索エンジンと Wikipedia で共通して収集できた呼称は少なく，両者を組み合わせることで，網羅的に呼称を収集可能である．

提案手法で取得した呼称の一部を表 7，表 8 に示す．検索エンジンを用いることで，東原亜希の呼称「デスブロガー」,「デスブログ」などの蔑称に近い呼称を抽出できた．加えて，入江陵介，北山宏光，氷川きよし，など Wikipedia に呼称が掲載されていない人物の呼称も，収集できた．

分類を誤った例を以下に示す．例 1 から例 4 は false negative（負例を正例と判定），例 5 は false positive（正例を負例と判定）である．

例 1　Web サービスでのアカウント名（faridyu，@faridyu）

例 2　「こと」が人名と呼称の関係以外を意味する（シンケンレッド，いとうあさ，娘のために産む）

例 3　形態素解析の誤り（くん，王子）

(注6)：http://www.csie.ntu.edu.tw/ cjlin/libsvm/
(注7)：http://tt.excite.co.jp/people/
(注8)：http://person.naver.jp/issue

例 4　正しい呼称を含む連接（チーム 4 のみなるん）

例 5　「○○の××」の形の文字列（瑛太の妻，演歌界のファッションリーダー）

例 1 の「faridyu」，はダルビッシュ有の Twitter のアカウント名を呼称と判別している．特定の Web サービスのアカウント名は，出現する Web ページのドメインに偏りがあると考えられ，分類器の素性に呼称候補を抽出した URL の情報を用いることで，正しく分類できると思われる．

例 2 は,「こと」が人物と呼称の関係を表す文脈以外にも出現することが原因である.「シンケンレッド」は「こと」が役名と役者名の関係を表す際にも使用されることから抽出された.「いとうあさ」及び「娘のために産む」は「こと」が人名の一部や名詞として用いられるため抽出された．これら,「こと」の多義性が原因の問題は，呼称候補の抽出に「こと」以外の表現を使用することで解決可能であると考えられる．

例 3 の「王子」は,「童顔王子」や「バランス王子」といった呼称が「童顔」と「王子」,「バランス」と「王子」といった 2 形態素に分割されたことが原因である．形態素単位の呼称候補抽出に加えて，文字単位の抽出も併用することで改善可能であろう．

例 4 の「チーム 4 のみなるん」は，正しい呼称「みなるん」を修飾する句まで含んでいるため，呼称ではない．例 5 の「瑛太の妻」は,「チーム 4 のみなるん」と類似しているが,「妻」のみでは，ある特定の人物の呼称とは成り得ず,「瑛太の」という名詞句によって修飾されて，初めてある人物を指し示す.「演歌界のファッションリーダー」も同じ形式である．例 4 と例 5 の分類は,「人名以外の参照のされ方」という本研究における呼称の定義では，人手でも揺れが生じる可能性がある．今後の課題として心理学や社会学の分野の成果を利用し，呼称の定義の明確化が挙げられる．

表 5　抽出できた呼称数

検索エンジン	Wikipedia	共通
824 件	464 件	80 件

表 6　呼称を抽出できた人物数

検索エンジン	Wikipedia	共通
428 名	229 名	168 名

表 7　検索エンジンで取得した呼称の例

人名	検索エンジン
入江陵介	王子，カエル王子，バランス王子，背泳ぎ王子
大場美奈	チーム 4 のみなるん，コロちゃん
加護亜依	あいぼん
北山宏光	くん，王子，ミツ，童顔王子，キタミツ
後藤真希	ゴマキ ごっちん
ダルビッシュ有	faridyu，@faridyu
土屋アンナ	あんさん
氷川きよし	平成の股旅野郎，股旅野郎，野郎，若様
東原亜希	デスブロガー，ブロガー，デスブロク，デスブログ
松坂桃李	レッド，シンケンレッド

表 8　Wikipedia から取得した呼称の例

人名	Wikipedia
入江陵介	
大場美奈	みなるん，コロちゃん
加護亜依	あいぼん，加護ちゃん
北山宏光	
後藤真希	ゴマキ
ダルビッシュ有	ダル
土屋アンナ	アンさん，アンナ
氷川きよし	
東原亜希	ひがが，あきあき，あき，ひがし
松坂桃李	

表 9　パラメータ

	γ	C
品詞情報無し	1.34×10^8	0.25
提案法	1.00	0.50

表 10　品詞情報無し

	Precision	Recall	F-measure
正例	0.69	0.65	0.67
負例	0.62	0.67	0.64

表 11　提　案　法

	Precision	Recall	F-measure
正例	0.83	0.70	0.76
負例	0.65	0.80	0.71

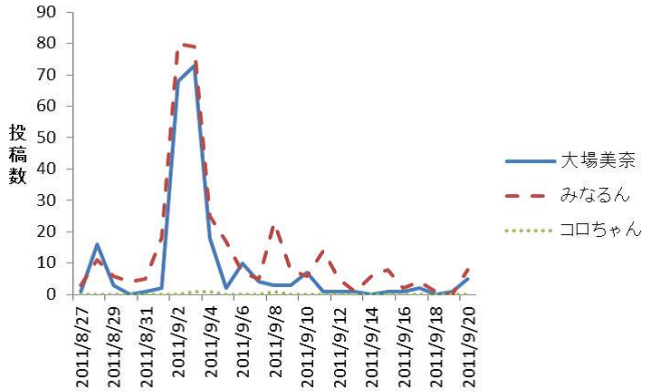

図 5　投稿数の推移:大場美奈

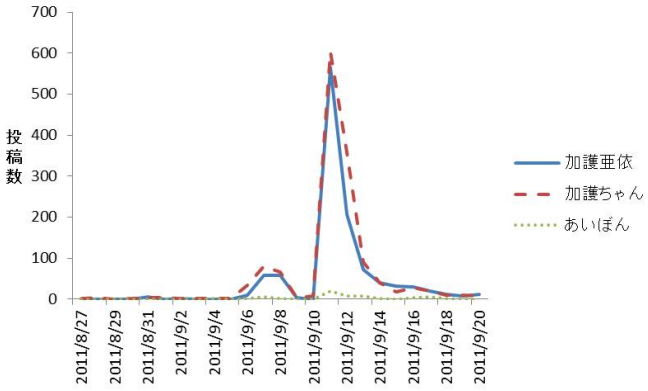

図 6　投稿数の推移:加護亜依

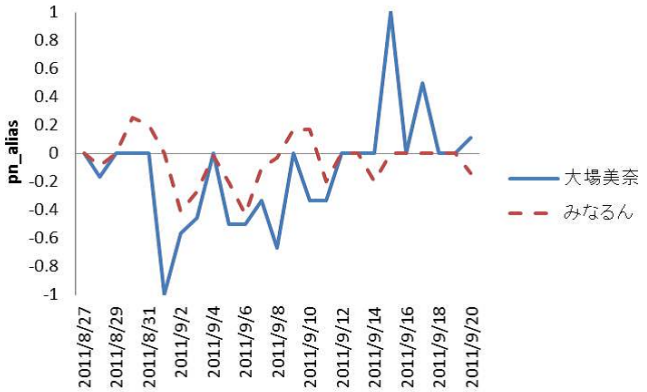

図 7　評価極性の変化:大場美奈

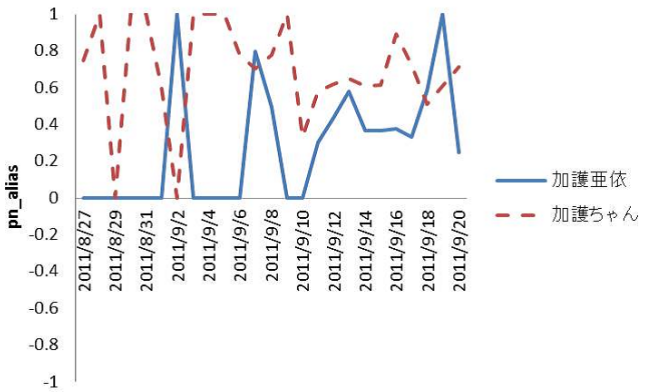

図 8　評価極性の変化:加護亜依

5.2　投稿数の推移

呼称ごとの投稿数の推移を図 5，図 6 に示す．

図 5，図 6 から，呼称でも十分に記事が投稿されていることが確認できる．図 5 における，9 月 2 日，3 日と図 6 の 9 月 10 日から 13 日が，投稿数のピークとなっている．これらの期間では大場美奈，加護亜依ともに比較的大きなイベントがあり，人名，呼称を問わず多くの記事が投稿されている原因となっている．

また，図 5 の 9 月 8 日と図 6 の 9 月 7 日，8 日では，人名と呼称で投稿数に差が見られる．これは，呼称でしか記事が投稿されない出来事が存在するためであると考えられる．特に，図 5 の 9 月 8 日の投稿は，9 月 2 日に謹慎を表明した大場美奈が，所属しているアイドルグループのレギュラー番組「AKBINGO!」に出演したことに関する投稿が多数を占めていた．この番組の視聴者にはアイドルグループのファンが多いと思われるため,「みなるん」を使用して記事を書く投稿者は，大場美奈を擁護する傾向があるといえる．

5.3　評価極性の変化

呼称ごとの評価極性の変化を図 7，図 8 に示す．グラフの縦軸は，式 5 で算出される各呼称の pn_alias の値である．記事が投稿されていない場合は pn_alias の値は 0 とした．また，十分に記事が投稿されていない，加護亜依の呼称「あいぼん」と大場美奈の呼称「コロちゃん」は除外した．

図 7，図 8 から評価極性は，呼称と人名では異なる値を取っていることがわかる．分析に使用した呼称では人名に比べてポジティブな記事の割合が多い傾向を示した．これは，今回使用した呼称「みなるん」,「加護ちゃん」が人名に比べて親密さやポジティブなイメージを持つためだと考えられる．呼称の評価極性は，1 日に投稿された記事の評価極性の平均で算出しているため，投稿数が少ない日付では極端な値になりやすい．加えて，現状では単なる単語の数え上げで評価極性を算出しているため，評価極性の算出手法の改善が求められる．

5.4　呼称を利用したマイクロブログ検索システムの考案

呼称を利用したマイクロブログ検索システムの概要を図 9 に，システムのインタフェースを図 10 に示す．実装環境は，クライアントの処理には jQuery を使用し，サーバは PHP 及び，MySQL を使用した．ユーザが選択した人物の呼称を，事前に作成した呼称 DB から取得し，人名と呼称を用いて Twitter API から記事を検索して提示する．呼称 DB の要素には，5.1 節で収集した 489 名，1208 件の呼称を登録した．人物に関する情報検索において，人名に加えて呼称を使用することで，幅広い話題の取得や同一の話題に対する多様な意見の収集が期待できる．

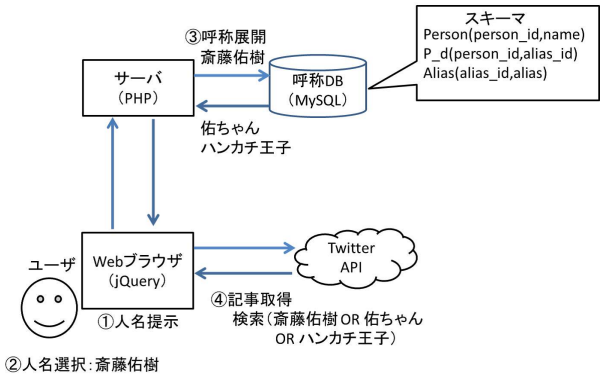

図 9　システム概要

図 10　ユーザインタフェース

6. おわりに

本論文では,「人名以外の参照のされ方」を呼称と定義し，人物に付けられた様々な呼称が持つ特徴を明らかにすることを目的に，呼称抽出及び呼称分析手法を提案した．SVM による呼称候補の分類の評価実験の結果，呼称候補の前後の品詞の頻度を素性に加えた提案手法では，正例で適合率 0.83 の精度を示し，品詞情報を使用しない場合に比べて精度が向上した．抽出した呼称を使用して，記事の投稿件数を分析したところ，呼称でも十分に記事が投稿されていることが確認できた．また，投稿件数のピークが人名と呼称で異なる場合があることが明らかになった．これは，出来事によっては呼称でしか投稿されないものがあるためと考えられる．記事の評価極性は，呼称と人名では異なる値を示した．分析に使用した呼称では，人名に比べてポジティブな記事の割合が多い傾向を示した．

今後の課題として，心理学や社会学の分野の成果を利用した人物の呼称の定義の明確化や，ポジネガ以外の呼称の特徴の検討，システムの改良などが挙げられる．

謝　辞

本研究の一部は科研費（21500091）の助成を受けたものである．ここに記して謝意を示す．

文　献

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka. Automatic discovery of personal name aliases from the web. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 23, No. 6, pp. 831–844, june 2011.

[2] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, Vol. 315, pp. 972–976, 2007.

[3] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A.Smith. From tweets to polls:linking text sentiment to public oppinion time series. *ICWSM-2010*, 2010.

[4] Aniket Rangrej, Sayali Kulkarni, and Ashish V. Tendulkar. Comparative study of clustering techniques for short text documents. *20th International World Wide Web Conference (WWW2011)*, p. 111, 2011.

[5] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 355–358, 2011.

[6] 外間智子, 北川博之. Web コーパスを用いた人物の呼称抽出. 夏のデータベースワークショップ (DBWS2006), 2006.

[7] 外間智子, 北川博之. blog からの人物に関する呼称を用いたトピック抽出. 第 18 回データ工学ワークショップ (DEWS2007), 2007.

[8] 若木裕美, 藤井寛子, 福井美佳, 住田一男. Web 情報を用いた人物の愛称抽出手法. 第 19 回データ工学ワークショップ（DEWS2008）, 2008.

[9] 大根千明, 松尾豊, 木戸冬子, 勝芳邦, 石塚満. Weblog からのタレントに関する好感度情報抽出. 第 22 回人工知能学会全国大会, 2008.

[10] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. pp.584–587, 2008.

[11] 本間大輝, Danushka Bollegala, 松尾豊, 石塚満. Web を用いた人物の別名抽出. NLP 若手の会第 2 回シンポジウム, 2007.