blogからの人物に関する呼称を用いたトピック抽出

外間　智子†　　北川　博之†,††

† 筑波大学システム情報工学研究科　〒 305-8577 茨城県つくば市天王台 1 – 1 – 1
†† 筑波大学 計算機科学研究センター　〒 305-8577 茨城県つくば市天王台 1 –1– 1
E-mail: †{tomokoh,kitagawa}@kde.cs.tsukuba.ac.jp

あらまし　口コミ的な性質の強いメディアである blog 上では、人物はフルネーム以外に種々の呼び名（呼称）で参照される. 呼称はそれを用いる書き手の、その人物に対する印象や評価を反映しており、フルネームとは違った文脈で用いられることも多い. 本研究では、Web から対象人物に対する呼称を抽出し、次に、人物がフルネームで参照されている Blog 記事群と、呼称で参照されている Blog 記事群をそれぞれ Blog 検索エンジンを用いて収集する. さらに文書クラスタリングを用いて各々の記事群からトピックを抽出し、抽出されるトピックにどのような違いが見られるか、また実際にフルネームと呼称では用いられる文脈が違ってくるのかについて分析する.
キーワード　ブログ, オブジェクト識別, トピック検出

Topic Detection about People from Blogsphere by Using Mnemonic Names

Tomoko HOKAMA† and Hiroyuki KITAGAWA†,††

† Graduate School of Systems and Information Engineering, University of Tsukuba　Tennodai 1–1–1, Tsukuba-shi, 305–8577 Japan
†† Center for Computional Sciences, University of Tsukuba　Tennodai 1–1–1, Tsukuba-shi, 305–8577 Japan
E-mail: †{tomokoh,kitagawa}@kde.cs.tsukuba.ac.jp

Abstract　People are referred to by various ways in weblog articles. For example, a person may be referred to by the full name, affiliation and title, or nicknames. These mnemonic names reflect writers' images or opinions for the person. In this paper, we extract mnemonic names for the target person. Then we collect weblog articles that include the full name and weblog articles that include each mnemonic name separately. After collecting weblog articles, we extract topics from each article set and analyse differences between extracted topics from each article set.
Key words　Blogspace, Object Identification, Topicc Detection

1. はじめに

インターネット技術の発達は、個人に、大手メディアを介さず情報を発信する多くの手段をもたらした. 中でもここ数年で爆発的に普及した blog は、世の中の関心をリアルタイムに反映する新しいメディアとして注目されるようになってきた. こうした背景のもと、blog からの有用な知識抽出を目指した様々な研究が進められている [5]〜[7], [12], [13].

一方で、組織、製品、人物などある特定のオブジェクトに関する Web からの動的かつ非公式な情報抽出・知識抽出への要求が高まってきている. その代表的なものとして、評判情報抽出 [14], [15] があげられる. blog は、こうした動的かつ非公式な情報抽出・知識抽出に非常に適した情報源であるといえよう.

口コミ的な性質の強い blog では、1 つのオブジェクトが正式名称以外の様々な呼び名で参照されることが多い. 例えば人物の場合、参照のされ方はフルネーム、姓のみ、肩書、ニックネームなど様々である. 本研究では、こうしたオブジェクトの「参照のされ方」を「呼称」と呼ぶ. ただし、自明な正式名称は「呼称」に含まない.

呼称の多くは動的で非公式なものであり、機械的に発見することは一般に容易ではない. 我々は、[16] において、Web から人物の呼称 (フルネーム以外の人物の参照のされ方) を抽出する手法を提案した. 呼称はそれを用いる書き手の、その人物に対する印象や評価を反映しており、フルネームとは違った文脈で用いられる. 例えばニュース記事等公式な文書では、人物を参照するのにフルネームが用いられるが、匿名の個人が書く blog

記事中ではフルネームはあまり使われず、一般的によく知られた呼称が用いられることが多い. したがって、blog からいわゆる「生の声」「評判」を抽出するためには、人物がフルネームで参照されている記事だけでなく、呼称で参照されている記事にも注目することが必要である.

また、呼称にも、好意的な文脈で用いられる呼称と批判的な文脈で用いられる呼称があり、各々の呼称がどのような文脈で用いられるのかを知ることは、その人物に関する評判を知るうえで非常に重要である. 例えば、[16] において前総理大臣「小泉純一郎」に関する複数の呼称が抽出されたが、その中には「純ちゃん」といった好意的な呼称と、「ポチ」「変人」といった蔑称的な呼称が含まれていた. これらは Web 全体より抽出された呼称であるが、blog においても書き手は、「小泉純一郎」に関して言及する際に、自身の立場や文脈に応じてフルネームや呼称を使い分けているであろう.

本研究では、人物のフルネーム及び [16] で抽出した呼称を含む記事群からそれぞれトピックを抽出し、抽出されるトピックの違いを分析する. 具体的には、まず目的人物の呼称を [16] の手法を用いて抽出し、次にフルネーム及び抽出した呼称で目的人物が参照されている blog 記事を収集する. さらに、収集した記事群をクラスタリングし、抽出されるトピックの違いを観察する. なお、トピックの時間的な遷移や、同じ時点におけるフルネームと呼称の出現する文脈の違いをみるため、数ヶ月程度の対象期間を設定し、その期間中に書かれた記事を用いてトピックの抽出を行う.

また、情報源となる記事の収集について、自分でクローリングを行う方法と、既存の blog 検索エンジンを利用する方法の 2 通りが考えられる. Web のクローリングには膨大なコストがかかる点、また最近では Web サービスとして blog 検索 API が公開され始めている点を考慮し、本研究では記事収集に既存の blog 検索エンジンを用いるものとする.

2. 関 連 研 究

2.1 トピック検出

新聞記事より話題の抽出を行うものに、TDT(Topic Detection and Tracking) の手法がある。タイムスタンプが付加されたニュース記事からトピックを検出する試みであり、文書クラスタリング手法を用いた手法など、いくつかのアプローチがある. 蓄積した過去の記事が利用可能であり、リアルタイム処理を目的としないのであれば、階層的クラスタリングに時間的要素を組み合わせた手法が高い精度を示すことが報告されている [9].

クラスタリングを用いずに記事集合からトピックを検出する手法として、他に burst の検出がある。これは、すべての単語の定常状態の出現頻度を保持しておき、それから大きく外れて高頻度で出現する単語をトピックワードとして提示する手法である [10], [13]. burst 情報を利用してトピックを検出する手法 [8] も提案されているが、burst を検出するためには大量のコーパスが必要となる. 本研究では特定の人物に関する話題を抽出することが目的のため、必要な量のコーパスが各対象について用意できるとは考えにくく、従って単語の burst を利用する手法の適用は難しい.

2.2 オブジェクト識別

データクリーニングや異種情報源統合のために、これまで、異なるデータベース上の重複レコードを抽出する Duplicate detection に関する多くの研究がなされてきた [2], [3]. その応用として、最近では、Web 上の異なる情報源を統合するため、表記のゆれを解決するオブジェクト識別の研究が行われている [1]. これらは基本的に、レコード属性の類似度を利用しており、データベースのスキーマや HTML のテーブルタグ等のメタデータを必要とする.

またプレーンテキスト (Web 文書) を対象としたものとして、同姓同名の人物を識別する研究もこれまでにいくつかなされている. [11] は、文書クラスタリングとプロファイル情報を利用し、Web 検索時に同姓同名の人物を識別する試みである. 本研究では、これら同姓同名の識別とは逆に、テキストからある人物に関する別の呼び名を抽出する.

3. blog からの人物に関するトピック抽出

3.1 トピック抽出の概要

本節では、トピック抽出の概要について述べる.

人物に関するトピック抽出は、以下の 3 フェーズで行う.

フェーズ 1 目的人物の呼称抽出

フェーズ 2 検索ワードの選択と blog 記事収集

フェーズ 3 時刻印を考慮したトピック抽出

フェーズ 1 では、目的人物のフルネームを手がかりにして、Web から呼称を抽出する.

次にフェーズ 2 では、フルネームおよび抽出した呼称を blog 検索エンジンにキーワードとして与え、blog 記事を収集する. その際に、以下のような問題がある. まず、呼称が一般名詞であったり、同じ呼称が複数の人物に対して用いられている場合、呼称はそれ自体では目的人物に関する記事を同定できない. また多くの blog 検索エンジンは新着順に記事を返し、かつ返す検索結果数には制限があるため、単純にフルネーム・呼称のみで検索すると、直近に書かれた記事ばかりが得られてしまう. そのため、目的人物について記述しており、かつ対象期間中に書かれた十分な量の記事を得るためには、検索クエリに何らかの別のキーワードを選択する必要がある. 以下、クエリに追加するキーワードを検索ワード、と呼ぶ.

フェーズ 3 では、フルネームと各呼称それぞれについて収集した blog 記事群から、トピックを抽出する. ここでは、我々が [17] で提案した、タイムスタンプを考慮したクラスタリングを用いる.

以下、3. 2, 3. 3, 3. 4 節で、各フェーズの詳細について述べる.

3.2 目的人物の呼称抽出

呼称抽出の詳細については、文献 [16] で述べた. 本節では、その概要を簡単にまとめる.

呼称抽出は、以下の 2 つのヒューリスティクスを背景とする。

（1） 文字列 *alias* が、*fullname* という名前の人物の呼称であることを述べる際、日本語では “*alias*「こと」*fullname*” と

表現する。(注1)

（2） 人物のフルネームと呼称は、同様なコンテクスト中に出現することが多い。ここで、コンテクストとはフルネーム及び呼称に隣接する文字列のことを指す。すなわち、Web コーパス中に文字列 *prefix*+*fullname*(注2), *fullname*+*suffix* があったとき、*fullname* の部分を *alias* と置き換えた文字列 *prefix*+*alias*, *alias*+*suffix* という文字列も Web コーパス中に出現する可能性が高い。

これらを踏まえ、次の手順で Web より人物の呼称を抽出する。

（1） “こと *fullname*” の直前に出現する文字列を Web ページ群より取得し、呼称候補 *cand* とする. (例: 文字列「日本のクールビューティーこと荒川静香」より、“日本のクールビューティー”, “クールビューティー”, “ビューティー” を呼称候補として抽出する.)(ヒューリスティクス 1)

（2） フルネームの直前に出現する文字列を prefix パターン、直後に出現する文字列を suffix パターンとする. フルネームを含む Web ページ群より prefix パターン及び suffix パターンを抽出する. (例: 文字列「トリノオリンピックで荒川静香が金メダルを獲得」より、“トリノオリンピックで” を prefix パターン、“が金メダルを獲得” を suffix パターンとして抽出する.) 次にそれぞれのパターンにフルネームとの共起度を考慮した重みを与え、重みの大きいものを「隣接パターン」として選択する.

なお上記手順 (1), (2) は独立した処理のため、並列処理可能である.

（3）「隣接パターン」を用いて呼称候補を評価する. 具体的には、ある呼称候補 *cand* を評価するために、*prefix*+*cand*, *cand*+*suffix* という Web 検索クエリを作成する. (例: 呼称候補 “クールビューティー” と prefix 隣接パターン “トリノオリンピックで” から Web 検索クエリ “トリノオリンピックでクールビューティー” を作成する.) 検索クエリは、隣接パターンの数だけできることになる. 作成したクエリを Web 検索エンジンにかけ、検索エンジンの返す「推定検索結果数」に応じて *cand* のスコアを計算する. (ヒューリスティクス 2)

（4） スコアが閾値を越える候補を「呼称」として抽出する.

呼称抽出の流れを図 1 に示す.

3.3 検索ワードの選択と blog 記事収集

3.1 節で述べたように、フルネームおよび呼称をキーワードとして blog 検索エンジンを用いて記事収集を行う際に、検索ワードを選択し、クエリに加える必要がある。追加する検索ワードは、できるだけ目的人物と関連が深いものを選択する. また、収集する記事に偏りが生じないよう、検索ワードを変えて複数回検索を行うこととする. 本節では、検索ワードの選択方法について述べる.

目標としては、多くの blog 記事中で、時間的な要素に左右されず、フルネームおよび呼称とコンスタントに共起する語を

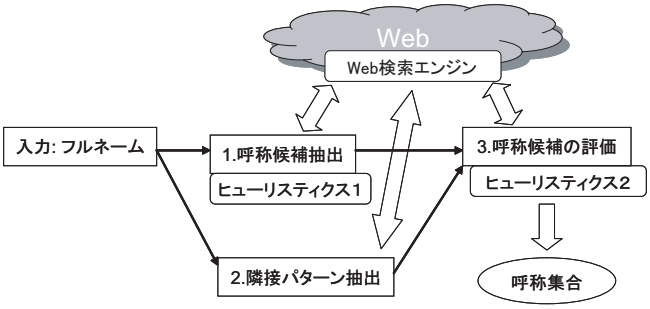

図 1　呼称抽出の流れ

検索ワードとして選択したい. 正確にそのような語を知ることは難しいが、Web 空間全体に目を向け、Web 検索エンジンを利用することで、ある程度適切な検索ワードが発見できる、と考えられる. なぜなら、リンクに基づくページのランクづけが行われている Web 検索エンジンの検索結果は、時間的な要素の影響が軽減されるからである. なお以下では、同姓同名を考慮せず、フルネームは目的人物を同定できる、と仮定する.

具体的には、まずフルネーム/呼称をクエリとして Web 検索を行い、フルネーム/呼称周辺に高頻度で出現する単語・熟語を検索ワード候補として抽出する. 次に検索ワード候補の中でフルネーム/呼称との共起度の高い語を検索ワードとして選択する. ただし、それのみで目的人物を同定する力の弱い呼称の場合は、検索ワード候補抽出および選択の際にフルネームを補助的に用いる. 以下の 2 節で、フルネームに対する検索ワード、呼称に対する検索ワードの選択方法についてそれぞれ述べる.

3.3.1 検索ワードの選択: フルネームの場合

まず、フルネームを含む Web ページ集合を、Web 検索エンジンを用いて取得する. 次に、フルネーム周辺に出現する単語・熟語の出現回数をカウントし、出現回数の多い語を検索ワード候補とする. 検索ワード候補中から、フルネームと共起度の高い語を選択する. 具体的には、フルネームを含む Web ページ集合と、検索ワード候補を含む Web ページ集合の重なりの度合を、その検索ワード候補のスコアとする (図 2). 式で書くと、次のように表せる.

$r1 = \#totalResults(fullname)$

$r2 = \#totalResults(term)$

$inter = \#totalResults(termANDfullname)$

$score(term) = inter/(r1 + r2 - inter)$

ここで、*fullname* はフルネームを、*term* は検索ワード候補を表す. また、*#totalResults(query)* は、*query* を含む全 *Web* ページ数を返す関数である(注3). 各候補のスコアを計算した後、スコアの高い上位 k 個を検索ワードとして選択する.

3.3.2 検索ワードの選択: 呼称の場合

3.1 節で指摘したように、呼称はそれのみでは目的人物を同定できないことが多いため、検索ワードの選択において、補助的にフルネームを用いる. おおまかな流れはフルネームの場合と同様である. まず、検索ワード候補を「呼称とフルネームの

(注1)：以下、対象人物のフルネーム文字列を *fullname* と表記する。

(注2)：演算子 + は、文字列の連結を表す.

(注3)：実際には、正確な Web ページ数を得ることは困難であるため、実験では Web 検索エンジンの返す推定検索結果数を用いた

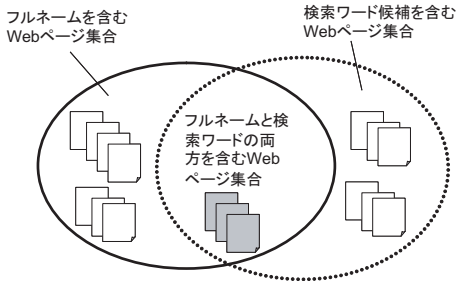

図 2　検索ワード候補のスコア

両方を含む」*Web* ページ群から抽出する．また、検索ワード候補のスコアは、「呼称とフルネームを含む」*Web* ページ集合と、検索ワード候補を含む *Web* ページ集合の重なりの度合、とする．式で書くと、以下のようになる．

$r1 = \#totalResults(mnameANDfullname)$

$r2 = \#totalResults(term)$

$inter = \#totalResults(termANDmnameANDfullname)$

$score(term) = inter/(r1 + r2 - inter)$

ここで、*mname* は呼称を表す．フルネームの場合と同様、スコアの高い上位 k 個を検索ワードとして選択する．なお、ここで得られる検索ワードは、フルネームと呼称の両方と共起しやすい語であり、両者をつなぐ語とみなすこともできる．

3.3.3　blog 記事収集

フルネーム/呼称と検索ワードをクエリとし、*blog* 検索エンジンを用いて記事収集を行う．すなわち、*“フルネーム (*呼称*) AND* 検索ワード *”* をクエリとして、k 回の検索を行うことになる．また、検索された記事のうち、対象期間外に書かれた記事は除去する．

3.4　時刻印を考慮したトピック抽出

次に、各フルネーム/呼称ごとに収集した記事群をクラスタリングし、トピックの抽出を行う．文書クラスタリングの手法を用いることで、出現単語の分布が似ている記事の集合を抽出することが可能である．しかし、現実のイベントに対応した「トピック」を抽出するためには、時間軸を意識することが必要となる．特に本研究のように、特定の人物に着目する場合は、どの記事同士の類似度も高くなってしまう可能性がある．たとえばあるスポーツ選手が参加した、まったく違う時期に行われた試合についての記事は、出現する単語という点ではよく似ているであろう．

我々は、*[17]* において、時刻印を考慮した文書クラスタリングを提案した．これは、凝集型の階層的クラスタリングを拡張したものである．凝集型の階層的クラスタリングは、一つの要素のみからなる初期クラスタ群から出発し、最も似ているクラスタペアの併合を繰り返すことで、クラスタを徐々に大きくしていく．*[17]* では、クラスタごとに「併合許容期間」を定めることで、凝集型クラスタリングに時間的な要素を導入している．「併合許容期間」とはクラスタ内で最も古い記事と最も新しい記事の時間差に応じて定められる期間である．クラスタ間類似度計算の後に、類似度の高いクラスタペアについて、両者の併合許容期間が重なっているかどうかをチェックしたうえで併合

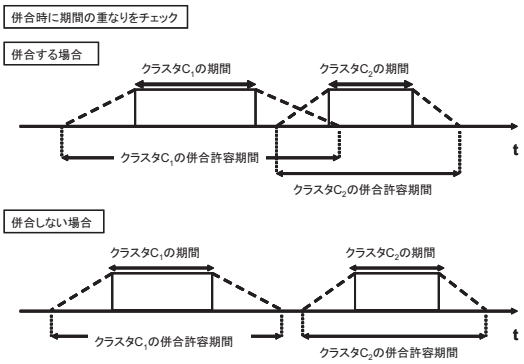

図 3　時間要素を考慮した階層的クラスタリング

するかどうかを決定する．図 *3* に、2つの類似クラスタを併合する場合としない場合の例を示す．

4.　実　　験

本節では、「小泉純一郎」と「荒川静香」の *2* 人について呼称を用いたトピック抽出を行った結果を示し、得られたトピックについて分析する．なお、*Web* 検索には *Yahoo!*検索 *Web* サービス(注4)*, blog* 検索には *Technorati API*(注5)を利用した．

4.1　呼 称 抽 出

呼称抽出フェーズでは、呼称候補抽出ステップ、隣接パターン抽出ステップでは、それぞれ *500* 件の *Web* ページを解析して候補およびパターンを抽出した．表 *1* に、呼称抽出フェーズで抽出された呼称を示す．不適切な呼称も少数含まれているが、おおむね妥当な呼称が抽出されていることがわかる．

4.2　トピック抽出:「小泉純一郎」

「小泉純一郎」については、表 *1* にみられるように多くの呼称が抽出された．本実験ではその中で、「小泉純一郎」*(*フルネーム*)* および「変人」、「ポチ」、「純ちゃん」の3つの呼称を用いてトピック抽出を行った．また、トピック抽出の対象期間は *2006* 年 *7* 月 *1* 日*-2006* 年 *12* 月 *31* 日とした．

まず、検索ワード選択で選択された語を表 *2* に示す．なお、ここでは検索ワード候補のうち、スコアの高かった上位 *30* 件を検索ワードとした．表より、呼称ごとに選択された検索ワードに違いがみられることがわかる．たとえば「変人」の場合、「田中真紀子」「派閥」といった語が上位にきており、「ポチ」の場合では「ブッシュ米大統領」「イラク」といった語がみられる．

次に、収集した記事群から抽出されたトピック *(*記事数の多い順に上位5件*)* を表 *3-6* に示す．ここで、表中の「ラベル」は、人手でつけたものである．なお、収集された記事数は各フルネーム/呼称ごとにまちまちであるが、トピック抽出の際には各フルネーム/呼称ごとにそれぞれランダムに *400* 件の記事を選択し、クラスタリングを行った．

表 *3-6* を比較すると、「靖国神社参拝」のようにフルネームおよび複数の呼称で抽出されているトピックもあるが、全体的にフルネームと各呼称ではトピックに違いがあることがわかる．

(注4)：http://developer.yahoo.co.jp/search/
(注5)：http://www.technorati.com/developers/api/

表 1　抽出された呼称

小泉純一郎	変人, ポチ, ライオン丸, 結城純一郎, 日本国民, 純ちゃん, 金日正, ジュンジュン, ライオン宰相
荒川静香	女王, しーちゃん, 静香様, クールビューティー, 開拓者

また、フルネームから得られたトピックの個々の記事をみると、新聞記事やニュースサイトからの引用が大部分を占めており、逆に呼称から抽出されたトピックの個々の記事は個人的なコメントや意見が書かれたものが多かった．したがって、フルネームと呼称は用いられる文脈に違いがあるといえよう．

次に、同じトピックでもフルネームと呼称が用いられる場合では、書き手の観点に違いがあるかを分析する．例として、フルネームおよび複数の呼称で観察されたトピック「靖国神社参拝」をとる．「靖国神社参拝」は、小泉純一郎前首相が *2006* 年 *8* 月 *15* 日に靖国神社を参拝したことに関し、賛否両論が激しく闘わされたトピックである．政治的な話題であり書き手の立場も非常に様々であるが、ここでは簡単のため、このトピックに関する全記事を今回の靖国神社参拝に「賛成」「反対」「中立 *(*新聞記事からの引用等含む*)*」に分け、それぞれの割合を比較した．以下、その結果を示す．なお、「靖国神社参拝」と関係のない記事 *(*ノイズ*)* は除いてある．

- 「小泉純一郎」
  - 賛成*: 13%,* 反対*: 14%,* 中立*: 73%*
- 「変人」
  - 賛成*: 25%,* 反対 *53%: ,* 中立*: 22%*
- 「純ちゃん」
  - 賛成*: 40%,* 反対*: 14%,* 中立*: 46%*

記事数がそれほど多くないため、この結果だけから結論づけることはできないが、書き手の立場とどの呼称が用いられるかとの間には、何らかの関連がみられるのではないか、と考察される．この点に関する詳細な検討は、今後の課題としたい．

4.3　トピック抽出:「荒川静香」

「荒川静香」については、「荒川静香」*(*フルネーム*)* および「しーちゃん」、「クールビューティー」の２つの呼称を用いてトピック抽出を行った．また、トピック抽出の対象期間は *2006* 年 *1* 月 *1* 日*-2006* 年 *12* 月 *31* 日とした．

まず、検索ワード選択で選択された語を表 *7* に示す．なお、「小泉純一郎」の場合と同様、検索ワード数は *30* 個とした．

次に、収集した記事群から抽出されたトピック *(*記事数の多い順に上位５件*)* を表 *8-10* に示す．なお、トピック抽出の際には各フルネーム*/*呼称ごとにそれぞれランダムに *400* 件の記事を選択し、クラスタリングを行った．

表 *9,10* の、各トピックの記事数の分布をみると、「しーちゃん」「クールビューティー」という呼称が使われているのは、トリノオリンピックの時期に集中していることがわかる．

次に、トピック「トリノオリンピック」について、フルネーム*/*各呼称を使っている *blog* 記事の書き手が、どのような観点からこのトピックに言及しているのか分析してみたい．「トリノオリンピック」自体は比較的大きなトピックなので、荒川静香選手がフィギュアスケート女子で金メダルを獲得した直後の *2/24-2/26* 日に書かれた記事に絞って分析する．この期間に書かれた記事は、*(*金メダルを獲得したという*)* 日記や記録に類する記事を除くと、「待ちに待ったトリノオリンピック初の」「アジア女性がフィギュア女子で獲得した初の」金メダルというように、特別な意味を持つ金メダルであることに着目した記事と、荒川選手の演技の「美しさ」「華麗さ」「完璧さ」など演技そのものに着目した記事の２パターンに大きく分けられることが観察された．そこで、フルネーム*/*各呼称を用いている書き手が、どちらの観点で記事を書いているのか調べ、割合を比較した．どちらにも言及している記事は、力点がどちらに強く置かれているか、で分類した．以下、結果を示す．

- 「荒川静香」*(24* 件 *2/24-26)*
  - メダル*: 67%,* 演技*: 4%,* その他 *29%*
- 「しーちゃん」*(80* 件 *2/24-26)*
  - メダル*: 18%,* 演技*: 34%,* その他 *48%*
- 「クールビューティー」*(129* 件 *2/24-26)*
  - メダル*: 15%,* 演技*: 43%,* その他 *42%*

特に、「クールビューティー」という呼称が使われている場合に、荒川選手の *(*フリーの*)* 演技について詳しく描写している記事が目立った．ただし、トピックの記事数がまちまちであるため、こうした傾向が本当にみられるのか、詳しく検討する必要があるだろう．

5.　まとめと今後の課題

本稿では、人物に関する呼称を用いて *blog* からのトピック抽出を行う手法を提案した．また、*blog* 検索エンジンを用いて実際に *blog* 記事を収集・トピック抽出実験を行い、どのようなトピックが抽出されるかを観察した．また、抽出されたトピックを分析し、フルネームと各呼称が用いられる文脈に違いがみられるか検討した．

今後の課題としては、目的人物が呼称で参照されている記事の収集精度の向上、フルネームと呼称が用いられる文脈の違いに関する詳細な検討、などがあげられる．また、今回はどのようなトピックがあるかを検出しただけであるが、*blog* という情報源の性質を考慮すると、各トピックが実際にどの程度話題になっているのか、話題の規模を推定する、といったテーマも興味深い．

謝辞　本研究の一部は、科学研究費補助金特定領域研究 *(#18049005)* の助成による。

文　献

[1] S.Tejada et al., Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification, In SIGKDD 2002.

[2] M.A.Hernandez et al., The merge/perge Problem for Large Databases, In SIGMOD 1995.

[3] S.Sarawagi et al., Interactive deduplication using active learning, In SIGKDD 2002.

[4] Russell Swan and James Allan. Extracting significant time varying features from text, CIKM'99

[5] R.Kumar et al., On the bursty evolution of Blogspace,

表 2　検索ワード: 小泉純一郎

フルネーム/呼称	検索ワード
小泉純一郎(フルネーム)	靖国神社参拝, 自民党総裁, 内閣総理大臣, 靖国神社, 靖国参拝, 衆院, 自民党, 総裁, 郵政民営, 公明党, 参拝, 首相官邸, 長官, 衆議院, 官邸, 安倍, 構造改革, ブッシュ, 演説, 憲法, 派閥, 選挙, 自衛隊, 議員, イラク, 国会, 秘書官, 批判, 石原, 就任
変人	自民党総裁, 田中真紀子, 内閣総理大臣, 郵政民営, 総裁, 靖国参拝, 衆院, 総理, 自民党, 派閥, 参院, 梶山静六, 安倍晋三, 構造改革, 森喜朗, 郵政, 政界, 政権, 小渕恵三, 総裁選挙, 内閣, 政治家, 民主党, 橋本龍太郎, 秘書官, 官邸, 選挙, 外相, 首相, 議員
ポチ	ブッシュ米大統領, 靖国神社参拝, ブッシュ大統領, 内閣総理大臣, 靖国参拝, 安倍晋三, 郵政民営, 衆院, 総裁, 自民党, ブッシュ, 参拝, 売国奴, 政権, 民主党, イラク, 憲法, 自衛隊, 日本国, 日米, 北朝鮮, 法案, 選挙, 大統領, 議員, マスコミ, 朝日新聞, 幹事, 批判, 戦争
純ちゃん	日本国首相, 靖国神社参拝, 自民党総裁, 内閣総理大臣, フジ特番, 靖国参拝, 郵政民営, 自民党, 靖国神社, 参拝, 政権, 秘書官, 総裁, 衆院, 遊説, 訪朝, 選挙, 北朝鮮, 会談, 議員, 国会, 大統領, ブッシュ, 麻生太郎, 安倍晋三, 外交, 発言, イラク, 国民, 記者

表 3　“小泉純一郎”(フルネーム) に関するトピック (上位 5 件)

ラベル	期間	記事数
靖国神社参拝	2006/08/09 - 2006/08/21	74
小泉内閣総辞職	2006/09/25 - 2006/09/30	7
対中韓関係	2006/09/25 - 2006/10/10	26
「政治家は使い捨て」発言	2006/11/07 - 2006/11/20	22
ブッシュ米大統領靖国批判	2006/12/14 - 2006/12/15	7

表 4　“変人” に関するトピック (上位 5 件)

ラベル	期間	記事数
靖國神社参拝	2006/08/15 - 2006/08/19	30
「小泉劇場」幕引き	2006/08/18 - 2006/08/30	10
自民党総裁選	2006/09/15 - 2006/09/29	48
田中真紀子氏のコメント	2006/10/06 - 2006/10/10	6
郵政増版組・復党問題	2006/11/04 - 2006/11/20	6

表 5　“ポチ” に関するトピック (上位 5 件)

ラベル	期間	記事数
「卒業旅行」訪米	2006/07/01 - 2006/07/12	23
自民党総裁選	2006/09/18 - 2006/09/28	11
北朝鮮核問題	2006/10/04 - 2006/10/24	12
変わり土鈴「ポチの家」	2006/11/05 - 2006/11/09	10
米中間選挙	2006/11/07 - 2006/11/13	13

表 6　“純ちゃん” に関するトピック (上位 5 件)

ラベル	期間	記事数
「卒業旅行」訪米	2006/07/01 - 2006/07/12	12
「ポスト純ちゃんまんじゅう」発売	2006/07/21 - 2006/07/28	12
靖国神社参拝	2006/08/14 - 2006/08/27	63
自民党総裁選	2006/09/16 - 2006/09/26	21
「晋ちゃんまんじゅう」発売	2006/10/05 - 2006/10/14	9

表 7　検索ワード: 荒川静香

フルネーム/呼称	検索ワード
荒川静香 (フルネーム)	フィギュアスケート女子 女子フィギュアスケート 安藤美姫 フィギュアスケート 金メダル 浅田真央 金メダル獲得 トリノオリンピック トリノ五輪 村主章枝 トリノ トリノ五 輪女子フィギュアスケート イナバウアー 世界選手権 女子シングル 高橋大輔 金メダリ スト メダリスト オリンピックメダリスト エキシビション トリノ冬季五輪 女子フィギ ュアスケートメダリスト フィギュアスケーター アイスショー トゥーランドット 本田武史 競技 紅白 演技 出場
しーちゃん	安藤美姫選手 あれだけ完成 女子シングル 金メダル獲得 トリノオリンピック スルツカ ヤ トリノ五輪 金メダリスト 世界選手権 メダリスト ファンサイト オリンピック プリ ンスホテル レミオロメン オフィシャルサイト アマチュア 村主章枝 金メダル 安藤美姫 トラックバック 恩田美栄 高橋大輔 演技 プロ転向 ミキティ 浅田真央 競技 滑り ジャンプ NHK
クールビューティー	フィギュアスケート女子 女子フィギュア 女優デビュー 五輪フィギュアスケートメダリ スト トリノ五輪フィギュアスケート 女子フィギュアスケート 金メダリスト トリノオリンピック フィギュアスケート プロスケーター スルツカヤ トリノ五輪 トリノ冬季五輪 メダリスト トゥーランドット 村主章枝 金メダル 金芽 トリノ イナバウアー 女弁護士 スケーター ミキティ トップアスリート テレビ朝日 安藤美姫 銀盤 氷上 活躍 ポーカーフェイス

表 8　“荒川静香”(フルネーム) に関するトピック (上位 5 件)

ラベル	期間	記事数
トリノオリンピック	2006/02/22 - 2006/03/10	55
プロ転向発表	2006/05/07 - 2006/05/17	12
グランプリファイナル, メダリスト・オン・アイス(注6)	2006/12/03 - 2006/12/31	40
2006 年のビッグニュース	2006/12/18 - 2006/12/31	52
紅白歌合戦ゲスト	2006/12/21 - 2006/12/27	12

(注): このトピックは、フィギュアスケートグランプリファイナルとメダリスト・オン・アイスの 2 つのサブトピックからなる. 本来は、別々のトピックとして検出されるべきものが混合されている.

表 9　“しーちゃん” に関するトピック (上位 4 件)(注7)

ラベル	期間	記事数
トリノオリンピック, 凱旋公演	2006/02/19 - 2006/03/12	179
「スマスマ」(TV 番組) ゲスト出演	2006/03/13 - 2006/03/29	10
世界選手権出場辞退	2006/03/22 - 2006/04/02	9
プロ転向発表	2006/05/04 - 2006/05/14	18

(注): “しーちゃん” に関しては、人手でラベルが付けられるトピックが 4 件のみだったため、4 件のみ示した.

表 10　“クールビューティー” に関するトピック (上位 5 件)

ラベル	期間	記事数
トリノオリンピック	2006/02/20 - 2006/03/11	218
女優デビュー	2006/05/25 - 2006/06/02	41
週刊誌グラビア登場	2006/05/29 - 2006/05/30	12
ファッションショーモデル出演	2006/07/18 - 2006/07/20	6
「雪見だいふく」TVCM 出演	2006/08/17 - 2006/08/18	6

WWW2003

[6] D.Gruhl et al., Information Diffusion Through Blogspace, WWW2004

[7] E.Aar et al., Implicit Structure and the Dynamics of Blogspace, Workshop on the Weblogging Ecosystems, WWW2004

[8] Gabriel Pui Cheong Fung et al., Parameter Free Bursty Events Detection in Text Streams, VLDB2005

[9] Yiming Yang et al., Learning Approaches for Detecting and Tracking News Events, IEEE Inteligent Systems 1999(Vol.14, No.4) pp.32-43

[10] Jon Kleinberg. Bursty and hierarchical structure in streams, SIGKDD2002

[11] 白砂健一, 小山聡, 田島敬史, 田中克己, *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, 第 *17* 回データ工学ワークショップ, *2006.*

[12] 中島伸介, 竹原幹人, 舘村純一, 日野洋一郎, 原良憲, 田中克己, *blog* 解析に基づく *Web* 情報検索の信頼性向上技術, 人工知能学会研究会資料 *SIG-SWO-A401-05*

[13] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学, *document stream* における *burst* の発見, 言語処理学会 第 *11* 回年次大会 *2005*

[14] 藤村滋, 豊田正史, 喜連川優, 電子掲示板からの評価表現および評判情報の抽出, 人工知能学会第 *18* 回全国大会, *2004.*

[15] 鈴木康裕, 高村大也, 奥村学, *Semi-Supervised* な学習手法による評価表現分類, 言語処理学会第 *11* 回年次大会, *2005.*

[16] 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 *2006(DBWS2006)*

[17] 外間智子, 北川博之, *blog* における人物に関する”旬な”話題の抽出電子情報通信学会 第 *17* 回データ工学ワークショップ *(DEWS2006)*