2006－DBS－140（II）（49）
2006／7／13

Webコーパスを用いた人物の呼称抽出

外間　智子†　　北川　博之†,††

† 筑波大学大学院 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@kde.cs.tsukuba.ac.jp, kitagawa@cs.tsukuba.ac.jp

あらまし　Web 掲示板や Weblog といったツールが普及するにつれ、Web は、世の中の関心を反映する新しいメディアとしても注目されるようになってきた。Web からの情報抽出・知識抽出の代表的なものに、評判情報の抽出がある。評判情報抽出のようにある特定のオブジェクト（組織、製品、人物など）に着目する場合、まずそのオブジェクトがどのように参照されているかという情報を基に、オブジェクトに関する Web ページを収集する必要がある。ここで問題となるのが、一般的な Web 文書では、あるオブジェクトは公的・正式な名称だけでなく様々な呼び名で参照される、という点である。本研究ではオブジェクトの正式名以外の「参照のされ方」を「呼称」と呼ぶ。例えば、人物であれば姓、名、所属と肩書の組合せ、ニックネームなどが考えられるだろう。本研究は、オブジェクトのうち「人物」に着目し、人物の呼称を抽出することを目的とする。本論文では、人物のフルネームが出現するパターンを手がかりに Web コーパスより人物の呼称を抽出する手法を提案し、実際の人物を対象に行った評価実験について述べる。

キーワード　Web マイニング, 知識抽出, オブジェクト識別

Mnemonic Name Extraction about a Person from Web Corpus

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 Science, University of Tsukuba, Tennodai 1–1–1 Tsukuba-shi, 305-8577 Japan
E-mail: tomokoh@kde.cs.tsukuba.ac.jp, kitagawa@cs.tsukuba.ac.jp

Abstract　The web has gained much attention as new media recently due to the proliferation of tools such as bulletin boards and weblogs. Reputation information extraction is one of the major research topics in information extraction and knowledge extraction from the web. Collecting web pages about a target object is needed for reputation information extraction. A big problem for collecting web pages is that objects are refered to by various ways in general web documents. For example, a person may be refered to by the full name, the first name, affiliation and title, or nicknames. In this paper, we propose a method for extracting these mnemonic names of a person from the Web coupus and show experimental results for some people.

Key words　Web Mining, Knowledge Extraction, Object Identification

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。その主なものとして、Web 上の掲示板や Weblog 等があげられる。様々な情報発信サービスが普及するにつれ、Web は、世の中の関心を反映する新しいメディアとしても注目されるようになってきた。

Web から有用な情報を抽出したいという需要は高く、これまで様々な研究がなされてきたが、代表的なもののひとつに評判情報の抽出がある [5], [6]。評判情報抽出では、ある特定のオブジェクト（組織、製品、人物など）に関して人々がどのような評価をしているか、という情報 (評価表現) をオブジェクトを指す文字列 (製品名など) の周辺テキストから抽出する。そのため、あるオブジェクトに着目する際には、まずそのオブジェクトがどのように参照されているのかという情報を基に、オブジェクトについての記述を集める必要がある。既存の評判情報抽出では、製品名などオブジェクトの正式名称周辺のテキストから、そのオブジェクトに関する評価表現を抽出している。

評判情報のような局所的な情報の収集だけでなく、特定のオブジェクトに特化した話題の抽出も今後重要になってくると考えられる。例えば人物を考えると、情報源として人物情報データベース、Wikipedia、公式 HP などがあるが、これらには基本的に静的・公式な情報が掲載されている。一方、その人物に関する最近の話題、というような情報は動的で非公式であるため、これらの情報源からは得られない。動的・非公式な情報を得るためには、blog 記事などを含めた全 Web スペースを考慮する必要がある。また非公式な情報であることから、「どのような話題があるか」に加え「その話題がどの程度関心を集めているか」も重要であろう。筆者らは [9] で blog を対象として人物に関するトピック抽出を行い、その規模 (記事数) を推測する手法を提案した。

このように特定のオブジェクトに関する情報収集や知識抽出を行う場合に問題となるのが、一般的な Web 文書を対象にした場合、1 つのオブジェクトが様々な方法で参照される、という点である。例えば人物の場合、ある人物の参照のされ方はフルネーム、姓、名、所属と肩書の組合せ、ニックネームなど様々である。本論文では以下、こうしたオブジェクトの正式名以外の「参照のされ方」を「呼称」と呼ぶ。特に、クレームや悪い評価などを記述する際には、フルネームではなく別の呼び名が用いられることも多い。

オブジェクトには組織などいくつかあるが、本研究では、「人物」に着目する。人物に関する呼称のうち、姓などは自明であるが、それ以外の呼称を収集することは一般に容易ではない。本研究の目的は、Web コーパスより人物の呼称を抽出することである。呼称の抽出には、人物のフルネームが出現するパターンを手がかりとして用いる。

本論文の構成は次のようになっている。第 2. 節で関連研究として Duplicate detection とオブジェクト識別について触れ、第 3. 節で提案手法の概要と詳細について述べる。次に第 4. 節で実際に何人かの人物を対象として行った実験の結果について述べ、第 5. 節で結論と今後の課題をまとめる。

2. 関連研究

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 [3], [4]。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。

[2] は、オブジェクト識別を Personal Information Management(PIM) に適用している。[2] は個人の PC 上のファイルを解析し、論文・人物・会議など複数オブジェクトの依存関係を利用してオブジェクト識別を試みる。基本的なアルゴリズムは、あらかじめ人手で作成したスキーマをもとにオブジェクト間の依存グラフを構築し、次にノード間類似度を計算・伝播させることで、同一オブジェクトを指すノードを発見する、というものである。

その他、日本語プレーンテキスト (Web 文書) を対象としたものとして、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手法も提案されている [7], [8]。ただし、本研究のように、テキストからある人物に関する別の呼び

名を抽出する研究はこれまでなされていない。

3. 人物の呼称抽出

3.1 基本アイディアと提案手法の概要

本節では、Web コーパスより、人物の呼称を抽出するための提案手法について述べる。提案手法は、以下の 2 つのヒューリスティクスを背景とする。

- 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)
- 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*, *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。

これらを踏まえた提案手法の大まかな流れは次のようになる。

（1） Web コーパスより、対象人物の呼称候補を抽出する。

（2） Web コーパスより、フルネームと隣接する文字列 (prefix および suffix パターン) を抽出する。さらにフルネームとパターンの関連の強さを基に重みづけを行い、重要なパターンを「隣接パターン」として選択する。

（3） 2 で抽出した隣接パターンと Web コーパスを用いて 1 で抽出した呼称候補を評価する。

（4） 3 で計算した評価値が閾値を超える候補を、対象人物の呼称として出力する。

提案手法の流れを、図 1 に示す。

以下では、呼称候補抽出、隣接パターン抽出、呼称候補の評価の詳細について述べる。

3.2 呼称候補抽出

日本語では、ある人物の別名や愛称を記述する際に、よく" ○○こと *fullname*" という表現が用いられる。したがって、「こと *fullname*」という文字列の直前に出現する文字列は、その人物の呼称である可能性が高い。また、それが一般的に使われる呼称であれば、Web コーパス上に何度も出現するであろう。以上を踏まえ、次のように呼称候補を抽出する。

（1）「こと *fullname*」という文字列をクエリとして Web 検索を行い、URL リスト (N 件) を取得する。

（2） URL リストの Web ページを取得・解析し、「こと *fullname*」という文字列の直前に出現する文字列<$t_1t_2...t_n$>を得る。(t_1, t_2, ..., t_n は形態素)

（3） <$t_1t_2...t_n$>の部分文字列 $t_1t_2..t_n$, $t_2t_3..t_n$, ..., $t_{n-1}t_n$, tn のうち、最初の単語が名詞であるものを呼称候補とし、取得した Web ページ群中での出現回数をカウントする。

（4） 抽出した呼称候補のうち、出現回数が 1 回のみのものを除去する。

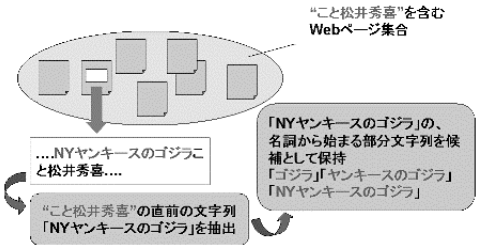

図 2 呼称候補抽出の例

呼称候補抽出の例を、図 2 に示す。図 2 は、メジャーリーグプレーヤーの「松井秀喜」の呼称候補を抽出する例である。

3.3 隣接パターン抽出と重みづけ

隣接パターン (フルネームと隣接する文字列) は、呼称候補と同様、Web コーパスから抽出する。基本的には人物のフルネームをクエリとして Web 検索を行い、フルネームが出現する Web ページを取得・解析してパターンを得る。ただし、同姓同名の人物が存在する可能性があるため、対象人物と関連の深いオブジェクト (所属組織等) 名を補足的にクエリに加える。

隣接パターンが得られたら、次に各パターンを、対象人物のフルネームとどの程度関連が深いか、という観点から重みづけする。重みづけには再び Web コーパスを用いる。

ある prefix 隣接パターン *prefix* の重みは以下のように計算する。*prefix* を含む Web ページ集合を R、*prefix*+*fullname* を含む Web ページ集合を $R1$ とする。理論的には $R1$ は R のサブセットであり、R1 が R に占める割合が高くなるほど *prefix* と *fullname*

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

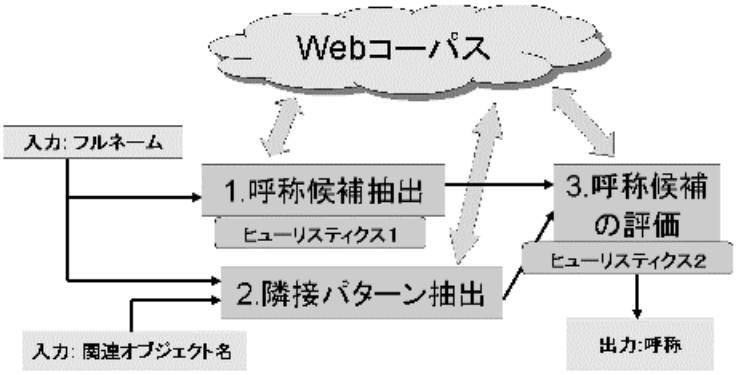

図 1　提案手法の概要

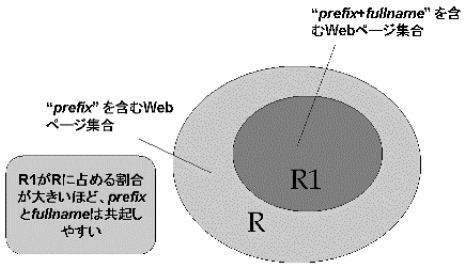

図 3　*fullname* と *prefix* の関連の強さ

が共起しやすい、すなわち関連が深いと考えられる (図 3)。以上を踏まえ、*prefix*+*fullname* を含む Web ページ数 (*R*1 の要素数) を *prefix* を含む Web ページ数 (*R* の要素数) で割った値を、*prefix* の重みとして採用する。

隣接パターン抽出と重みづけの手順を以下に示す。

（1） 対象人物と関連の深い（同姓同名の人物が存在した場合に、対象人物を一意に特定できる）オブジェクト名を *rel_obj* とする。*fullname* AND *rel_obj* をクエリとして Web 検索を行い、URL リスト (N 件) を取得する。

（2） URL リストの Web ページを取得・解析し、*fullname* に隣接する文字列 *prefix* および *suffix* を抽出する (図 4)。ここで、各 *prefix*, *suffix* の単語数は m 単語に固定する。

（3） 抽出したすべての prefix について、次のように重み w を計算する。*prefix* をある prefix とするとき、

$r = searchResults(prefix)$

$r1 = searchResults(prefix + fullname)$

$w(prefix) = r1/r$

（4） 抽出したすべての suffix について、次のように重み w を計算する。*suffix* をある suffix とするとき、

$r = searchResults(suffix)$

$r1 = searchResults(fullname + suffix)$

$w(suffix) = r1/r$

ここで、searchResults(query) は、query を含む全 Web ページ数を返す関数である。(注2)

（5） 重みづけをした prefix、suffix の中から、それぞれ重みの大きい上位 k 件を隣接パターンとして選択する。

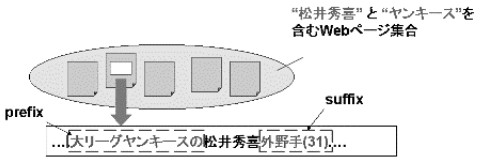

図 4　prefix, suffix パターン抽出

3.4　呼称候補の評価

次に、抽出した呼称候補を、隣接パターンと Web コーパスを用いて評価し、候補の絞り込みを行う。評価には、3.1 節で述べた 2 つ目のヒューリスティクスを用いる。すなわち、ある呼称候補 *cand* が、隣接パターンの直前 (後) に Web コーパス中に出現すれば、*cand* は実際に対象人物の呼称である可能性が高い、

(注2)：実際には、正確に R, $R1$ を得ることは不可能である。実験では推定値として Yahoo!API の totalResultsAvailable を利用した。

とみなす。具体的な評価手順は以下に示す。

（1） *cand* のスコアを 0 に初期化する。

（2） すべての隣接パターンについて、以下を繰り返す。

（a） パターンが prefix 隣接パターン *prefix* の場合、文字列 *prefix*+*cand* を生成する。パターンが suffix 隣接パターン *suffix* の場合、文字列 *cand*+*suffix* を生成する。

（b） 生成した文字列をクエリとして Web 検索を行い、トータル検索結果数を得る。得られた検索結果数に、パターンの重み (w(*prefix*) もしくは w(*suffix*)) を掛け合わせた値を *cand* のスコアに足し込む。これは、隣接パターンの直前 (後) に出現する回数が多く、またその隣接パターンの重みが大きいほど、*cand* が実際に呼称である可能性が高い、と考えられるためである。

（3） スコアが閾値以上であれば、*cand* を対象人物の呼称として返す。

呼称候補評価手順の疑似コードを以下に示す。

```
score(<cand>)=0
for <prefix> in allPrefixPattern
  query="<prefix><cand>"
  score(<cand>)+=searchTotalResults(query)
                 * w(<prefix>)
end
for <suffix> in allSuffixPattern
  query="<cand><suffix>"
  score(<cand>)+=searchTotalResults(query)
                 * w(<suffix>)
end
```

4. 実　　験

提案手法の有効性を検討するため、実際の人物を対象に実験を行った。

4.1 パラメータ設定

- 呼称候補抽出
  - 解析する Web ページ数: 500(件)
  - 抽出する文字列の単語数: 5
- 隣接パターン抽出
  - 解析する Web ページ数: 500(件)
  - 隣接パターンの単語数: 3
  - 隣接パターン数: (prefix, suffix それぞれ) 20 個
- 呼称候補の評価
  - 呼称として抽出する候補のスコア閾値: 0.01

4.2 実験結果および考察

8 人の人物について、提案手法を用いて呼称の抽出を行った。それぞれの人物について、抽出された呼称候補、隣接パターン、候補の評価後に呼称として抽出された候補を表 1–8 示す。

表より、各対象人物の呼称として、おおむね妥当な文字列が抽出されていることがわかる。各対象人物の呼称候補中には、呼称とはいえない文字列も多く含まれているため、隣接パターンを用いた呼称候補の評価が候補の絞り込みに効果的に働いたと考えられる。ただし、妥当な呼称とはいえない文字列 (表 3 の” 事件ホリエモン” など) を抽出していたり、抽出すべきと考えられる呼称候補 (表 5 の” 今信長” など) を見逃している場合もみられる。また抽出された隣接パターンの重み、呼称のスコアにかなりのばらつきがみられ、適切な閾値が対象によって変わってしまうため、重み/スコア算出方法の改善、自動的な閾値設定を検討する必要がある。

今回は「こと」という単語を手がかりとして、名詞から始まる文字列をのみ呼称候補としたが、形容詞から始まる文字列も考慮する、「こと」以外の言語的ヒューリスティクスを利用する等、呼称候補の広げ方も検討していきたい。

5. 結論と今後の課題

Web コーパスを用いて、対象人物の (フルネーム以外の) 呼称を抽出する手法を提案し、予備実験を通してその有効性と問題点を検討した。今後の課題としては、呼称候補抽出および隣接パターン抽出の改善、手法の評価方法の検討、有名人以外の人物への拡張、人物以外のオブジェクト (組織など) への拡張が挙げられる。

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。

表 1　ターゲット: 松井秀喜 (関連オブジェクト名: ヤンキース)

呼称候補	ゴジラ, ヤンキースのゴジラ, 契約したゴジラ, 松井, 今年ついにゴジラ, 野球界のゴジラ, ゴジラ松井, ビッグ松井,09 にしこり, のはゴジラ,...(全 41 個)
prefix 隣接パターン (重み)	北陸中日新聞挑戦 (0.99), 正式辞退する (0.94), 大リーグヤンキースの (0.72),...
suffix 隣接パターン (重み)	早期回復祈願 (0.96), の大リーグ日記 (0.94),55 の伝説 (0.94),...
抽出された呼称 (スコア)	松井 (25270.23), ゴジラ (26.47), ゴジラ松井 (16.22)

表 2　ターゲット: 安藤美姫 (関連オブジェクト名: トリノ)

呼称候補	ミキティ, ミキティー, 位ミキティー, アイドルミキティ, スケートのミキティ, フィギュアのミキティ, 氷上のミキティ, 中のミキティ,...(全 23 個)
prefix 隣接パターン (重み)	ミキティーこと (0.84), 全日本女王の (0.56), 名言のプレゼント (0.51),...
suffix 隣接パターン (重み)	宮里藍浅尾 (0.98), が結婚トリノ (0.97), 中野荒川の (0.97),...
抽出された呼称 (スコア)	ミキティ(6.74), ミキティー (1.84), アンミキ (0.04)

表 3　ターゲット: 堀江貴文 (関連オブジェクト名: ライブドア)

呼称候補	リエモン, ホリエモン, もん, ほりえもん, 事件ホリエモン, りえもん, 社長ホリエモン, ホリえもん, 夜ホリエモン, 社長ほりえもん,...(全 15 個)
prefix 隣接パターン (重み)	として同社社長 (0.78), 日同社社長 (0.58), 術神原弥奈子 (0.50),...
suffix 隣接パターン (重み)	社長逮捕新 (0.99), 社長のオンザエッジ (0.98), のカンタン儲かる (0.98),...
抽出された呼称 (スコア)	ホリエモン (1236.65), ほりえもん (17.71), 社長ホリエモン (3.60), ホリえもん (2.62), 社長ほりえもん (0.45), 堀えもん (0.35), 事件ホリエモン (0.10)

表 4　ターゲット: 坂本龍一 (関連オブジェクト名: YMO)

呼称候補	教授, 世界のサカモト, アホマン 2 号, 世界の教授, ゲストは教授, キョージュ,YMO の教授,04 教授, 音楽家教授, 大好きな教授,...(全 38 個)
prefix 隣接パターン (重み)	アイテム細野晴臣 (0.75), 歌で復帰 (0.69),YMO 細野晴臣 (0.4),...
suffix 隣接パターン (重み)	総合情報局 (0.99), ライブジャパンツアー (0.97), からリスナーの (0.85),...
抽出された呼称 (スコア)	教授 (46.23), キョージュ(0.49), 龍一教授 (0.29), 我らが教授 (0.12),YMO の教授 (0.01)

表 5　ターゲット: 小泉純一郎 (関連オブジェクト名: 自民)

呼称候補	純ちゃん, ライオンハート, アイドルジュン様, 政界の変人, ポチ, 今信長, そったれバッハ (全 7 個)
prefix 隣接パターン (重み)	内閣総理大臣 (0.06), で講演し (0.007), 訪問中の (0.004),...
suffix 隣接パターン (重み)	の率いる自民 (0.5), 首相が今年 (0.42), 首相の靖国神社 (0.35),...
抽出された呼称 (スコア)	純ちゃん (0.02)

表 6　ターゲット: 小澤征爾 (関連オブジェクト名: ウィーン)

呼称候補	オザワ, 世界のオザワ (全 2 個)
prefix 隣接パターン (重み)	ニューイヤーコンサート 2002(0.02), ニューイヤーコンサート２００２ (0.02), 管弦楽団指揮 (0.003),...
suffix 隣接パターン (重み)	さん 7 月復帰 (0.97), セレクション音楽の (0.97), 音楽塾オペラ (0.93),...
抽出された呼称 (スコア)	オザワ (1.77)

表 7　ターゲット: 西村博之 (関連オブジェクト名: 2 ちゃんねる)

呼称候補	ひろゆき, 開祖であるわたくし, 日記の開祖, 社長日記の開祖, 人ひろゆき, ひろゆき氏, のひろゆき (全 7 個)
prefix 隣接パターン (重み)	東京プラス社長 (0.93), ひろゆきこと (0.85), あるわたくしこと (0.67),...
suffix 隣接パターン (重み)	から田代正 (0.6),Seesaa 社長へ (0.33), 氏は 24(0.31),...
抽出された呼称 (スコア)	ひろゆき (42.06), のひろゆき (20.20), ひろゆき氏 (3.84), 人ひろゆき (0.77)

表 8　ターゲット: ジーコ (関連オブジェクト名: 日本代表)

呼称候補	ペレ, 神様, サッカーの神様 (全 3 個)
prefix 隣接パターン (重み)	日本協会打診 (0.90), 豪州と初戦 (0.78), 痛すぎる敗戦 (0.23),...
suffix 隣接パターン (重み)	監督日本外国特派員協会で (0.99), 監督肩落とす (0.98), スタイル進化する (0.96),...
抽出された呼称 (スコア)	ペレ (11.46), 神様 (0.13), サッカーの神様 (0.04)

文　　献

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

[2] Xin Dong et al., Reference Reconciliation in Complex Information Spaces, In SIGMOD 2005.

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

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

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

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

[7] 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.

[8] 木村塁, 戸田浩之, 田中克己, 検索結果スニペットのクラスタリングによる同姓同名人物の特定, 第 17 回データ工学ワークショップ, 2006.

[9] 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, 第 17 回データ工学ワークショップ, 2006.