

ニュースサイトによる人物や組織に関する記述の特徴分析

石田 晋[†] 馬 強^{††} 吉川 正俊^{††}

[†] 京都大学工学部情報学科 〒 606-8501 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 京都市左京区吉田本町

E-mail: [†]ishida@db.soc.i.kyoto-u.ac.jp, ^{††}{qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし 我々は、ニュース理解を支援するため、Web 上の様々なニュースサイトの記事集合における、ある特定の人物や組織とよく同時記述される人物、組織および事象などを、その人物や組織に関する記述の特徴として抽出し提示する手法を提案する。記事文の主語、目的語、述語といった単語の役割を考慮して、人物や組織に関する記述を求め、それを用いて人物や組織の記述によく利用される記述表現を抽出する。同時に、その他のサイトにおけるその人物や組織に対する記述と比較して、そのサイトに特有の記述を抽出する。さらに、我々は、抽出された特徴的記述を時系列的に提示する手法を提案して、サイトごとの記述特徴とその時間的な変化の可視化を試みる。国内の主要新聞社を対象とした実験では、特徴的記述の抽出が確認できた。

キーワード ニュースサイト、記述特徴、人物、組織、テキストマイニング、情報抽出

Analyzing Features of Different News Agencies' Descriptions of People and Organizations

Shin ISHIDA[†], Qiang MA^{††}, and Masatoshi YOSHIKAWA^{††}

[†] Undergraduate School of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 Japan

^{††} Graduate School of Informatics, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8591 Japan

E-mail: [†]ishida@db.soc.i.kyoto-u.ac.jp, ^{††}{qiang,yoshikawa}@i.kyoto-u.ac.jp

Abstract News agencies report news from different viewpoints and with different writing styles. In this paper, we propose a method to extract the characteristic descriptions, especially about people and organizations, written by news agencies. To extract the characteristic descriptions of a given person or organization by a news agency, we analyze co-occurrence relationships among words which appear in the same sentence based on their SVO roles. Then, we extract a description which is often used by the news agency but not commonly used by the others. We carried out an experiment by using the articles from three major Japanese newspaper agencies. The experimental results show that our method can elucidate the different features of each agency.

Key words news agency, characteristic description, person, organization, text mining, information extraction

1. はじめに

現在 Google ニュース [1] や Yahoo! ニュース [2], あらたに [3] のような複数のニュースサイトから記事を集めてまとめて表示するようなサービスや、ニュースサイトの記事を並べて比較して見せるようなサービスは数多く存在する。しかし、ニュースサイトごとに定常的に持っている独特の観点や取り上げる事象といった特徴を分析し、違いを明らかにして見せるようなサービスがまだ少ない。例えば、他のニュースサイトに比べて

自民党に対して否定的な記述が多い、中国に関する記事では毒餃子事件と一緒に取り上げがちであるといった特徴である。こういった特徴を意識せずに記事を読むと、無意識に一方的な観点到るなどの弊害につながる可能性がある。よって、ニュースサイトの特徴を明瞭にし、ユーザーがニュースサイトごとの特徴を意識できるようなサービスが必要であると考えられる。

特定の人物や組織に関する記述には各サイトの観点や意見の違いが出ることが多々ある。例えば、Z という政治家に対して P 新聞のニュースサイトは肯定的、Q 新聞のニュースサイトは

否定的な記述表現が多く見られた。我々は各ニュースサイトの記事集合からこのような表現を分析し、人物や組織に対する特徴的記述を抽出する手法を提案する。また、我々はある人物や組織と同時によく取り上げる事象や人物・組織の違いから各サイトの特徴的記述を抽出する。さらに、我々は、特徴的記述を時系列的に提示する手法を提案し、サイトごとの人物や組織に対する特徴的記述の遷移を明らかにする。これらの特徴的記述を用いて、ニュースサイトによる人物や組織に関する記述の特徴を分析できる。ニュースサイトごとの特徴を分析することで、そのニュースサイトらしい記事(そのサイトで、よくある書き方や報道の仕方の記事)、またはそのニュースサイトにしてはめずらしい記事(そのサイトにおける、今までの書き方や報道の仕方とは大きく異なる記事)を評価できる。

具体的に、特定の人物や組織に対する特徴的記述を抽出するため、我々は、その人物や組織と同文内に現れる表現について分析する。その際我々は特に主語、目的語、述語といった文内での単語の役割に注目する。特定の人物・組織が文内で主語や目的語の役割で登場すれば、残りの要素(目的語、述語または主語、述語)はその人物・組織に関する記述と考えることができるので、そのような記述を抽出し、それを用いて、特徴的記述を以下の2つの基準に基づいて抽出する。

- サイト内の記事集合内で計算された特徴量: サイト内での記述の多さを示し、そのサイトでよく使われる記述を抽出する指標である。

- その他のサイトの記事集合と比較して計算された特徴量: 他のサイトと比べての記述の珍しさを示し、他サイトではあまり使われていない記述を抽出する指標である。

つまり、他のサイトにあまり現れない頻出記述をそのサイトの特徴的記述として抽出する。

さらに、我々は、人物や組織に対する特徴的記述を時系列的に比較して見せる提示方法を提案する。我々はトピック(人物や組織が関わる事象)間で人物や組織に対する特徴的記述が変化し得ることに注目し、トピック単位での特徴的記述を抽出する。特徴的記述をサイトごとに時間軸に沿って並べることで、人物や組織に関する記述のトピック単位での特徴とニュースサイトごとの差異を視覚的に見ることができる。

以下、本論文の構成を説明する。まず2章で関連研究を紹介する。次に3章で特徴的記述の抽出手法とその提示方法について述べる。4章でプロトタイプシステムと実験について述べる。5章でまとめを述べる。

2. 関連研究

2.1 ニュースの差異に関する研究

我々は、ニュースサイトごとの記述の特徴を分析することで、ニュースサイトごとの差異を発見することを目的としている。

ニュースサイトの記事の差異を抽出する研究は数多く行われている。このような既存研究は同一トピックに対して差異のある記事を見せるものや各サイトごとの意見のパラツキを見せるものが多い。

瀬本ら[4]は、類似WebサイトのWebページを同時に比

較表示するブラウザである Comparative Web Browser(CWB)を提案している。異なるニュースサイト間で同一トピックに関する類似記事をキーワードにより発見し、記事を比較して見せている。馬ら[5][6]は関連ニュース報道から各記事で取り上げられている話題と視点を、記事で出現する語の頻度や品詞などに基づいて抽出することにより、同じトピックについてのニュースの視点の多様性を提示している。濱砂ら[7]は同一トピックに関する関連ニュースの記事から特徴語を求めた後に、語のセンチメント値を計算している。センチメント値とは語の印象を表す数値で、この値により記事の特徴を求め、各ニュースサイトの観点の違いを見せている。青木ら[8]は同一トピックに関する関連ニュース記事を比較して、キーワードにより記事内で言及している事象と言及していない事象を抽出している。これにより各記事で重視している部分と軽視している部分といった意図が分かり、記事ごとの意見の違いを見せている。

これらの研究はいずれも記事ごとの差異の抽出を目的としている。しかし、抽出できる差異は各サイトの同一トピックに関する記事間での差異であって、あくまでも一時的な差異である。これまで一時的な差異ではなく、あるニュースサイトがよく見せる記述やよく取り上げる事象に注目して、ニュースサイトごとに常々持っているような差異を見せるような研究はほとんど見当たらない。

そこで我々は定常的な差異が出ることが多い人物や組織に関する記述に注目し、ニュースサイトの人物や組織に関する記述についてそのサイトの記事集合内での比較と他サイトの記事集合との比較を行うことで、そのサイトでよく見られる特徴的記述を抽出し、ニュースサイトごとの差異を示す。

2.2 人物記述の抽出に関する研究

人物に関するキーワードを抽出する研究を挙げる。森ら[9]は人物に関する情報を表すキーワードを、Web上から語の共起情報を用いて抽出する手法を提案している。外間ら[10]はBlogから特定の人物に関する記事を収集し、クラスタリングを行うことで特定の人物に関する話題を抽出している。飯田[11]らは従属節、主節、埋め込み文などといった文・文章中の構造を考慮して、人物などのゼロ代名詞に対する先行詞候補の位置の分類を行っている。その結果人物が省略されている文でも、ある人物について書かれた文と同定することができる。

これらの研究では人物に関するキーワードを人物が登場する記事に含まれる語句として抽出しているが、そのキーワードが人物に対して記述されたものとは限らない。そこで我々は、主語、目的語、述語といった単語の役割に注目することで、人物に対しての記述を抽出し、さらにその記述をニュースサイトごとに収集することでニュースサイトの人物に関する記述の特徴を示す。

2.3 発信者の特徴分析

他に情報の発信者特徴を分析している研究として、Blog発信者に関する研究も挙げておく。関口ら[12]はBlog発信者をBlog記事内で使う語句から同じ興味を持つ者同士にクラスタリングし、各クラスターで話題となっている語を抽出している。中島ら[13]はBlog発信者の記事で使われているキーワードが

ら、そのキーワードが表す話題に対する発信者のマニア度を算出し、それに基づき Blog エントリのランキングを算出する方法を提案している。

いずれも発信者による特徴の分析といった面で本論文と関連している。しかし、これらの研究で抽出する特徴は、発信者の興味や嗜好といったもので、発信者ごとの観点の差には注目していない。本論文では、他のニュースサイトとの比較も取り入れることで、ニュースサイトの観点を記述の特徴から抽出する。

3. 人物や組織に対する特徴的記述の抽出と提示

前述したように、ニュースサイトにおける人物や組織に対する特徴的記述は、他のサイトではあまり用いられてない、そのサイトにおいてよく使われる記述表現である。このような記述表現を抽出するため、まず、我々は、それぞれのサイトにおけるその人物や組織に関連する記事集合をトピック（人物、組織が関わる事象）ごとに分類する。そして、トピックごとに、その人物、組織の記述によく使われる表現を、主語や目的語や述語といった単語の役割を考慮して抽出する。さらに他のサイトと比較して、自サイトの独自表現を抽出する。

3.1 記述構造

文内の、主語（～が）、目的語（～を、～に）、述語（～する）から成る構造を文の記述構造とする。人物や組織に関する記述は、人物・組織が主語、目的語として登場する記述構造を元に取り出す。例えば、ある特定の人物・組織 X を仮定すると、 X が主語となる「 X が～」が現れる記述構造や、目的語となる「 X を～」 「 X に～」が現れる記述構造である。人物・組織が主語である場合と目的語である場合に依じて、記述構造から以下の 6 パターンの記述を得る。ここでは主語は S 、目的語は O 、述語は V と記している。

(1) 人物・組織が S の場合

- $S-V$ 記述（人物・組織が V した）
- $S-O$ 記述（人物・組織が O を）
- $S-O-V$ 記述（人物・組織が O を V した）

(2) 人物・組織が O の場合

- $O-S$ 記述（人物・組織を S が）
- $O-V$ 記述（人物・組織を V した）
- $O-S-V$ 記述（人物・組織を S が V した）

以上 6 パターンの記述を人物・組織記述と呼び、人物や組織に関する記述とする。各パターンごとに次のような意味を持つ。

- $S-O$ 、 $O-S$ 記述

人物・組織と同時記述される人物・組織や事象を示す

- $S-V$ 、 $O-V$ 記述

人物・組織とその行動を示す

- $S-O-V$ 、 $O-S-V$ 記述

人物・組織と同時記述される人物・組織とその行動を示す

3.2 トピックに基づくニュース記事分類

人物や組織に関する記述を抽出するため、人物や組織に関する記事を取得する。人物や組織に関する記事は記事内の 1 文に人物・組織が現れる記事とする。我々は、人物や組織に対す

表 1 人物・組織 X が現れる記述構造

X が主格	X が目的格、与格
X が ～を ～する	～が X を (に) ～する
(S) (O) (V)	(S) (O) (V)

る特徴的記述を、トピック単位で抽出する。トピックは特定の人物・組織が関わる事象や事件とする。例えば、 A 大統領という人物の記事を分類して、就任というトピックと辞任というトピックで特徴的記述を抽出すれば、 A に関する記述の就任時と辞任時の変化を見ることができる。そのため、キーワードと時間を用いて、人物や組織に関する記事を分類する。本研究では、1 週間などの固定時間長でニュース記事を分類しているが、石川ら [14] の手法を用いることが考えられる。

3.3 特徴的記述の抽出

3.3.1 自サイトにおける特徴抽出

人物や組織に対する特徴的記述を抽出するために、人物や組織が現れる記事に対して、記事の各文ごとに構造解析を行い、文から記述構造を取得する。この内、人物や組織に関する記述となる人物や組織が登場する記述構造を取得する。人物・組織は名詞であるため文中で S または O としてのみ現れる。よって、人物・組織が S の場合と人物・組織が O の場合で、記述構造を元に 3.1 節で述べた 6 パターンの人物・組織記述を取得する。パターンによって違う意味を持っているので、パターンごとによく出現する人物・組織記述を抽出する。ニュースサイト N_j のトピック t での人物・組織記述 $d_i(p, x)$ の出現率 r_x を下式で求める。

$$r_x(N_j, t, d_i(p, x)) = \frac{cooc_x(N_j, t, d_i(p, x))}{\sum_{i=0}^{n_x} cooc_x(N_j, t, d_i(p, x))} \quad (1)$$

但し p は人物・組織名、 x は記述パターンを表す。 n_x はパターン x での記述の総数である。 $cooc_x(N_j, t, d_i(p, x))$ はパターン x での人物・組織記述 $d_i(p, x)$ の出現数を表す。

人物・組織記述の出現率 r_x が求めれば、 r_x が高い記述が自サイトの記事集合における特徴的な記述となる可能性が高い。

3.3.2 他サイトとの比較による特徴抽出

他サイトの記事との比較による自サイトの記事の独特の記述を求める。特定の人物・組織を含む記事集合を全サイトから獲得し、各サイトごとに前項で述べた手法を用いて人物・組織記述を抽出する。他サイトとの比較により、自サイトの人物・組織記述の特徴を、自サイトの記述の出現率を他サイトでのその記述の出現率の合計値で割った値 isf (inverse site frequency) で計算する。ニュースサイト N_j のトピック t での人物・組織記述 $d_i(p, x)$ の isf を下式で求める。

$$isf(N_j, t, d_i(p, x)) = \frac{r_x(N_j, t, d_i(p, x))}{\sum_{j=0}^m r_x(N_j, t, d_i(p, x))} \quad (2)$$

但し、 m は比較するニュースサイト数を表す。

3.3.3 特徴的記述の抽出

ニュースサイト N_j でのトピック t における人物・組織記述

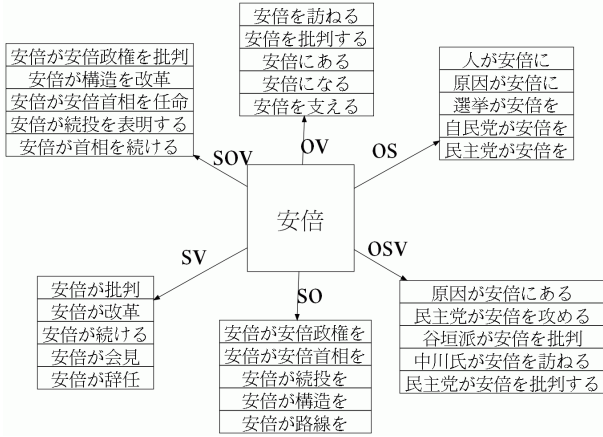


図 1 トピックごとの特徴的記述例

$d_i(p, x)$ の特徴量 f は, r_x と isf を用いて下式で求める.

$$f(N_j, t, d_i(p, x)) = r_x(N_j, t, d_i(p, x)) \times isf(N_j, t, d_i(p, x)) \quad (3)$$

人物・組織 p に対する全ての語の人物・組織記述の特徴量を求め, パターンごとに特徴量の値によりランク付を行う. 各パターンで上位にある記述を, そのサイトの人物・組織 p に対する特徴的記述とする.

3.4 提示方法

我々は, サイトごとの特徴的記述を時系列的に表示する手法を提案して, サイト間の差と時間的な変化を明らかにする. 提案提示方法の概観を図 3 に示す. トピックごとの特徴的記述を, 縦軸がサイト, 横軸が時間の空間に配置する. これで, 縦軸の比較によりサイト間の記述の差異を, 横軸の比較によりあるサイトの記述の時間的推移を見ることができる. 実際に人物・組織に対する特徴が時間的に変化する様子を観察できた例を 4.2 節に記す.

トピックごとの特徴的記述は図 1 のようなグラフで表現する. 中央のノードは人物または組織名を, 周辺のノードは 6 パターンごとの特徴的記述を, それぞれ表す. 各パターンの記述はそれぞれ 3.1 節に記したような特徴を持つので, 各パターンごとに 5 記述ずつ計 30 記述をグラフにより示す. 図 1 の例では 2007 年 7 月の参院選時の P 新聞社の安倍元首相に対する, 特徴量ランク 5 位内の特徴的記述を示している.

4. プロトタイプシステムと実験

3 章の手法を用いて, プロトタイプシステムを構築し実験を行った.

4.1 プロトタイプシステム

図 2 にプロトタイプシステム構成図を示す. プロトタイプシステムでは, 大手の新聞社 3 社 (P 社, Q 社, R 社) の 2007 年度記事約 57 万件のデータを利用して記事データベースを構築している. 記事データベースでは (新聞社名, 日付, ジャンル, タイトル, 本文) といったスキーマを用いて, 各記事を格納している. また, 転置ファイルを作成することで, 任意の日付の任意の人物や組織が含まれる記事を検索可能にしている.

システムは入力として特定の人物または組織名を受け取り,

記事データベースからその人物または組織名が現れる記事を検索する. その後, 記事を 1 週間や 1 か月といった時間単位に分類する.

次に各トピックの記事ごとに, 文を日本語の述語項解析器 syncha[15] を用いて解析する. syncha は文章を入力すると, 文の述語項構造解析により, 述語と項の関係 (ガ格, ヲ格, ニ格) を出力する. システムでは syncha の解析結果から, ~ がのガ格の語を S, ~ を, ~ にのヲ, ニ格の語を O, 述語を V として記述構造を取得する. その内, 人物・組織が現れる人物・組織記述を抽出し, 記述データベースに格納する. トピック単位で 3.3 節の手法を用いて記述の特徴量を計算し, これもまた記述データベースに格納する. つまり記述データベースには (新聞社名, トピック, 記述, 特徴量) といったデータが格納される.

そして各新聞社ごとに特徴量により記述をランキングし, 各パターンでランキング 5 位以内の特徴的記述を図 1 のようなグラフとして作成する. システムの出力のキャプチャ画像を図 3 に示す.

4.2 実験

様々な人物や組織に対して, 及び様々なトピック分類法を用いて実験を行った.

4.2.1 実験 1: 週単位の特徴的記述抽出

実験 1 では 2007 年 1 月から 3 月までの 3 ヶ月間の記事を, 週ごとに分類して記述を分析した.

対象人物・組織は下の通りである.

- (1) 安倍晋三^(注1)
- (2) 自民党
- (3) 小沢一郎^(注2)
- (4) 北朝鮮

まず, 各新聞社で独自の記述が抽出できているかの指標として, 人物・組織 p のパターン x における特徴的記述の重複数を, 以下のような被覆率を用いて評価する. 被覆率が低ければ, 各新聞社の p の x における記述の重複している数が少なく, 各社ごとに独自の記述が抽出できていることになる. 表 2 に各人物・組織における記述のパターンごとの被覆率を示す.

$$cov(p, x) = \frac{1}{n} \times \sum_{k=1}^n \frac{|c_k(p, x)|}{|D_k(p, x)|} \quad (4)$$

n は記事を分割したクラスタ数 (この場合は週数), $D_k(p, x)$ は p の x におけるクラスタ k での全社の特徴的記述集合, $c_k(p, x)$ は p の x におけるクラスタ k 内の特徴的記述で各社で重複している記述の集合をそれぞれ表す.

表 2 に示す通り, 被覆率は小沢の OV パターンで 0.102 となった以外, 全て 0.1 以下となった. これは 10 件に 1 件ほどの割合で各社の記述が重複していることを示し, 各社で独特の記述を抽出できたと思われる.

次に, 抽出した特徴的記述が, 実際にどの程度各社の特徴を表しているかを調べるため, 抽出精度を計算した. 抽出精度は

(注1): システムの入力としては「安倍」を入力した

(注2): システムの入力としては「小沢」を入力した

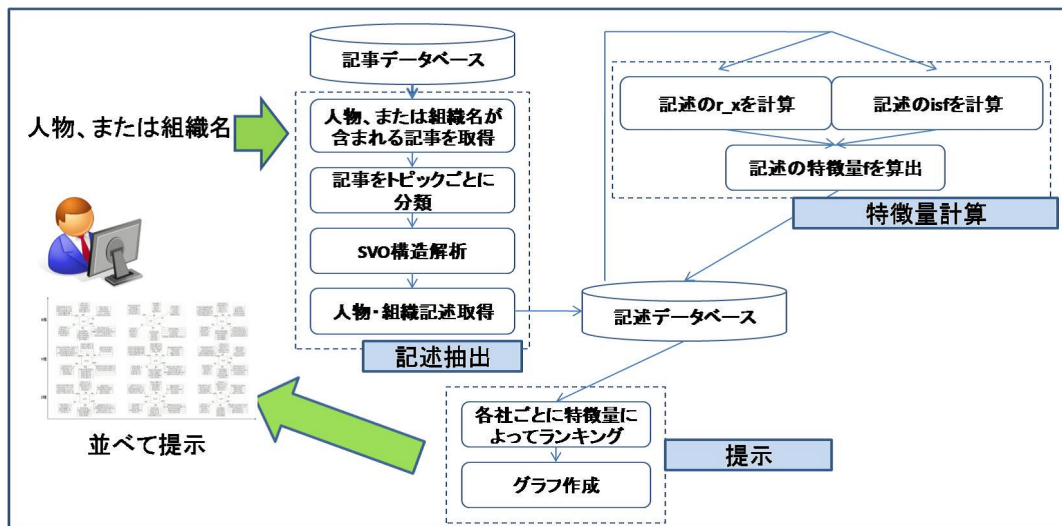


図 2 プロトタイプシステム構成図

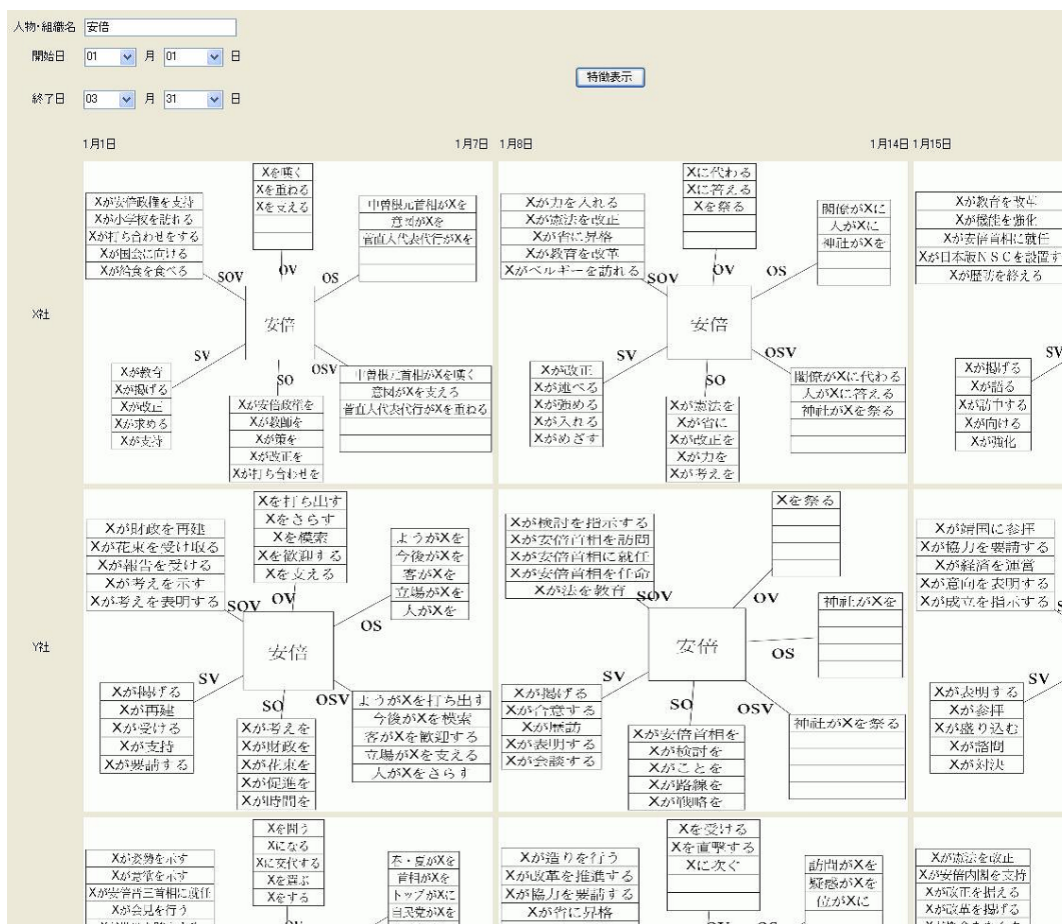


図 3 プロトタイプシステムの出力キャプチャ画像

システムが出力した特徴的記述の内、筆者が各社の特徴を表していると判断した記述数の割合である。ニュースサイト N でのトピック t (この場合はトピックは各週) の人物 p のパターン x における記述の抽出精度を、以下の式で計算した。

$$precision(N, t, p, x) = \frac{|D_s(N, t, p, x) \cap D_u(N, t, p, x)|}{|D_s(N, t, p, x)|} \quad (5)$$

$D_s(N, t, p, x)$ は N の t での p の x におけるシステムが出力した特徴的記述集合、 $D_u(N, t, p, x)$ は筆者が各社の特徴を表

していると判断した記述集合を表す。

表 3 に週ごとに計算した抽出精度の平均値を示す。各パターンの平均抽出精度は OSV パターンが 0.5475 と最も高く、次いで OS パターンが 0.475 と続く。一方 SV パターンや OV パターンはそれぞれ 0.255, 0.23 と低い値となった。これは現在の各パターンでの特徴の現れやすさに差があることを示している。今後、これについて検討し、特徴量計算手法の改善を行う予定である。

表 2 週単位で抽出した特徴的記述の被覆率

	SV	SO	OV	OS	SOV	OSV
安倍	0.08	0.05	0.06	0.04	0.04	0.03
自民党	0.1	0.05	0	0.01	0.03	0
小沢	0.02	0.01	0.06	0.03	0.02	0
北朝鮮	0.08	0.11	0.03	0	0.03	0

表 3 各人物・組織に関する特徴的記述の抽出精度

	SV	SO	OV	OS	SOV	OSV
安倍	0.27	0.16	0.28	0.46	0.33	0.47
自民党	0.26	0.28	0.3	0.58	0.29	0.72
小沢	0.21	0.23	0.11	0.39	0.22	0.42
北朝鮮	0.28	0.45	0.23	0.47	0.41	0.58
平均	0.255	0.28	0.23	0.475	0.3125	0.5475

表 4 安倍に関する記述

安倍の教育に関する記述数			
	1 月 1 週-1 月 4 週	2 月 1 週-2 月 4 週	3 月 1 週-3 月 5 週
P 社	9	1	0
Q 社	1	1	6
R 社	1	2	1
安倍に対する否定的記述数			
	1 月 1 週-1 月 4 週	2 月 1 週-2 月 4 週	3 月 1 週-3 月 5 週
P 社	2	7	6
Q 社	0	2	0
R 社	1	0	2
安倍と共に記述されている人物、組織、国数			
	1 月 1 週-1 月 4 週	2 月 1 週-2 月 4 週	3 月 1 週-3 月 5 週
P 社	9	15	25
Q 社	5	12	0
R 社	12	16	4

最後に、抽出した各新聞社の記述が、それぞれ特徴的であるかどうかの分析を行った。

まず、安倍に対して抽出できた記述を分析する。「安倍が教育を」や「安倍が教師を」などの教育に関する記述を分析すると、表 4 に示す通り、P 新聞社の 2007 年 1 月の記述では教育に関する記述が 9 件と他より多いが、徐々に減っていき、3 月の記述では今度は Q 社が 6 件と他より多くなっている。また、R 社は概して教育に関する記述は少ない。安倍と教育というペアに注目している時期や度合が新聞社によって異なることが分かる。

また、「安倍を批判する」や「安倍を非難する」など安倍に対する否定的な記述を分析すると、表 4 に示す通り、P 社は 15 件と他より圧倒的に多い。「批判」という記述の出現数だけを見ても、Q 社、R 社は合計 2 件であったのに対し、P 社は 8 回と明らかに多い。これは P 社の安倍に対する特徴と考えられる。

最後に安倍と共に記述されている人物名、組織名、国名数を分析した。人物名は個人名、組織名は自民党などの政党や団体、国名は中国などの国名をカウント（ただし、「安倍が安倍首相を」などの記述は SVO 解析の誤検出とみなしてカウントしていない）したところ、表 4 に示す通り、P 社が合計 49 件と総じて記述件数が多く、逆に Q 社は合計 17 件と総じて少ないことが分

表 5 自民党と共に記述されている人物・組織数

人物	1 月 1 週-1 月 4 週	2 月 1 週-2 月 4 週	3 月 1 週-3 月 5 週
P 社	19	21	30
Q 社	18	15	16
R 社	11	14	12
組織	1 月 1 週-1 月 4 週	2 月 1 週-2 月 4 週	3 月 1 週-3 月 5 週
P 社	2	0	2
Q 社	6	11	7
R 社	2	0	3

かった。また、記述されている人物・組織名にも特徴がみられ、P 社は小沢や福島などの野党の党首が、Q 社は有名でない個人名が、R 社は中川幹事長がそれぞれ独自に現われている。このように取り上げる人物・組織にも各新聞社で特徴が見られた。

以上のように安倍に対する各社ごとの特徴的な記述をみることであった。他の人物・組織に対する記述に対しても分析を行う。

- 自民党

自民党と共に記述される組織数に違いが見られた。P 社が合計 4 件、R 社が 5 件に対し、Q 社は 24 件と他社より圧倒的に多く、中でも「公明党が自民党を支援する」といった公明党を取り上げる記述がほとんどであり、公明党を取り上げる記述の 80 % が Q 社によるものであった。

- 小沢

P 社に批判的な記述がある程度見られるが、各新聞社に一貫した特徴は見られなかった。

- 北朝鮮

全体的に北朝鮮に批判的な記述が見られる。人物に関する記述を調べると、Q 社は 3 月に 14 件と他社より多く記述している。北朝鮮と共に様々な国の人物を記述しており、これは一つの傾向と考えられる。また、国に関する記述は、全社とも、米国や中国などの記述が総じて多かったが、その中でも P 社は 32 件と他社より多く見られた。

以上の結果より、安倍や自民党、北朝鮮に対しては各社の記述に特徴が見られることが分かる。しかし、小沢に対しては特徴を見出すことは出来なかった。

4.2.2 実験 2: トピック単位での特徴的記述抽出

実験 2 では対象人物・組織に関連するトピックを選定し、トピックごとの記述を分析した。

対象人物として安倍晋三を選び、トピックとして下の 5 事象を選定した。各事象の期間は事象の起きた日から 1 週間としている。

- (1) 松岡農水相問題
- (2) 久間防衛相失言問題
- (3) 参院選
- (4) 安倍首相海外訪問
- (5) 安倍首相辞任

システムの入力として安倍を入力し、トピックごとに特徴的記述を抽出した。

表 6 トピックごとに抽出した特徴的記述の被覆率

	SV	SO	OV	OS	SOV	OSV
安倍	0.246	0.15	0.1	0	0.132	0

表 7 特徴的記述のトピックごとの抽出精度

	SV	SO	OV	OS	SOV	OSV
松岡農水相問題	0.27	0.4	0.36	0.62	0.53	0.62
久間防衛相失言問題	0.27	0.4	0.33	0.4	0.4	0.6
参院選	0.33	0.2	0.47	0.53	0.53	0.93
安倍首相海外訪問	0.33	0.13	0.38	0.23	0.53	0.4
安倍首相辞任	0.07	0.2	0.38	0.23	0.53	0.4
平均	0.25	0.27	0.37	0.49	0.48	0.64

トピックごとの特徴的記述の被覆率と抽出精度を表 6,7 に示す。週ごとの記述に比べて、トピックごとの記述のほうが被覆率が全体的に高い傾向にある。これは例えば安倍首相辞任時にどの新聞社でも「安倍が辞任」という記述が多いように、同トピックについての各社の記述は似通うことが多いためだと考えられる。また抽出精度のパターンごとの違いは、週ごとの記述でもトピックごとの記述でも同様の違いがみられた。これにより OSV や OS パターンに特徴が現れやすいという傾向が明確になった。

表 8～12 にトピックごとの違いが見られる代表的記述を示す。

トピックごとに、抽出できた記述の新聞社ごとの違いを分析する。

● 松岡農水相問題

P 社の特徴的記述には「安倍が松岡氏をかばう」と安部の松岡氏への擁護している面がみられる一方、「安倍を批判」など否定的な面もみられる。

Q 社は「高村氏が安倍を支援する」と他にはない肯定的記述が見られる。

R 社は「安倍が唇をかみしめる」と安部の感情的な部分を記述している。

● 久間防衛相失言問題

P 社は「民意が安倍を離れる」と失言による国民の安倍離れを指摘している面が目立つ。

Q 社は「加藤が安倍を批判する」と否定的な面をみせている。

R 社は「選挙が安倍を問う」と選挙に関する記述が目立つ。

● 参院選

P 社は批判という記述が 30 件中 5 件と非常に多い。また、「原因が安倍にある」という記述もあり、総じて安倍に対して否定的な記述が多かった。

Q 社は「民主党が安倍を追い詰める」や「民主党が安倍を倒す」など民主党寄りの記述がある。その中でも「人が安倍を支持する」と他とは違い支持という記述がある。

R 社は「小池百合子が安倍に」や「公明党が安倍に」など独特の人物・組織と安倍を取り上げている。また「選挙が安倍を信任」と選挙に関する記述がまたも現われている。

● 安倍首相海外訪問

P 社はここでも批判という記述が 27 件中 4 件と他より目立っている。

表 8 松岡農水相問題における代表的記述

P 社	Q 社	R 社
安倍を批判	高村氏が安倍を支援する	安倍が唇を入りする
安倍が松岡氏をかばう	安倍が憲法を改正	安倍が密葬に参列する
責任が安倍にある		安倍が唇をかみしめる
マガジンが安倍を批判		安倍が参院選を
		安倍が教育を

表 9 久間防衛相失言問題における代表的記述

P 社	Q 社	R 社
民意が安倍を離れる	安倍が批判	選挙が安倍を問う
安倍が支持を訴える	加藤が安倍を批判する	野党が安倍を罷免
	安倍を批判する	安倍が要因を抱えこむ
		安倍が足を引っ張る
		田上富久市長が安倍に抗議する

表 10 参院選における代表的記述

P 社	Q 社	R 社
安倍が批判	民主党が安倍を追い詰める	選挙が安倍を信任
安倍を批判する	民主が安倍を倒す	有権者が安倍に突きつける
原因が安倍にある	人が安倍を支持する	国民が安倍を見る
民主党が安倍を攻める	安倍が赤城農相を更迭する	公明党が安倍に
谷垣派が安倍を批判		小池百合子が安倍に
民主党が安倍を批判する		安倍が反省する

表 11 安倍首相海外訪問における代表的記述

P 社	Q 社	R 社
太田代表が安倍を批判する	安倍が腰を立てる	同派が安倍を支持する
国民が安倍に憤る	安倍が態度を表明	問題が安倍を直撃する
安倍が態度を保留する	安倍が参拝する	安倍を支持する
安倍を批判する		氏が安倍を支える
安倍が批判		安倍が支持

表 12 安倍首相辞任における代表的記述

P 社	Q 社	R 社
竹中平蔵総務相が安倍を批判する	安倍が退陣を表明	北朝鮮が安倍をののしる
舛添厚労相が安倍を批判する	号外が安倍を	岸信夫が安倍を訪ねる
英国各紙が安倍を報道する	読売新聞が安倍を	小坂憲次が安倍を訪れる
安倍が憲法を改正		安倍が慶応大学病院に入る
安倍を否定		安倍を支持する

Q 社は「安倍が参拝する」と他とは違った面をとらえている。P 社の「安倍が態度を保留する」とは対照的な「安倍が態度を表明」という記述が見られる。

R 社は「安倍を支持する」など安倍に肯定的な記述が見られる。

● 安倍首相辞任

P 社の記述には批判が 3 件、否定が 1 件と否定的な記述が多い。また「英国各紙が安倍を報道する」と安倍と英国を同時に取り上げている。

Q 社には「号外が安倍を」や「安倍が退陣」など表現が他社よりも独特である。

R 社は「北朝鮮が安倍をののしる」と安倍と北朝鮮を同時に取り上げている。

以上のようにトピックごとに各社の記述の違いを見ることができた。

4.3 考 察

実験 1 のように週単位といった細かい単位で記述を抽出すると、一貫した記述の特徴が抽出しやすく、特徴の小さな変化も観察できた。各新聞社が同じトピックを 1 週以上にわたって書く傾向にあることが原因だと思われる。しかし、小沢のように一貫した特徴が出ないものもあった。これは小沢の記事を週単位で分割すると記事数が少ないため、抽出できる記述も少なすぎて、特徴が抽出しにくくなるのが理由と考えられる。

実験2のように選定したトピック単位で記述を抽出すると、トピックと人物の関連が抽出しやすく、各社の意見の違いが抽出しやすかった。

全体での問題点を挙げる。

- 「安倍が安倍首相を任命」や「家が安倍を改革」など文章として意味をなさない記述が多い。これは記事から記述構造を取得する精度に原因があり、記事文解析の手法の改良が必要である。特に「安倍首相の支持率が低下している」という文に対して「安倍首相が安倍首相を支持」と解析するなど明らかな解析ミスが見受けられるので改善する必要がある。

- 現在の特徴量計算法では、「責任が安倍に」、「安倍にある」、「責任が安倍にある」など各パターンで同じような記述が上位にきてしまうので、記述パターン間の包含関係などを考慮した計算法の改良も必要である。

5. ま と め

本論文ではニュースサイトの特定の人物または組織に対して記述を取り出すのに、記事文の主語、目的語、述語といった単語の役割を考慮した記述構造を用いた。自サイト内の記事集合から見た特徴と、他サイトの記事集合と比較した特徴から計算される特徴量を用いて特徴的記述を抽出する手法を提案した。また、これらの記述の時系列的特徴を考慮した提示方法を提案した。本手法を用いた実験の結果から、各サイトごとの記述の特徴を示すことができた。

提案手法の応用として、あるニュースサイトの記事に対して、記事内の人物、組織に対する記述を抽出し、そのサイトの人物、組織に対する特徴的記述と比較して、記事のらしさ（そのサイトの記事で、よくある書き方や報道の仕方であること）、めずらしさ（そのサイトにおける、今までの書き方や報道の仕方とは大きく異なること）を評価し、数値で付与することによるニュース検索のランキング手法も考えられる。

今後の課題として、より多くのニュースサイトや、より多くの人物、組織に対して実験を行い、提案手法によって特徴的記述を抽出できるか調査する予定である。また、評価実験を行い、提案手法がユーザーにとって有用であるか評価を行う必要がある。

謝 辞

本研究の一部は、科研費（20700084 と 20300042）の助成を受けたものです。また、本研究の一部は、NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです。

また、本研究では CD-読売新聞 2007 記事データ集、CD-朝日新聞 2007 記事データ集、及び CD-毎日新聞データ集 2007 年版を利用しました。データの利用を許可して下さった読売新聞殿、朝日新聞殿、毎日新聞殿に心より感謝致します。

文 献

- [1] Google ニュース : <http://news.google.co.jp/>
- [2] Yahoo ニュース : <http://headlines.yahoo.co.jp/>
- [3] くらべる一面 : 新 s あらたにす (日経, 朝日, 読売) : <http://allatany.jp/>

- [4] Akiyo Nadamoto, Katsumi Tanaka : A comparative web browser (CWB) for browsing and comparing web pages, Proc. of the 12th international conference on World Wide Web, pp.727-735, 2003
- [5] 馬強, 吉川正俊 : 話題と視点に基づくニュースコンテンツの多様性分析システム, IEICE SIG W12 第 13 回研究会論文集, pp.19-24, 2008
- [6] Qiang Ma, Masatoshi Yoshikawa : Topic and Viewpoint Extraction for Diversity and Bias Analysis of News Contents, Proc. of APWebWAIM2009 (to appear), 2009
- [7] 濱砂佳貴, 河合由起子, 熊本忠彦, 田中克己 : センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案, DEWS2008 論文集, B6-4, 2008
- [8] 青木伸也, 湯本高行, 角谷和俊, 新居学, 高橋豊 : 関連ニュース記事集合内の特異箇所注目した発信者意図の抽出, 情報処理学会研究会報告, 2008-DBS-146, pp. 187-192, 2008
- [9] 森純一郎, 松尾豊, 石塚満 : Web からの人物に関するキーワード抽出, 人工知能学会論文誌, Vol. 20, No. 5, pp. 337-345, 2005
- [10] 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, DEWS2006 論文集 1B-i4, 2006
- [11] Iida, R. Inui, K. and Matsumoto, Y : Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution, Proc. of COLING/ACL, pp. 625-632
- [12] 関口裕一郎, 川島晴美, 奥田英範, 奥雅博 : ブログ発信者の特徴を利用した話題抽出手法, 日本データベース学会 Letters, Vol. 5, No. 1, pp. 9-12, 2006
- [13] 中島伸介, 稲垣陽一, 草野奉章 : 高信頼性情報の提示を目指した熟知度に基づくプログランピング方式の提案, 日本データベース学会論文誌, Vol. 7, No. 1, pp. 257-262, 2008
- [14] 石川佳治, 北川博之 : 忘却の概念に基づくクラスタリング手法の改良方式, DBSJ Letters, Vol.2, No.3, pp.53-56, 2003
- [15] syncha : <http://cl.naist.jp/ryu-i/syncha/>