# 記述の主観性を考慮したニュース発信者の特徴分析とその応用

石田　晋† 馬　強† 吉川　正俊†

† 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，ニュースの偏り分析のために，ニュース発信者の過去の記事集合から特定のエンティティに対する記述を集めて，そのエンティティに対する発信者ごとの特徴ベクトルを生成し，分析する手法を提案する．まずニュース発信者の主観が現れている記述と現れていない記述を分けるために，特定のエンティティと共に使われている語を利用して記事文から主観的記述と客観的記述を抽出する．そして主観的記述が特定のエンティティに対して肯定的であるか，否定的であるか，客観的記述が他のニュース発信者に比べてどれほど独特であるかといった尺度により，発信者の特徴ベクトルを生成する．そして，さらに，特徴ベクトルを用いた各ニュース発信者の特徴が顕著に表れているトピックの推薦手法について議論する．

**キーワード**　テキストマイニング，信頼性，ニュース分析，発信者，主観性

# Analyzing Features of News Agencies by Considering Subjectivity of Descriptions

Shin ISHIDA†, Qiang MA†, and Masatoshi YOSHIKAWA†

† Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606–8501 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 discover features of a news agency for a certain entity(a person, organization, and country) by analyzing descriptions written about the entity. First, we classified descriptions into subjective description and objective description by analyzing the co-occurring words with the entity. Then, we generate a feature vector by measuring whether the subjective description is positive or negative and how the objective description is distinct. We discuss about an application to recommend a topic where the feature of news sender is expressed prominently.

**Key words**　Text Mining, News agency, Feature analysis, Subjectivity

## 1. はじめに

現在，Web 上では様々なニュース発信者から数多くのニュースが発信されている．Google ニュース [1] やあらたにす [2], FairSpin [3] のような複数のニュース発信者の記事を集めてまとめて見せるサービスや，ニュース発信者ごとの記事を並べて比較するサービスは数多く存在する．しかしこれらのサービスではニュース発信者が持っている特徴に注目して，発信者が定常的に持っている独特の観点や取り上げる事象を分析し，違いを明らかにするようなことは行っていない．例えば，他の発信者に比べてある政党に対して否定的な記述が多い，ある国に関する記事ではある人物を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情報を受け取ると，無意識に一方的な観点に陥るなどの弊害につながる可能性がある．よって，ニュース発信者によって生じるバイアスを明瞭にし，ユーザーが発信者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．

ニュース発信者の特徴に関する既存研究 [4]〜[6] は存在する．しかし既存研究では，ニュース発信者ごとの差異を定量的に分析し評価を行っていない．そこで，我々はニュース発信者の特徴の差異を定量的に分析するために，ニュース発信者の特徴ベクトルを生成する．

特定のエンティティに関する記述には発信者の観点や意見の違いが出ること多々ある．これは特定の政治家や政党，国に関する記述に関しては顕著であり，そういった記述は発信者の特徴を表しやすいものと考えられる．我々は各発信者の記事集合からこのような記述を分析し，発信者の特定エンティティに対する記述特徴を発見する手法を提案する．

ニュース発信者のエンティティに関する記述は次の 2 通りに分けられると考えられる．

- エンティティに対する主観的記述

エンティティに対する直接評価

- エンティティに対する客観的記述

エンティティに関する事象，事実の記述

例えば「安部首相が悪い」といった記述は発信者の安部首相に対する否定的な意見が現れている．一方「安部首相が北朝鮮を非難する」といった記述は単なる事実を述べた記述である．我々は前者を発信者の主観的記述，後者を客観的記述と呼ぶ．主観的記述とは発信者の主観が現れているような記述であり，このような記述は発信者の特徴を直に表しているものと考えられる．客観的記述は事実を述べた記述である．客観的記述は単に事実を述べているだけであり，発信者の特徴には無関係であると考えられるかも知れないが，我々は発信者の事実の取り上げ方に特徴が表れていると想定している．例えばある発信者が同じ事実を何度もとりあげるといった特徴や他の発信者が取り上げていないような事実を取り上げているといった特徴である．このような事実の取り上げる回数や事実の網羅度を分析するために客観的記述は重要である．

具体的に特定のエンティティに関する記述を抽出するために，まず我々はそのエンティティに対して用いられている表現について分析する．エンティティに対して用いられている形容詞，副詞などは発信者の意見が現れていると考え，発信者の主観的記述に分類される．またエンティティが何をどうしたなど，エンティティが現れる「主語-目的語-述語」の表現は，エンティティに関する事実を述べていると考えられ，客観的記述に分類される．このようにエンティティに用いられる表現のパターンによって主観的，客観的記述を抽出する．注意したいのは，ここでは文単位ではなく，表現単位で記述を抽出する．よって同じ文から主観的記述，客観的記述両方を抽出することもある．

続いて抽出した主観的記述と客観的記述を利用して，発信者の特徴ベクトルを生成する．特徴ベクトルを生成するにあたって，発信者の特徴を測るためのいくつかの指標を設定し，それぞれの指標に基づいて数値を算出し，ベクトルを生成する．まず一つ目の項目は，主観的記述に対する分析で，対象エンティティへの発信者の肯定，否定度を数値で表現する．次に二つ目の項目は，客観的記述に対する分析で，発信者が取り上げている対象エンティティに関する事実がバランスよいものか，偏ったものでないかを表す被覆度で数値化する．三つ目の項目は発信者が対象エンティティに対して主観的に書きやすいか，事実を取り上げる程度かを判定するためのもので，記述数で数値化する．これらの指標により発信者のエンティティに対する特徴ベクトルを生成する．

生成した特徴ベクトルを用いた応用例として，我々はニューストピック推薦アプリケーションについて議論する．発信者の過去の傾向と異なる記事は，希少であり注目に値する．1 日のトピックの中で各発信者の記事の特徴が過去の特徴と著しく異なるようなトピックは，読む価値があると考えられる．そのようなトピックを推薦するために，我々は記事から各エンティティに対する発信者の特徴ベクトルを計算し，発信者の過去の特徴ベクトルとのベクトル空間上の距離を計算する手法を提案する．

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. 章で特徴ベクトルの生成手法について述べる．4. 章で特徴的トピックの提示について述べる．6. 章で結論と今後の展望を述べる．

## 2. 関 連 研 究

我々はニュース発信者の特徴分析を行うが，発信者の特徴に注目した研究は多い．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 [4] 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．これらの研究の問題点として，記事内や文内で共起するキーワード間の関係を考慮していないので，キーワードがどのオブジェクトに対して記述されたものか明確でない点がある．よってニュース発信者の言葉の使い方は明らかになっても，特定オブジェクトに対する意見の傾向は明らかにならない．

著者らが以前行った研究 [5], [6] では，ニュースサイトの人物や組織に対する記述を主語，目的語，述語といった文内の語の役割に注目して抽出し，サイト内での特徴とサイト間でみた特徴を用いた特徴量を計算し，特徴的記述を提示した．サイトによっては特定の人物，組織に対して一貫の傾向を見出すことに成功した．しかし問題点として，記事文内での事実を述べている部分と意見が現れている部分というを区別していない点がある．両者では言葉の使われ方が異なるので，区別して分析しないと，結果として出される傾向の精度が低くなる恐れがある．

よって我々は，主観的記述，客観的記述という二種類の記述を分類することで，ニュース発信者のより精度の高い傾向を抽出する．他の関連研究として，ニュースの差異に関する研究と，文書の主観性，客観性の分類に関する研究を挙げる．

### 2.1 ニュースの差異に関する研究

我々は，ニュース発信者の記述の特徴を分析することで，発信者ごとのニュースの差異を発見することを目的としている．

ニュース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発信者ごとの意見の違いを見せるものが多い．

灘本ら [8] 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 [9] 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．

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発信者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．我々はニュース発信者の持っている特徴を分析し，一時的でなく恒常的な差異を見出す．

### 2.2 文書の主観性，客観性の分類に関する研究

文書の主観性，客観性に注目して分類を行う研究はこれまでにも行われている．Fin ら [10] は bag-of-words や品詞情報，人手で作成した主観性をもつ語のリストを用いて，ニュース記事を事実を述べられている記事と発信者の意見が述べられている記事に分類する手法を提案している．松本ら [11] は，助動詞や助詞といった文末表現を利用してウェブページの主観，客観度の判定を行い，ウェブ文書を分類するシステムを提案している．形容詞や形容動詞と比較して，文末表現を用いたほうがトピックに依存しない分類ができることを実験で示している．

これらの研究はウェブ検索における検索精度向上のためにウェブページの主観，客観性分析を行っている，一方，我々は，ニュース発信者による情報の偏り分析のために，発信者の特定エンティティに対する記述の主観，客観性に注目した．よって記事内の各文から主観的記述や客観的記述を取り出すので，主観，客観分析を文書単位ではなく文単位で行っている．

## 3. 発信者の特定エンティティに対する特徴分析

前述したように，ニュース発信者の特定エンティティに対する特徴を分析するために我々は記事文から発信者の特定エンティティに対する主観的記述，客観的記述を抽出する．そのために，まず，我々は発信者の特定エンティティに関する記事のトピック分類を行う．そしてトピックごとに特定エンティティに対する発信者の主観的記述と客観的記述をそれぞれ抽出する．最後に，抽出した記述を 3 つの尺度で数値化し，特徴ベクトルを生成する．

### 3.1 記事のトピック分類

我々は発信者の特徴ベクトルをトピック単位で生成する．トピックとはあるエンティティに関する事象や事件に関するニュース記事の集合である．トピック単位で特徴ベクトルを生成する理由は，同じトピックに関する記事は似た記事であり，特徴を測る単位としてふさわしいからである．またトピック単位でベクトルを生成することで，トピック間のベクトルの変動やある特徴を持ちやすいトピックなどを発見することもできる．トピック分類の手法であるが，これは現時点では既存手法を利用する予定である．具体的には Google ニュースのトピック分類を利用することを考えている．

### 3.2 記述の抽出

トピック分類を行った後，発信者ごとに記事から特定エンティティに対する主観的記述，客観的記述を抽出する．まず主観的記述，客観的記述を以下のように定義する．

- エンティティに対する主観的記述

エンティティに対する直接評価

- エンティティに対する客観的記述

表 1　エンティティX が現れる客観的記述

| 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) |

エンティティに関する事象，事実の記述

その際，発信者が特定エンティティに対して直接記述している部分と間接的に記述している部分の 2 種類をそれぞれ考慮する．

#### 3.2.1 主観的記述の抽出

「麻生は首相にふさわしい」など，発信者が特定エンティティを形容する言葉には発信者の主観が現れていると考えられる．一方,「麻生が注目を集めることになりそうだ」など文の内容が発信者の推定であるような場合も，発信者の主観が現れていると考えられる．よって前者を直接的な主観的記述，後者を間接的な主観的記述と区別して，それぞれ抽出する．

まず我々は特定エンティティに対する直接的な主観的記述を特定エンティティに対して用いられている形容詞や副詞といった品詞の語とする．これらの品詞を主観的品詞と呼ぶ．そのような主観的記述を取り出すために，我々は記事文から特定エンティティが現れる文を抽出し，その文を構文解析することで，構文木を作成する．これは主観的記述の分析に適した構文構造を定義して行う．続いて構文木上で特定エンティティに係る語の構造パターンを定義し，それに基づいて特定エンティティに対して用いられている語を特定する．特定された語の中で，形容詞や副詞といった主観的品詞の語を，発信者に対する主観的記述とする．

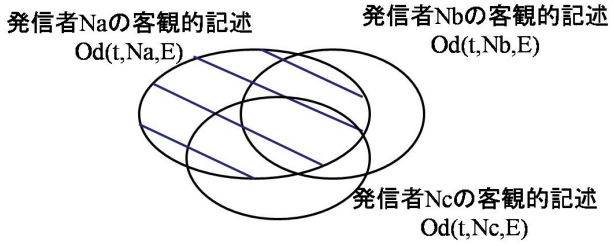
#### 3.2.2 客観的記述の抽出

客観的記述は特定エンティティに関する事実，事象を述べた記述である．客観的記述を抽出するために，我々は文の述語項構造を利用する．文の述語項構造とは文内の主語（〜が），目的語（〜を，〜に），述語（〜する）からなる構造で，誰が何をどうしたという事実を述べている部分である．特定エンティティに関する客観的記述はエンティティが主語または目的語として登場する述語項構造を元に取り出す．例えば，ある特定のエンティティX を仮定すると，X が主語となる「X が〜」が現れる述語項構造や，目的語となる「X を〜」「X に〜」が現れる述語項構造である．構文解析器を用いて，記事から述語項構造を取り出し，客観的記述を抽出する．

### 3.3 特徴ベクトルの生成

抽出した主観的記述，客観的記述から発信者の対象エンティティに対する特徴ベクトルを生成する．まず，主観的記述に対して感情分析を行うことにより，特定エンティティに対する発信者の肯定度を算出する．客観的記述に対しては他の発信者と比較することにより，発信者の特定エンティティに対する被覆度を算出する．最後に主観的記述の割合を求めることで主観度を算出する．以上 3 つの尺度より，発信者の特定エンティティに対する特徴ベクトルを生成する．

- 肯定度
- 被覆度
- 主観度




**発信者Naの被覆度**

$$\mathrm{cov}(t, Na, E) = \frac{|\,od(t, Na, E)\,|}{|\,od(t, Na, E) \cup od(t, Nb, E) \cup od(t, Nc, E)\,|}$$

図 1　発信者の客観的記述の被覆度

#### 3.3.1　肯　定　度

主観的記述には発信者の意見が現れている．そこで抽出した各主観的記述に対して，主観的品詞の語がエンティティに対して肯定的であるか，否定的であるかを分析する．そのために我々は，主観的品詞の語について感情分析により肯定度，否定度を求め，主観語辞書を作成する．主観語辞書では肯定的な語には正の値，否定的な語には負の値を割り当てる．例えば,「麻生がすばらしい」という記述では，すばらしいという語は肯定的な語なので，正の値があてられる．あるトピック t の記事におけるニュース発信者 $N_j$ の特定エンティティ E に対する肯定度 pn を下式で求める．

$$pn(t, N_j, E) = \sum_{i=1}^{n} score(sd_i(t, N_j, E)) \qquad (1)$$

但し $sd_i(t, N_j, E)$ は t における $N_j$ の E に対する主観的記述を表し，n はその総数を表す．$score(sd_i(t, N_j, E)$ は，主観語辞書での $sd_i(t, N_j, E)$ にあてられているスコアを表す．

#### 3.3.2　被　覆　度

そのエンティティに対してどれほど独特な事実をとりあげているかは発信者の特徴を表す一つの指標であると考えられる．これはいつもそのエンティティに対して局所的な事実しかとりあげていない発信者とエンティティに対して偏りなくとりあげている発信者には特徴に違いがあるという考えに基づいている．このような事実の被覆度は客観的記述の割合から求められる．あるトピック t の記事におけるニュース発信者 $N_j$ の特定エンティティ E に対する被覆度 dist を下式で求める．ここで比較するニュース発信者を $N_k$ と $N_m$ とする．

$$dist(t, N_j, E) = \frac{|od(t, N_j, E)|}{|od(t, N_j, E) \cup od(t, N_k, E) \cup od(t, N_m, E)|} \qquad (2)$$

但し $od_i(t, N_j, E)$ は t における $N_j$ の E に対する客観的記述を表す．

発信者の客観的被覆度を図説したものを，図 1 に示す．

#### 3.3.3　主　観　度

記事内でそのエンティティに対してどれほど記述しているか，またどの程度の割合で主観が入っているのかということも発信者の特徴を表現するのに重要なファクターである．発信者ごとに，エンティティに関する記述数は異なる．よって発信者ごとに主観的記述数と客観的記述数の割合から，発信者の主観度を求める．. あるトピック t の記事におけるニュース発信者 $N_j$ の特定エンティティ E に対する主観度を下式で求める．

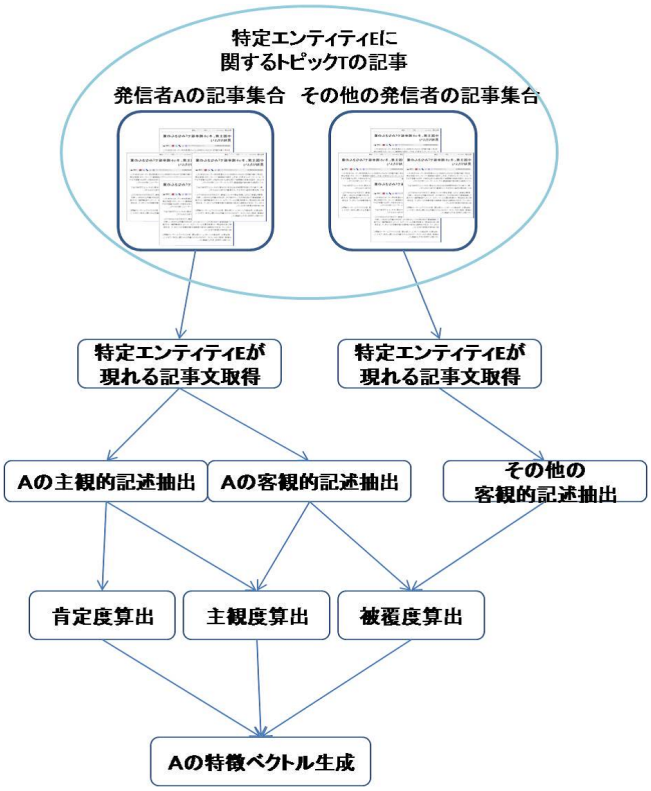



図 2　特徴ベクトル生成の流れ

$$sr(t, N_j, E) = \frac{|sd(t, N_j, E)|}{|sd(t, N_j, E)| + |od(t, N_j, E)|} \qquad (3)$$

#### 3.3.4　特徴ベクトルの生成

以上述べた三つの指標により，あるトピックにおける発信者の特定エンティティに対する特徴ベクトルを生成する．ベクトル生成の流れは図 2 に示す．トピック t における発信者 $N_j$ のエンティティ E に対する特徴ベクトル F は次のように表される．

$$F(t, N_j, E) = (pn(t, N_j, E), dist(t, N_j, E), sr(t, N_j, E))$$

## 4.　特徴的なトピックの提示

我々は，発信者の特徴ベクトルを利用した応用例として，特徴的なトピックの提示を提案する．ここでいう特徴的なトピックとは発信者の特徴や発信者間の特徴が過去との傾向と著しく異なるようなトピックとする．例えば発信者全体が麻生に否定的である状況の中である日のあるトピックにおいては肯定的な発信者が多かったとする．そうするとそのトピックには過去の発信者の特徴という観点から見れば，特徴的であり読むに値するものであると考えられる．つまり過去との傾向といかに異なるかという過去との異なり度でもって特徴的なトピックを発見するのである．

まず一日のニュースをトピック分類し，トピックごとの記事集合を得る．そして各記事集合から，文内に登場するエンティティを抽出し，発信者ごとに 2 種類の特徴ベクトルを生成する．まず記事集合内で特徴ベクトルを作成し，これを現在の特

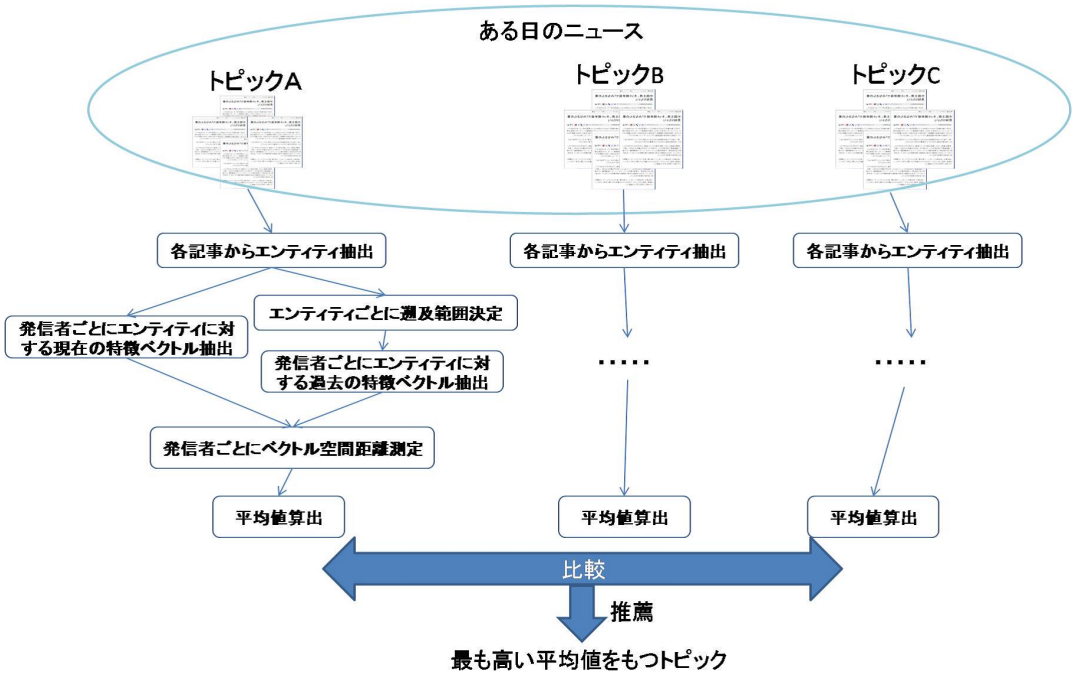



図 3　特徴的なトピックの提示

徴ベクトルとする．続いてエンティティごとに遡及範囲を求める．過去の傾向は推移していくもので常に一定ではないため，過去の範囲を決める必要がある．よって遡って，発信者ごとの傾向が大きく変化した点までを遡及範囲とする．遡及範囲までの記事集合で発信者ごとに特徴ベクトルを生成し，これを過去の特徴ベクトルとする．そして発信者ごとに過去と現在の特徴ベクトルの空間距離を測り，平均値を求める．さらに，これをエンティティごとに行いトピックごとに平均値を求める．これらの平均値が最も大きいトピックを，過去の傾向と最も異なるトピックとして推薦する．

特徴的トピックの提示のプロセスフローを図 3 に示す．

## 5. 実　　験

本手法による記事から発信者の主観的記述，客観的記述の抽出精度を測るために実験を行う．まず大手の新聞社 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た．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．

次に各トピックの記事ごとに，文を日本語構文解析システム KNP [12] を用いて解析する．KNP の形態素解析より品詞情報を得ることができ，また構文解析により主観的記述の抽出，格解析により客観的記述の抽出が可能となる．今後，抽出した主観的記述，客観的記述の精度評価を行う予定である．

## 6. 結論と今後

本論文では，ニュース発信者の特定エンティティに対する特徴を測るために，主観的記述と客観的記述を用いて，発信者のエンティティンに対する特徴ベクトルを生成した．さらに，特徴ベクトルを用いたニューストピック推薦システムについて議論した．

今後は，自然言語処理による主観的、客観的記述の抽出精度を測定，特徴ベクトルを用いたトピック抽出の評価を行う予定である．

## 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くださった読売新聞殿，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] くらべる一面 ： 新 s あらたにす（日経・朝日・読売). http://allatanys.jp/.
[3] Fairspin. http://fairspin.org/.
[4] 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に注目した発信者意図の抽出. *2008-DBS-146*, pp. 187–192, 2008.
[5] S. Ishida, Q. Ma, and M. Yoshikawa. Analysis of news agencies ' descriptive features of people and organizations. *Proc. 20th Int. Conf. Database and Expert Systems Applications, LNCS 5690*, pp. 745–752, 2009.
[6] S. Ishida, Q. Ma, and M. Yoshikawa. Analysis of news agencies' descriptive feature by using svo structure. *Fourth International Conference on Digital Information Management (ICDIM),*, 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, 2008.
[8] A. Nadamoto and K. 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.

[9] Q. Ma and M Yoshikawa. Topic and viewpoint extraction for diversity and bias analysis of news contents. *Proc. of APWebWAIM2009, LNCS 5446*, pp. 152–160, 2009.

[10] A. Finn, N. Kushmerick, and B. Smyth. Genre classification and domain transfer for information filtering. *Proc. of ECIR-02, 24th European Colloquium on Information Retrieval Research*, 2002.

[11] 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄. 文末表現を利用したウェブページの主観・客観度の判定. DEIM2009 論文集, A5-4, , 2009.

[12] 日本語構文解析システム knp. http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html.