2008－GN－67 (2)
2008／3／21

読者に影響を与えるブログ記事発見手法

宮田 章裕 川島 晴美 奥田 英範
日本電信電話株式会社 NTT サイバーソリューション研究所

本稿では，読むことで新たな情報を得たり，共感等の感情を抱いたりできる記事を “読者に影響を与えるブログ記事” と定義し，これを発見する手法を提案する．提案手法では，コメント・トラックバックといったブログ記事の読者が起こした行動の痕跡を分析することで，その記事が読者にどのような影響を与えたのか推定するアプローチを採る．コメント・トラックバック等に関する 20 個の素性 (反響素性) を定義し，“読者に影響を与えるブログ記事” が持つ反響素性の特徴を利用して機械学習を行うと，文書特徴を学習対象としたベースライン手法と比べ，大幅に少ない素性で高い識別性能を得られ，記事のジャンルの違いによる影響を受けにくいことが確認できた．

A Method of Detecting Influential Blog Entries

Akihiro Miyata Harumi Kawashima Hidenori Okuda
NTT Cyber Solutions Laboratories, NTT Corporation

We present a method of detecting *influential blog entries*, defined as entries where readers can obtain new information or feel empathy. We consider the “footprints” left behind by readers, which primarily consist of comments and trackbacks. Our method analyzes and detects *influential blog entries* by evaluating twenty comment and trackback attributes using a machine learning based approach. A comparison of our method against the baseline method (text-based approach) results in better performance while using fewer attributes.

1 はじめに

ブロガーが実際に使用，あるいは強い関心を抱いた商品の紹介がブログ記事として投稿される機会が増えてきた．これらの記事では好評も悪評も含めた消費者の率直な意見が多く述べられ，他のメディアからは得られないような情報が含まれていることが多い．

しかし，すべてのブログ記事が有益であるとは限らない．商品紹介記事の例で言えば，商品を買った感想や利点・欠点が記されている場合もあるが，商品を買ったと述べているだけの記事や，自分が欲しい商品をリスト化しているだけの記事も少なくない．これがまさに “ブログ記事は玉石混淆” と称されている現状であり，ユーザにとって有益な記事を見つけ出すことは容易ではない．

この問題を解決するために，本研究では分析対象記事の読者が起こした行動の痕跡 (コメント・トラックバック等) を分析することで，読者がその記事にどのような影響を受けたのか推定する手法を採る．本研究では，有益な記事の一例として，読むことで新たな情報を得たり，共感等の感情を抱いたりできる記事を “読者に影響を与える記事” と定義し，機械学習手法を用いてこのような記事を発見する手法を提案する．

以降，2 章で機械学習を用いたブログ記事分析の関連研究を紹介し，3 章で提案概念を述べる．4 章で検証を行い，5 章で結論を述べる．

2 機械学習を用いたブログ記事分析

ウェブ上に溢れているブログ記事の中に埋もれた情報を発見するために，様々な分析技術が提案されている．ここでは本研究と関連が深い事例として，機械学習を用いてブログ記事分析を行う先行研究を紹介する．

Xiaochuan らは機械学習を利用して，ブロガーの専門・職業に関連したトピック指向の記事と，ブロガーの日常的なできごとや感情を綴った記事の分類を行っている．[1]．彼らは記事本文中に出現する単語を抽出して学習データを作成し，Naive Bayes Classifier や Support Vector Machine 等の機械学習手法を利用して記事の自動分類を試みている．

Beibei らは，ブログ空間上から消費者の率直な意見や危険思想の兆候を発見するために，文書中に出現する各単語の TF/IDF 値から成る特徴ベクトルを作成し，これを利用してブログ記事群のクラスタリングを行っている [2]．ここで注目すべきは，彼らが記事本文の単語だけでなく，そこに付与されたコメント中に出現する単語も利用している点である．コメント中には，ブログ記事本文の中でも特に読み手が関心を持った部分が再掲されていたり，書き手による説明・補足が追記されていたりすると彼らは考察している．彼らの実験結果によると，コメントを考慮することでクラスタリング性能は大幅に上昇している．

3 読者に影響を与えるブログ記事発見手法の提案

3.1 研究目標

人は意思決定の際に，情報収集を行うことが多い．例えば，欲しい商品があれば商品販売元のウェブサイトや雑誌を閲覧して商品の詳細を調べる．しかし，企業やマスメディアから発せられる情報には消費者視点の意見が含まれていない場合や，企業側に不都合な情報は隠蔽される場合がある．また，発信主体が限られているため，情報に不足・偏りが生じるおそれもある．

これに対し，ブログ記事には消費者の率直な意見が含まれていることが多い．これは，情報発信者である消費者が多くの場合，営利を求めておらず，日々感じていることを自発的に表現しているためと思われる (「この店は美味しい．また行きたい.」,「この商品は失敗．買って損した.」等)．また，ブログ開設数は国内で 1000 万件を超えており，情報の不足や偏りも生じにくいと思われる．つまり，消費者が購入前の商品を吟味する際等には，ブログ記事は非常に有益な情報源であると言える．

しかし，ブログ記事は玉石混淆と称されることも多く，その質は様々である．例えば商品 X の名前を含む記事 A と記事 B があっても，記事 A では商品 X を使用した感想を詳細に述べている一方，記事 B では「商品 X が欲しい」と述べられているだけで読者は何の情報も得られない場合がある．また，これは商品 X が人気商品である場合に多く見受けられるが，記事内容は商品 X とは全く関係無いにも関わらず，文中に商品 X の名前を多数含めることで多くのユーザに閲覧してもらうことを狙う悪質な記事もある．このため，商品 X の名前をクエリとしたブログ記事検索を行っても，検索結果の記事から有益な情報が得られない場合が少なくない．

そこで本研究では，玉石混淆のブログ記事の中から有益な記事を見つけ出すことを目指す．特に，ユーザが幅広いジャンルの商品に対する消費者視点の情報・感想を把握するシーンを想定する．このシーンにおける “有益な記事” には，その記事を読むことで商品に関する新たな情報を得られたり，商品に対する感情 (「興味が湧いた」,「欲しくなくなった」等) を抱いたりできるという特徴があると思われる．本研究では，このような記事を “読者に影響を与えるブログ記事” と定義し，非特定分野においてこれを発見する手法の確立を目標とする．

3.2 既存手法の問題点

ブログ記事の特徴分析を行う問題に対しては言語処理のアプローチが多く採られる．言語処理は研究が活発な分野であり，実用的な技術も数多く登場している．しかし，本研究の目標を達成するためには，“ドメイン・言語への依存”，“更新のコスト”，“言語処理技術の限界” の問題があると思われる．

例えば，特定分野のウェブ上の文書から肯定・否定表現を抽出する手法がある [3, 4]．この手法を用いれば，単に「商品 X を買った.」と述べるだけの記事ではなく，「商品 X を買った．デザインが良い.」等と意見を述べている記事を発見でき，このような記事は読者に影響を与えやすいと思われる．ただし,「車」や「デジカメ」等の分野ごとに抽出ルールを構築しなければならないため，多くの分野に対応するためにはコストがかかる．また，日本語と英語のように言語が違えば抽出ルールも異なるので，複数言語に対応させるためにはその分抽出ルールも作成しなければならない (ドメインの依存)．

一方，非特定分野の対象から意見表現 (肯定・否定・評価・願望・見解等) を抽出する手法 [5] や比較表現を抽出する手法 [6, 7] も提案されている．しかし，他のメディアと比べて，ブログ記事には次々と生まれる新語やインフォーマルな表現 (“良い” の意味で主に若者が使う「ヤバい」,「アツい」,「マチガイない」等) が出現しやすいため，抽出ルールを常に更新し続けなければ精度を高く保つことは難しいと思われる (更新のコスト)．

さらに，意見・比較表現等を含む記事を発見できたとしても，人間が書いた自然言語が分析対象である以上，その記事の価値を判定することは容易ではない．例えば，あるブログ記事で「商品 X は面白い.」と述べられていたとしても，この意見は妥当な根拠に基づいて述べられているのか，読者の心を揺り動かす内容なのか，といったことは現在の言語処理技術だけで判定することは難しい (言語処理技術の限界)．

3.3 提案手法

本研究では，ブログ記事の読者が起こした行動の痕跡を分析することで，その記事が読者にどのような影響を与えたのか推定する手法を採る．我々はこの手法を**反響特性分析** [8] と呼び，読者によって記事に送信されたコメント・トラックバックを多面的に分析することで，その記事がどのような反響を呼んでいるのか (“幅広い人から長期間参照されている” 等) 判定するアプローチを採ってきた．反響特性分析には下記 3 つの特徴があり，研究目標達成のために有効であると考えられる．

1. **ドメインへの依存が少ない**
   商品紹介記事に対してコメント・トラックバックが送信される様子は，商品分野，記事が書かれている言語の影響を受けにくいと思われるので，分析モデルを分野・言語ごとに作成するコストを低減できる．

2. **更新のコストが少ない**
ブログ空間上に次々と新しい言葉が登場している状態と比べれば，人々が商品紹介記事に対してコメント・トラックバックを送信する振る舞いは大きく変化しないと思われるので，一旦分析モデルを作成すれば頻繁にモデルを更新する必要性は小さい．

3. **言語処理技術の弱点を補える**
現在の言語処理技術では記事内容は読者の心を揺り動かす内容なのか，といった側面まで推測することは困難であり，この点は人間が実際に記事を読んで判断するしかない．反響特性分析は，読者が様々な判断を下した結果起こした行動を分析するので，記事に対する人間の判断結果を推定しやすい．

表 1 に示すのは，反響特性分析で利用するコメント・トラックバックの属性 (以降「反響素性」) である．(17)〜(20) は記事本文の属性であるが，これは記事属性とコメント・トラックバック属性の関係 (“記事の長さが短いのにコメントを送信した人数が多い” 等) も重要と考えているため利用している．“読者に影響を与えるブログ記事” を発見するためには，これらの記事が持つ反響素性のモデルを抽出する必要がある．反響素性は素性の数が多いので，効率良く的確なモデル抽出を行うために本研究では機械学習手法を用いる．

表 1: 反響素性

(1)CM 送信者数	CM を送信した人数
(2) 平均 CM 送信数	CM 送信者 1 人あたりの平均 CM 送信数
(3)CM リンク率	CM 総数における，CM 送信者の URL が書いてある CM 数の割合
(4)CM タイトル空欄率	CM 総数における，CM タイトルが空欄の CM 数の割合
(5)CM 送信者空欄率	CM 総数における，CM 送信者名が空欄の CM 数の割合
(6) 平均 CM 文字数	CM に含まれる平均文字数 (絵文字は除く)
(7) 平均 CM 絵文字数	CM に含まれる平均絵文字数
(8) 平均 CM 内リンク数	CM に含まれるハイパーリンク数
(9) 初 CM 受信経過時間	記事が投稿されてから最初の CM を受信するまでに経過した時間
(10) 最終 CM 受信経過時間	記事が投稿されてから最終の CM を受信するまでに経過した時間
(11) 平均 CM 間隔	CM 受信時間間隔の平均値
(12)TB 送信者数	TB を送信した人数
(13)TB リンク率	TB 総数における，記事にリンクした上で送信されている TB 数の割合
(14) 初 TB 受信経過時間	記事が投稿されてから最初の TB を受信するまでに経過した時間
(15) 最終 TB 受信経過時間	記事が投稿されてから最終の TB を受信するまでに経過した時間
(16) 平均 TB 間隔	TB 受信時間間隔の平均値
(17) 記事文字数	記事に含まれる文字数 (絵文字は除く)
(18) 記事絵文字数	記事に含まれる絵文字数
(19) 記事画像数	記事に含まれる画像数 (絵文字は除く)
(20) 記事リンク数	記事に含まれるハイパーリンク数

(CM=コメント，TB =トラックバック)

4 評価実験

4.1 実験概要

この実験の目的は，機械学習を用いて反響特性分析を行うことで，“読者に影響を与えるブログ記事” を発見できるかどうか確認することである．今回の実験では，予め “読者に影響を与えるブログ記事” とそうでない記事を収集しておき，これらに対する提案手法/ベースライン手法の識別性能を比較検証する．機械学習手法は，Naive Bayes Classifier(NB)，C4.5，Support Vector Machine(SVM) を利用する．SVM では多項式カーネルを用い，次数は 1，2 の 2 通りを検証する．

4.2 学習データ/テストデータ

“読者に影響を与えるブログ記事” の学習データ/テストデータを作成するために，表 2 に示す商品名を含むブログ記事をそれぞれ最大 500 件収集した* ．なお，各商品はいずれも 2006〜2008 年に話題になったものであり，ブログ記事でも頻繁に取り上げられていた．

収集した記事の中から，スパム記事 (公序良俗に反する内容の記事)，機械生成と思われる記事 (他のブログ記事のコピーだけで構成されている，内容が文章として全く意味を成していない等) を人手で取り除いた．また，反響特性分析はコメント・トラックバック情報を利用するので，コメントまたはトラックバックが 1 つも無い記事は除外し，実験結果に悪影響が及ぶのを避けるためスパム記事・機械生成と思われる記事から 1 つでもコメント・トラックバックを受けている記事も取り除いた．

そして，残った記事 790 件を実験者が実際に閲覧し，そのブログ記事が商品に関して読者に影響を与えている場合は “Positive”，そうでない場合は “Negative” というラベルを付与した．判定基準を明確にするために，判定は図 1 に示すフローチャート (番号は表 3 参照) に基づいて行った．このチャートでは判定できない/判定に迷う場合は “Invalid” というラベルを付与した．その結果，Positive は 271 件，Negative は 163 件，Invalid

* 記事 URL は goo ブログ検索 (http://search.goo.ne.jp/，文書適合度順，goo ブログのみを検索対象に指定) を利用して取得した．

表 2: 分析対象とした商品名

商品名	検索用キーワード
Coca Cola Zero (清涼飲料水)	「Coca Cola Zero」OR 「コカ コーラ ゼロ」
ICE CUCUMBER (清涼飲料水)	「コーラ」AND 「キュウリ」
McWrap (ファーストフード)	「McWrap」OR 「マックラップ」
Segreta (シャンプー)	「Segreta」OR 「セグレタ」
Wii Fit (玩具)	「Wii Fit」OR 「Wii フィット」
人生銀行 (玩具)	「人生銀行」
クロックス (サンダル)	「クロックス」
iPod touch(音楽プレイヤー)	「iPod touch」

は 356 件となった．下記に，表 3 中の 5(1) 下線部を満たす/満たさないコメントの例を示す．なお，原文掲載は問題が生じかねないので，文意を損ねない範囲で省略・改変を行っている．

[5(1) 下線部を満たすコメントの例]

- こんなコーラあるんだ !?キュウリ味なんて想像できないなぁ。面白いこと考える人がいるね〜。(ICE CUCUMBER の紹介記事へのコメント)
- このブログでお孫さんが遊んでいらっしゃるのを拝見して、これなら私にもできそうと、早速買いました。(Wii Fit の紹介記事へのコメント)
- どこでも売り切れみたいですね。私はオンラインショップで買いました。そういえば、北千住の駅ビルで売っているのを見かけましたよ♪ (クロックスの紹介記事へのコメント)

[5(1) 下線部を満たさないコメントの例]

- はじめまして。私もこの島に引っ越してきました。過去記事も拝見させていただきます。(ICE CUCUMBER の紹介記事へのコメント)
- こんばんは。今度自分が関わっている車の展覧会が開催されるのでぜひ来てください！(Wii Fit の紹介記事へのコメント)
- あけましておめでとう。またこのブログ覗きにくるよ♪ (Wii Fit の紹介記事へのコメント)

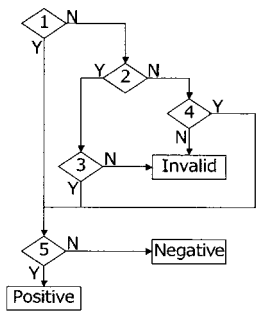

図 1: 判定フローチャート

表 3: 判定フローチャートの項目詳細

1	記事 A のタイトルに対象商品名 (または同義語) が含まれている.
2	A の本文に含まれるテーマは 1 つである.
3	A の本文に含まれるテーマは対象商品に関するものである.
4	A の本文に含まれるテーマの半数以上は対象商品に関するものである.
5	次の条件を 1 つ以上満たす. (1) コメントの半数以上が対象商品に関する意見・経験・情報を含んでおり，かつ，A から情報を獲得した旨・A に対する共感や感想・A を踏まえた上での追加情報 のどれかを含んでいる. (2) トラックバック元の記事の半数以上が対象商品に関する意見・経験・情報を含んでおり，かつ，A へのリンク・A の引用・A への言及のどれかを含んでいる.

作成した Positive/Negative ラベル付きの記事群全体をデータセット D1 とする．D1 の中で，コメントを 1 つ以上受信している記事のみを抽出し，これを D2 とする．D2 の中で，「ICE CUCUMBER」の紹介記事からなるデータセットを D3,「Wii Fit」の紹介記事からなるデータセットを D4 とする．D3 と D4 に重複している記事は無い．各データセットの情報を表 4 に示す．

表 4: データセット

	紹介している商品	CM/TB 条件	記事数 (P/N)
D1	すべて	CM > 0 OR TB > 0	434(271/163)
D2	すべて	CM > 0	365(262/103)
D3	ICE CUCUMBER	CM > 0	98(52/46)
D4	Wii Fit	CM > 0	96(63/33)

(CM=コメント,TB =トラックバック)

4.3 複数分野の記事に対する検証 (1)

ここでは，複数分野の記事を含む D1 に対して 3-fold の交差検定を行う．提案手法では，D1 内の各記事の反響素性 (1)〜(20)(表 1 参照) を用いて特徴ベクトル (以降「反響特徴ベクトル」) を作成し，交差検定を行う (RSP-1)．一方，ベースライン手法は関連研究 [1] と同様，各記事に出現する単語を用いて特徴ベクトルを作成する．具体的には，記事本文中に出現する単語 (名詞・動詞・形容詞・副詞†) を抽出し，各単語の TF/IDF 値を素性とする特徴ベクトル (以降「文書特徴ベクトル」) を作成して交差検定を行う (TXT-1).

† 形態素解析ソフトウェアには mecab-0.96，mecab-ipadic-2.7.0-20070801 を改変せずに用いた．名詞には未知語も含め，非自立語，代名詞，接尾語，数詞，記号，英字 1 文字の語は除去した．動詞・形容詞は自立語のみ利用した．

検定結果を表 5 に示す．F は $2 \times Precision \times Recall/(Precision + Recall)$ で表される評価尺度である．表5を見ると，どの機械学習手法においても RSP-1 が TXT-1 より高い Precision，Recall，F 値を示していることが分かる．‡

つまり，複数分野の記事を分析する際は，文書特徴を利用するよりも，読者の行動情報を利用する方が高い識別性能が得られることが分かる．ただし，RSP-1 ではコメント・トラックバック・記事関連の素性を利用しているのに対して，TXT-1 では記事部分の文章しか利用していないので単純に比較することはできない．

表 5: 複数分野の記事に対する検証結果 (1)

		NB	C4.5	SVM(d=1)	SVM(d=2)
RSP-1	Precision	0.704	0.810	0.789	0.754
	Recall	0.629	0.756	0.676	0.674
	F	0.664	0.782	0.728	0.711
TXT-1	Precision	0.536	0.513	0.535	0.539
	Recall	0.538	0.514	0.535	0.539
	F	0.537	0.514	0.535	0.539

4.4 複数分野の記事に対する検証 (2)

RSP-1 と TXT-1 では性能にかなり差が出たため，ベースライン手法の拡張を行う．具体的には，ベースライン手法においても記事本文だけでなく，コメント中の文章も利用して検証を行う§．データセットは，すべての記事が 1 つ以上のコメントを持つ D2 を用いる．ベースライン手法はコメント・記事本文中に出現する単語を用いて文書特徴ベクトルを作成する (TXT-2)．提案手法もベースライン手法に合わせてトラックバックに関する反響素性を利用せず，コメント・記事関連の反響素性 (1)〜(11)，(17)〜(20) だけを用いて反響特徴ベクトルを作成する (RSP-2)．そして，RSP-2/TXT-2 の各手法で D2 に対して 3-fold の交差検定を行う．

検定結果を表 6 に示す．この表を見ると，C4.5 を用いたときに RSP-2 の F 値は最大 (0.682) となり，このとき TXT-2 の F 値は 0.615 であり性能に差が出ている．TXT-2 で最も高い性能を示したのは SVM(d=1) であるが，この場合の RSP-2・TXT-2 の F 値はそれぞれ 0.653，0.632 となっており，やはり RSP-2 の方が高い性能を示している．また，両者には識別に要する処理時間 (識別モデル構築時間を含む) に大きな差があり，例えば C4.5 の識別器では RSP-2 が 0.37sec，TXT-2 が 24.40sec となっている．これは，特徴ベクトルの次元数の差 (RSP-2 の反響特徴ベクトルは 15 次元，TXT-2 の文書特徴ベクトルは 9184 次元) が大きく影響していると思われる．

表 6: 複数分野の記事に対する検証結果 (2)

		NB	C4.5	SVM(d=1)	SVM(d=2)
RSP-2	Precision	0.545	0.722	0.864	0.616
	Recall	0.545	0.646	0.524	0.532
	F	0.545	0.682	0.653	0.571
TXT-2	Precision	0.608	0.613	0.652	0.628
	Recall	0.599	0.617	0.614	0.622
	F	0.603	0.615	0.632	0.625

つまり，複数分野の記事を分析する際に，ベースライン手法を拡張してコメント・記事部分を利用する TXT-2 と，提案手法を一部縮小して同じくコメント・記事関連の素性だけを利用する RSP-2 を比較しても，提案手法の方が概ね良好な識別性能を得られ，その処理速度は数十倍高速であることが分かる．

4.5 単一分野の記事に対する検証

我々は 3.2 節において言語処理アプローチには “ドメインへの依存” が問題になると述べた．つまり，あるデータセット A で学習したモデルを別のデータセット B に適用すると，ベースライン手法では識別性能が下がり，提案手法ではさほど識別性能が下がらないと考えている．これを確認するために，単一分野の記事からなるデータセットを用いて検証を行う．まず，「ICE CUCUMBER」の紹介記事からなる D3 に対して提案手法/ ベースライン手法で交差検定を行う (RSP-3/TXT-3)．次に，学習データを D3，テストデータを「Wii Fit」の紹介記事からなる D4 に設定して提案手法/ベースライン手法で検証を行う (RSP-4/TXT-4)．なお，各特徴ベクトルの作成方法は RSP-2/TXT-2 と同様である．

RSP-3/TXT-3 の結果を表 7，RSP-4/TXT-4 の結果を表 8 に示す．TXT-3 は TXT-2 と比べて大幅に識別性能が向上し，RSP-3 を上回っている．これは，1 つの商品だけに関するデータセット D3 を対象にしたため，Positive の記事・コメントに出現する単語が限定的 (「味」，「爽やか」，「美味しい」等) で，文書特徴の学習が効率良く作用したためだと思われる．一方，表 8 を見ると，TXT-4 の識別性能が TXT-3 と比べて大幅に下がっていることが分かる．これは，D3 と D4 では商品ジャンルが違い (D3:清涼飲料水/D4:玩具)，Positive の

‡ 紙面の都合により詳細は省くが，RSP-1 で最高の F 値を示した C4.5 が抽出したモデルでは，「反響素性 (1) が 5 人以上」や「反響素性 (6) が 31 文字以上」が Positive と判定されるための主要因となっていた．

§ トラックバック中の文章 (つまり，トラックバック元の記事本文) も利用して検証を行うのが理想的であるが，本実験の環境ではトラックバック元の記事は保有していなかったので今回は利用しない．

記事・コメントに出現する単語の傾向に差があったためだと思われる．D4 中の Positive の記事には「ゲーム」，「ヨガ」，「頑張る」等が多く出現していた．

つまり，商品ジャンルが異なれば出現単語等の文書特徴も異なるが，ベースライン手法は文書特徴を学習対象としているために識別性能はドメインに依存してしまったのだと判断できる．これに対し，RSP-3 と RSP-4 を比較すると，学習データとテストデータの商品ジャンルが同じ場合でも異なる場合でも，一定の識別性能を保持できていることが分かる．

表 7: 検証結果 3

		NB	C4.5	SVM(d=1)	SVM(d=2)
RSP-3	Precision	0.621	0.632	0.687	0.613
	Recall	0.618	0.626	0.677	0.574
	F	0.619	0.629	0.682	0.593
TXT-3	Precision	0.728	0.683	0.733	0.747
	Recall	0.719	0.683	0.722	0.548
	F	0.723	0.683	0.727	0.632

表 8: 検証結果 4

		NB	C4.5	SVM(d=1)	SVM(d=2)
RSP-4	Precision	0.560	0.502	0.660	0.591
	Recall	0.548	0.501	0.633	0.551
	F	0.554	0.502	0.646	0.570
TXT-4	Precision	0.565	0.524	0.570	0.643
	Recall	0.556	0.527	0.569	0.564
	F	0.560	0.525	0.570	0.601

5 おわりに

本研究では，“読者に影響を与えるブログ記事” を発見する技術の確立を目指し，反響特性分析と機械学習を組み合わせた手法を提案した．分析項目としてコメント・トラックバック等に関する 20 個の反響素性を定義し，“読者に影響を与えるブログ記事” が持つ反響素性を機械学習するアプローチを採った．

検証実験では，複数分野の記事を分析対象とした場合，提案手法が記事部分のみを利用したベースライン手法よりも識別性能が高いこと，コメント・記事部分を利用したベースライン手法の数十倍の速度でより高い識別性能を実現できることが確認できた．特に，学習対象が少ない素性で済むことは，大規模なデータマイニング等を行う際に有利である．また，単一分野の記事を分析対象とした場合，学習データとテストデータの分野が異なっても提案手法は一定の識別性能を保持できることが確認できた．このように，提案手法は分野への依存が少ないので，複数の分野や未知の分野が混在したデータに対しても一定の識別性能を発揮できるものと思われる．

今後は，反響特徴ベクトルの素性の追加・取捨選択を行うことで識別性能の向上を目指し，商品紹介以外の分野における適用可能性も検証する．また，提案手法と言語処理技術を組み合わせることで，識別性能の向上やコメント・トラックバックが無い記事への対応も可能になると思われるので，検討を重ねていく方針である．

参考文献

[1] Ni, X., Xue, G. R., Ling, X., Yu, Y. and Yang, Q.: Exploring in the Weblog Space by Detecting Informative and Affective Articles, *Proceedings of 16th International World Wide Web Conference (WWW2007)* (2007).

[2] Li, B., Xu, S. and Zhang, J.: Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments, *Technical Report No.462-06, Department of Computer Science, University of Kentucky* (2006).

[3] Dave, K., Lawrence, S. and Pennock, D. M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, *Proceedings of 12th International World Wide Web Conference (WWW2003)* (2003).

[4] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 意見抽出のための評価表現の収集, 自然言語処理, Vol. 12, No. 2, pp. 203–222 (2005).

[5] 古瀬蔵, 廣嶋伸章, 山田節夫, 片岡良治: ブログ記事からの意見文検索, 研究報告 - 自然言語処理, Vol. 2006, No. 124, pp. 121–128 (2006).

[6] 佐藤敏紀, 奥村学: blog からの比較関係抽出, 研究報告 - 自然言語処理, Vol. 2007, No. 94, pp. 7–14 (2007).

[7] 倉島健, 別所克人, 内山俊郎, 片岡良治: 比較評価情報の抽出とそれに基づくランキング手法の提案, 第 18 回データ工学ワークショップ (DEWS 2007) (2007).

[8] 宮田章裕, 松岡寿延, 岡野真一, 山田節夫, 石打智美, 荒川則泰, 加藤泰久: 反響特性分析を利用したブログ記事検索手法, 情報処理学会論文誌, Vol. 48, No. 12, pp. 4041–4050 (2007).