利用者と書き手の類似性を考慮したブログ上の評判分析

一本　真志† 　藤本　悠† 　大原　剛三†

† 青山学院大学理工学部情報テクノロジー学科　〒 252–5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: †a5808006@aoyama.jp, {yu.fujimoto,ohara}@it.aoyama.ac.jp

あらまし　本研究では，利用者と趣味や嗜好の類似したブロガーの記事に着目し評判分析を行うことにより，類似性を考慮しない既存の手法よりも利用者にとって有益となる評判情報を得ることを目的とする．具体的には，評判情報を知りたい対象に関連するキーワードを blog 記事から抽出し，それらに基づいて利用者と類似したブロガーを特定する．利用者の嗜好は関連キーワードに対して評価をつけてもらい判断し，ブロガーの嗜好は過去の blog 記事から評判情報を抽出することにより特定する．また本稿では，提案手法と通常の手法で特徴的な差異がみられるかを調べるために行った実際の blog 記事を対象とした両手法での分析結果の比較についても報告する．

キーワード　評判分析, ブログマイニング, 嗜好判別

Masashi ICHIMOTO†, Yu FUJIMOTO†, and Kouzou OHARA†

† College of Science and Engineering, Aoyama Gakuin University
5-10-1 Fuchinobe, Chuo-ku, Saganihara-shi, Kanagawa 410–2415 Japan
E-mail: †a5808006@aoyama.jp, {yu.fujimoto,ohara}@it.aoyama.ac.jp

1. はじめに

近年，ブログや掲示板，SNS（Social Networking Service）などの CGM（Consumer Generated Media）が消費者の間に浸透したことにより，個人は情報の受け手となるだけでなく，手軽に情報を発信できる立場となった．CGM 上で個人が発信する情報には，さまざまな商品やサービス，TV 番組や映画などに対する主観的な評価や意見がある．こういった情報源を有効に利用するため，近年，CGM から必要な評判情報を自動的に収集・分析する評判分析の研究が盛んに行われている [1]．たとえば，評価表現とその表現の極性の対をまとめた評価表現辞書を用いたもの [4, 7]，判別器を用いて文中の表現が評判情報であるかどうかを判定する手法 [5, 6] などが提案されている．

加えて，多くの企業が blog を対象とした評判分析サービスを提供している [2, 3]．これらのサービスでは，前述のような要素技術を用いて blog 上での対象物に関する発言が肯定的（Positive）か否定的（Negative）かの判定（P/N 判定）及びその割合の表示，時系列による対象物の話題度の推移，関連語の抽出などを行っている．これらのサービスによってユーザーは話題の商品，人物，出来事についての世間での評判を知ることができる．しかし，世間のさまざまな価値観や考えを持った人たちの意見が統合されるため，分析結果が一般化され特徴的な情報が得られにくいという問題点もある．

そこで本研究では，趣味や価値観などが類似したブロガー内での評判分析を行い，利用者にとってより有益となる情報を得るための手法を提案する．この手法によって,「愛煙家たちの煙草増税に対する意見」のような，ある特定の集団に特徴的な意見や評判を得ることが期待できる．具体的な手法としては，評判情報を知りたい対象に関連するキーワードを抽出し，利用者にそれら関連キーワードに対する評価付けを行ってもらう．その上で関連キーワードへの評価傾向が類似しているブロガーを特定し，そのような類似ブロガー集団内での評判分析をする．また本稿では，提案手法と通常の手法で特徴的な差異がみられるかを調べるために行った実際の blog 記事を対象とした両手法での分析結果の比較についても述べる．

2. システム概要

本研究では，利用者と類似した嗜好を持つブロガーに着目した評判分析システムを提案する．提案システムの処理の流れを図 1 に示す．図に示すように，提案システムは大きく次の 4 つの要素から構成される．

- 前処理系（blog 記事取得，データベース構築）
- 関連語抽出・統合モジュール
- 類似ブロガー探索モジュール
- 評判分析モジュール

提案システムでは，まず前処理としてクローラーを用いて blog 記事を取得し，名詞データベース（以下，名詞 DB），および評判情報データベース（以下，評判情報 DB）を構築する．

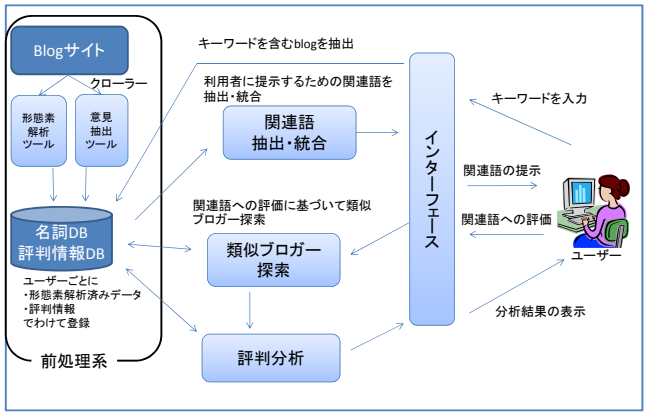

図 1　提案システムの流れ

表 1　抽出ツールの出力形式

ファイル名	文 ID	評価保持者	評価タイプ	評価表現
sample.txt	1	[著者]	メリット＋	ビタミンが豊富だ.

実際の利用時には，利用者がキーワードを入力すると，関連語抽出・統合モジュールがデータベースからキーワードを含む blog 記事を抽出し，さらにそのキーワードに対する関連語を抽出する．次に，抽出した関連語を利用者に提示し，評価付けを行ってもらい，類似ブロガー探索モジュールで関連語に対して類似した評価を blog 記事中で記述しているブロガーを抽出する．そして，評判分析モジュールが抽出されたブロガーを対象とした評判分析を行い，その結果を利用者に提示する．以降の節では，提案システムの個々の要素について説明する．

2.1　評判分析モジュール

提案システムにおける評判分析モジュールでは，独立行政法人情報通信研究機構 NICT により公開されている意見（評価表現）抽出ツール [8] を用いている．本ツールは，1 行につき 1 つの文が書かれたテキストファイルを入力として，テキスト中のそれぞれの文に評判情報が含まれているかどうかを判定し，その文に評判情報が含まれていた場合，以下の情報を出力する．

（1）　評価情報を表す表現（評価表現抽出）
（2）　評価情報の意味的な分類（評価タイプ分類）
（3）　評価情報の評価極性（評価極性判定）
（4）　評価情報を発信する主体（評価保持者抽出）

上記（3）の評価極性とは，対象への評判の記述が肯定的（ポジティブ）であるか否定的（ネガティブ）であるかを表す．この抽出ツールでは，評判情報の抽出および評価極性の分類には条件付き確率場を利用した手法を用いている [11, 12]．表 1 は評判情報を含む例文“ ほうれん草はビタミンが豊富だ ”を入力としたときの意見抽出ツールの出力例を示したものである．各項目はタブ区切りで出力される．評価極性は，評価タイプの後ろに，ポジティブならば“ ＋ ”，ネガティブならば“ － ”を付けることにより表される．

2.2　前 処 理 系

提案システムでは，前処理として blog 記事をあらかじめクローラーを用いて自動取得し，テキストの整形等の処理を行った後にデータベースに保存する．データベースは，利用目的別に名詞 DB と評判情報 DB の二つを構築する．クローラーは，ブロガーの ID をもとに指定した URL ディレクトリ以下にあるコンテンツを探索して HTML テキストを取得し，HTML タグを手がかりに blog 記事本文のみを抽出する．

前処理系では，さらに blog 記事の整形を行う．HTML タグや，記号の羅列があると解析に悪影響があるため，この段階で取り除いておく．顔文字に関しては，ブロガーの評価を表す場合もあるが，評価表現を伴わずに単独で用いられることは少ないこと，および文末の顔文字が意見抽出ツールの誤動作の一因となる場合があることから，ここでは同様に前処理にて取り除いた．また，blog 記事には文中に不必要な改行が入ることが多いため，改行も取り除いておく．

名詞 DB は，関連語抽出・統合モジュール，類似ブロガー探索モジュールで利用する．前処理後の blog 記事を形態素解析器にかけ，評価対象となり得る語のみを格納する．形態素解析器には，前述の NICT の意見抽出ツール内でも用いられている JUMAN [9] を用いている．なお，ここでいう評価対象となり得る語とは，JUMAN 品詞体系のうち，品詞細分類の普通名詞・固有名詞・組織名・地名・人名の 5 つとする．

一方，評判情報 DB は，類似ブロガーを対象とした評判分析を行う際に利用する．前処理後の blog 記事を NICT の意見抽出ツールにかけて評判情報を抽出し，表 2 のように「評価対象，評価極性，評価表現」の形式で格納する．なお，意見抽出ツールでは評価対象までは出力されないため，日本語構文解析システム KNP [10] を用いて評価表現との係り受けを求め，評価対象を特定する．

表 2　評判情報 DB の格納方法

評価対象	評価極性	評価表現
商品 A	＋	デザインが好き
商品 B	＋	画質が良い
商品 C	－	すぐに壊れた

2.3　検索キーワードの関連語の主成分分析による統合

提案システムでは，利用者の行動履歴などから事前に嗜好を推定するような処理は行わず，利用者が検索キーワードの関連語に直接評価をつけるという形でその嗜好を特定する．これは，ジャンルを絞らずに利用者とブロガーの類似性を推定する場合よりも，検索キーワードに関連したものに対する嗜好が類似したブロガーによる評価のほうが，より参考になるという仮定に基づいている．また，ここでいう関連語とは，本文中にキーワードを含む blog 記事に頻繁に出現する単語と定義する．名詞 DB からキーワードを含む blog 記事を抜き出して文書・単語行列を作成し，TF-IDF 値の高いものを関連語とする．

さらに本研究では，関連語をそのまま利用者に提示せず，利用者への負担軽減と類似ブロガー数確保のため，主成分分析により複数の関連語を統合する．具体的には，対象とする blog 記事数を m としたとき，各関連語を blog 記事ごとの TF-IDF 値により構成される m 次元ベクトルで表現し，そのベクトルの次元を主成分分析により $n(< m)$ 次元に縮約する．そして，縮

約された次元の空間上で単語ごとのユークリッド距離を求め，検索キーワードから一定範囲内にある関連語を k 個のクラスタに分割する．利用者には，各クラスタの代表語を提示し，各語に対してポジティブなイメージを持っているか，ネガティブなイメージを持っているかの評価付けを行ってもらう．利用者に提示する代表語は，クラスタ内の単語で最も多く評判情報が述べられている単語とする．

2.4 類似ブロガーの探索

類似ブロガーの探索は，システム利用者による関連語への評価付けの後に行う．ここでの類似ブロガーとは，関連語に対する評価（極性）がシステム利用者と同一であるブロガーとする．具体的には，評判情報 DB から利用者が関連語に対して付けた評価極性と同一の極性の評価を対応する関連語について記述しているブロガーを抽出し，類似ブロガーとする．ブロガーが関連語に対して複数回評判情報を記述している場合，ポジティブ/ネガティブの意見数の多い方を評価極性として採用する．類似ブロガーを探索する際には，検索キーワードと距離の近い関連語に対する評価を述べているブロガーから探索し，分析を行うために十分なブロガー数が確保できない場合には，徐々に距離の遠い関連語に対する評価を記述しているブロガーも探索範囲に含める．具体的には，検索キーワードから距離の近い順に関連語（w_1，w_2，w_3）があった場合，まず w_1 について評判を記述しているブロガーを探索し，ブロガー数が最低人数に満たない場合，順次 w_2，w_3 についての評判を記述しているブロガーを探索していく．

3. 評価実験

実際のブログ記事を利用し，提案システムにより得られる利用者と類似したブロガーを対象とした評判分析の結果が，対象ブロガーを限定しないで評判を分析した結果とどの程度異なるのかを確認した．実験データとして，ブログサイト Ameba [13] から事前に収集した blog 記事を使用した．具体的には，ジャンル別ランキングの中から，カテゴリ「高校生，大学生，アラサー，アラフォー，アラフィフ，お酒・ワイン，ラーメン，グルメ，モグモグ」のランキング上位者から計 1,520 名のブロガーの blog 記事を収集し実験データとした．全ブロガーの総記事数は 1,423,492 となった．実験で対象としたカテゴリは，各年代から均等にブロガーを集めるために年代別カテゴリを対象とし，さらに実験に用いるキーワードへの評価が集まりやすいと推測される食関係のカテゴリも採用した．今回は評判分析を行う検索キーワードとして「キムチ」「ココア」の 2 つを用いた．利用者に提示する関連語のクラスタ数は 2 とし，今回は実験のため，関連語クラスタへの評価付けはポジティブ（P），ネガティブ（N）の全組み合わせ 4 パターン（P–P，P–N，N–P，N–N）とした．

まず，キーワード「キムチ」で検索した結果，キーワードを含む blog 記事は 507 件存在した．実際には 1 人のブロガーが複数の記事を書いており，最終的に「キムチ」に対してポジティブまたはネガティブという評価を記事中に記述していたことが判別できたブロガーの人数は 129 名となった．「キムチ」に対する関連語として抽出されたものは「豚，豆腐，鍋，チーズ，野菜，ニラ」となった．クラスタリングの結果，クラスタ 1（豚，豆腐：代表語「豚」），クラスタ 2（野菜，鍋，チーズ，ニラ：代表語「野菜」）が得られた．図 2 に「キムチ」に対する類似ブロガーに限定した場合と全体での評判の分布を示す．図の上段 4 つが提案システムによる評判の分布であり，下段が全体での評判となる．全体評判における P/N 割合は，ポジティブ 69 人，ネガティブ 37 人，中立 23 人となり，好評意見の割合は約 53%となった．これに対し，提案手法でクラスタ 1，クラスタ 2 に対してポジティブな評価を記述していたブロガー集団（P–P）の P/N 割合は，ポジティブ 44 人，ネガティブ 14 人，中立 11 人となり，好評意見の割合が約 64%となり，好評意見の割合が増加した．これより，クラスタ 1，クラスタ 2 の関連語についてポジティブなイメージを持っている人は，キムチも好きである可能性が高いと考えられる．また，その他の評価パターンのブロガー集団は非常に人数が少なくなった．これは，各クラスタに含まれる関連語そのものに対しての評判のうち，好評意見が占める割合が高かったためだと考えられる．

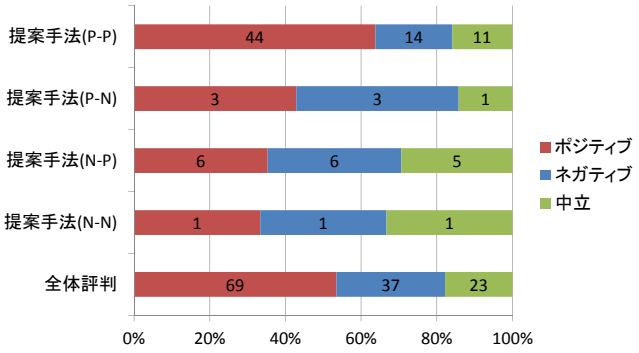

図 2 「キムチ」分析結果の P/N 割合

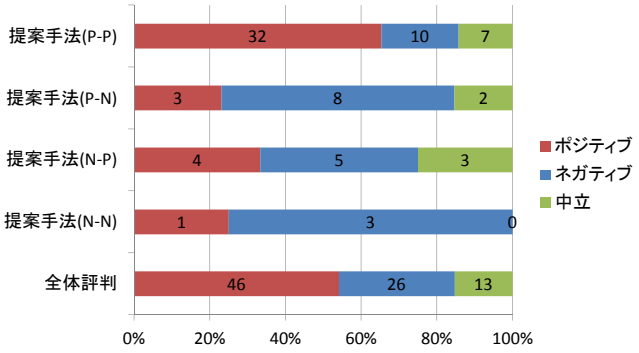

図 3 「ココア」分析結果の P/N 割合

次に，キーワード「ココア」で検索した結果では，キーワードを含む blog 記事は 361 件存在し，実際に「ココア」に対する評価を記事中に記述していることが判別できたブロガーは 85 人であった．「ココア」に対する関連語として，「ミルク，ホット，生クリーム，マシュマロ，ケーキ，紅茶」が得られ，クラスタリングの結果，クラスタ 1（紅茶，ミルク，ホット，マシュマロ：代表語「紅茶」），クラスタ 2（生クリーム，ケーキ：代表語

「生クリーム」）が得られた．ココアについても，クラスタ 1，クラスタ 2 へのポジティブ–ネガティブの全評価パターンおよび全体評判の分析結果を図 3 に示す．全体評判における P/N 割合は，ポジティブ 46 人，ネガティブ 26 人，中立 13 人となり，好評意見の割合は約 54%となった．これに対し，提案手法でクラスタ 1，クラスタ 2 に対してポジティブな評価を記述していたブロガー集団（P–P）の P/N 割合は，ポジティブ 32 人，ネガティブ 10 人，中立 7 人となり，好評意見の割合が約 65 %となり，好評意見の割合が増加した．一方，否定評価を含むその他の評価パターン（P–N，N–P，N–N）では，好評意見の割合の減少がみられた．特に，クラスタ 1 にポジティブ，クラスタ 2 にネガティブな評価を記述していたブロガー集団（P–N）では，肯定の割合が約 23% となり，全体評判と比べて大きく減少した．このことから,「乳製品は好きだが甘いものが嫌いな人は，ココアが好きではない可能性が高い」という解釈ができる．

以上の結果から，提案システムを用いた類似ブロガーに着目した評判分析により，全体評判ではわからない特定の集団における特徴的な評価傾向が得られることが確認できた．一方で，以下に挙げるような問題点も見つかった．まず，分析対象となるブロガー数が十分に集まらなかったことが挙げられる．提案手法では類似ブロガー集団に限定して分析を行うため，通常よりも分析対象となるデータが少なくなってしまう．この問題は事前により多くのブロガーの記事データを収集しておくことで，ある程度緩和されると考えられる．次に，関連語の抽出精度の問題がある．今回,「ココア」に対する関連語として「ホット」が抽出された．これは関連語としては正しいかもしれないが，評価対象としては解釈しづらい単語である．今後はより精度の高い関連語の抽出方法の検討が必要である．

4. ま と め

本稿では，評判分析を行う際に利用者にとってより有益となる情報を得るために，利用者と趣味や嗜好の類似したブロガーの記事に着目した手法を提案した．提案システムでは，類似ブロガーを探すために，評判情報を知りたい対象に関連したキーワードに対する評価傾向を手掛かりとしている．このアプローチにより，利用者は自分と価値観が似た人たちからなる集団の中での評判情報を知ることが可能になる．本論文では，評価実験を通してブロガーの関連語への評価傾向が検索キーワードに対しての評価と関連があることを確認した．

今後，提案システムの有用性を検証するために，類似ブロガー集団における評判の分布が実際に利用者の嗜好と一致する傾向があるかどうかの評価を行う必要がある．また，関連キーワードを抽出する方法として，ここでは単語の TF-IDF 値を基に主成分分析，クラスタリングを行ったが，より適切な関連キーワードを抽出するためには，ほかの手法の採用も含めてさらに議論が必要だと思われる．

謝　　辞

意見（評価表現）抽出ツールの利用を許諾いただいた独立行政法人情報通信研究機構に感謝する．

文　　献

[1] 稲葉真純, 長野伸一, 長健太, 溝口裕美子, 川村隆浩: CGM 分析技術の現状と課題, 人工知能学会研究資料, SIG-SWO-A603-06, pp.06-1–06-8 (2007).
[2] NTT レゾナント: goo ブログ記事評判分析, http://blog-hyoban.goo.ne.jp/ (2011).
[3] 株式会社きざしカンパニー: kizasi.jp, http://kizasi.jp/ (2011).
[4] 那須川哲哉，金山博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会研究報告，自然言語処理研究会報告, Vol.2004, No.73, pp.109–116 (2004).
[5] 鈴木泰裕，高村大也，奥村学: Weblog を対象とした評価表現抽出, 人工知能学会研究会資料, SIG-SW&ONT-A401-02, pp.02-1–02-10 (2004).
[6] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell: Text classification from labeled and unlabeled documents using EM, Machine Learning, Vol. 39, No. 2/3, pp. 103–134 (2000).
[7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: テキストマイニングによる評価表現の収集, 情報処理学会研究報告, NL154-12, pp.77–84 (2003).
[8] 独立行政法人情報通信研究機構情報分析研究室: 意見（評価表現）抽出ツール, http://alaginrc.nict.go.jp/opinion/ (2011).
[9] 京都大学 大学院情報学研究科 知能情報学専攻 知能メディア講座 言語メディア分野: 日本語形態素解析システム JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/ (2011).
[10] 京都大学 大学院情報学研究科 知能情報学専攻 知能メディア講座 言語メディア分野: 日本語構文解析システム KNP, http://nlp.ist.i.kyoto-u.ac.jp/ (2011).
[11] Tetsuji Nakagawa, Takuya Kawada, Kentaro Inui, Sadao Kurohashi: Extracting Subjective and Objective Evaluative Expressions from the Web, In Proceedings of the Second International Symposium on Universal Communication, pp.251-258 (2008).
[12] 中川哲治, 乾健太郎, 黒橋禎夫: 隠れ変数を持つ条件付き確率場による依存構造木の評価極性分類, 情報処理学会研究報告，自然言語処理研究会報告, 2009-NL-192(10), pp.1–7 (2009).
[13] Cyberagent: Ameba ブログ, http://www.ameblo.jp/ (2011).