Weblogからのタレントに関する好感度情報抽出

大根 千明† 松尾 豊‡ 木戸 冬子† 勝 芳邦†† 石塚 満†
†東京大学大学院情報理工学部系研究科電子情報学専攻 ‡東京大学大学院工学系研究科 †† ヤフー株式会社

1 はじめに

近年，特別な言語を使わなくとも，簡単に自分の意見をインターネット上で公開できる電子掲示板やWeblog，SNSなどのサービスが大変普及している．これらのサービスにより，Web上での意見公開を行う為の敷居が低くなり，以前に比べ，Web上へ公開されている情報量が格段に多くなっている．その為，従来のアンケート等による情報収集に代わって，Web上に公開されている情報の収集・分析のみで意見収集を行う事が現実的になりつつある．上記のようにインターネット上のメディアに各自公開されている「定型化されていない文章の集まりを収集，自然言語解析の手法を使って単語やフレーズに分割し，それらの出現頻度や相関関係を分析して有用な情報を抽出・要約する」といった作業を「テキストマイニング」と呼ぶ．現在，こういった手法はTVやメディアでは話題のキーワードを検索したり今後の注目ワードを探してくるものとして注目されている．本稿では,Weblog上の情報の中で，一般の人に書き込み易い話題であり書き込み数も多い「タレント」を対象として好感度を抽出し，多値評価することを目的とする．またこれが可能になると，対象を電気製品や書籍などに変えて出すことも出来るほか，良い評価の対象に対する広告を自動的に貼ることもできる．本稿の構成は以下のようになっている．第2節で評判評価についての概要を説明し，第3節でWeblogからの情報抽出方法，第4節にまとめを述べる．

2 評判評価

情報抽出は，大量文書・テキストからの有用な情報・知識発掘をするテキストマイニングの一つである．膨大な情報の中からいかに必要な，質の高い情報を選別し，取得することができるかが問題になる．情報抽出の中でも，特に評判・意見情報を抽出することが重要になってきている．

評価・意見情報とは，例えば，ある特定の対象商品にたいして「良い・悪い」「速い・遅い」「簡単・難しい」「好き・嫌い」といった評判に関する肯定的否定的ラベルのついたテキストのことである．評判評価は，良い悪いといった評価表現を抽出すると同時に，この評価が肯定的意見あるいは否定的意見として強いかどうかを「評価値」で判定する[1]．例えば,「Aは良い」「Aは良いのか？」「Aは良いとは思えない」といった表現がある場合，肯定評価を正値，否定評価を負値として，この順で値が小さくなっていく．評判評価は，テキスト中の構文解析による評判評価表現の抽出，評価値の計算，肯定・否定の分類といった一連の処理の流れで行なわれる。

現在，分類に関しては最も単純なP/N分類（Positive/Negative Classification）と呼ばれる肯定・否定のどちらの勘定を含んでいるかを特定する2値分類の研究が最も多い．このPN分類では、「欲しい」「いい」「クール」などを“ポジティブな表現”とし,「欲しくない」「嫌い」「ダメ」などを“ネガティブな表現”として、対象の文章にそれぞれの表現のうちどちらが多く含まれているかを判定するものである[2][3]．しかし，現在知られている手法では,「良い」といったように，それ自体に肯定・否定の感情が込められているものに対しては良好な結果を得られるが,「短い」といったような，評価対象や対象の部分などによって肯定・否定が異なるような表現に対しては精度の高い判定は難しい．これは，例えば「バッテリ駆動時間が短い」という表現はネガティブな評価として,「バッテリ充電時間は短い」という表現はポジティブな評価として判定する必要があるため，それらを機械的に分析することは難しい課題となっている．

3 Weblogからの情報抽出

3.1 イメージを表す語の抽出

現在，評判分類に関しては，対象に対して「良いか，良くないか」を判定する2値分類が主である．しかし，肯定的・否定的であるかは各自が判断するべきことであるため，「いい」や「クール」を全て“ポジティブな表現”とひとまとめすることはできないと考える．そこで，今回は評価にそれぞれ軸を設け，多値分類することを考える．また，実際に対象を評価する際には，今回の分析対象のタレントをとっても，評価軸はかっこいい，面白い，ダサい，などが挙げられるが，この評価対象とは，そもそも対象(今回でいう「タレント」)を指しているわけではなく，対象に付随する名詞を指している．例えば，良い・悪いと単体で使われていれば，タレントの「性格」や「演技」の評価をしており，可愛い・かっこいい等は「容姿」を評価しているこ

Abstraction of Goodwill Information for Performer from Weblog
†Chiaki One ‡Yutaka Matsuo †Huyuko Kido ††Masayoshi Katsu †Mitsuru Ishizuka
†Graduate School of Information Science acd Technology, The University Tokyo
‡School of Engineering, The University Tokyo
††Yahoo Japan Corporation

とになる．よって，対象を「タレント」に絞ることで，評価軸を「性格」,「演技」,「容姿」,「好感」の4つとする．

更に，Weblog上でタレント名がどのような形容詞と共起するかによって，タレントの世間的評価というものが取得できる．タレントが今までに出演したドラマやCM，共演者，不祥事，といったタレントを取り巻く環境を取り出すことでも，タレント自身の世間的評価に関連付けることができる．

3.2 抽出方法

抽出方法としては，いくつかの方法が挙げられる．

- 単純にタレント名と共起する数の多い形容詞を取り，共起する回数で評価優位をつける．
- タレント名の係り受け関係にある形容詞を抽出する．
- 対象名の近場の形容詞を取り出す．

係り受け関係を取る理由として，例えば,「Aはかわいい」というとAは「かわいい」という評価がつくが,「Aがきている服がかわいい」といった場合には，直接Aが「かわいい」と言及している訳ではないからである．タレントのイメージ取得なので，取得する単語は形容詞とする [1].

3.3 別名・同姓同名

タレントなど一般的に世の中に普及しているものには，愛称のような別名が存在する場合が多い．たとえば木村拓哉ならばキムタク，といったようにその別名がブログなどでは一般的に使われることも多くなる．これらが指すものは同一人物であるが，ただ単にタレントの名前をテキスト処理をするのみでは別名で書かれているテキストを取得できない．そこで，Wikipediaを使って，同一人物の一致を図る．

3.4 提案アルゴリズムと結果の例

以下に提案アルゴリズムを述べる．

Yahooの2008年1月16日～25日に収集されたWeblogのデータ10万件を対象に，タレントの評判に関する語を抜き出す．

一つめのデータ抽出アプローチとしては，タレントの氏名をキーワードにして検索をし，対象と共起する単語を抽出する．図1に，名前と共起する回数の多い語をWeb上から抽出した結果と，Wikiで取れる別名，そしてWeblogのデータから関連語を抜き出した結果である．

しかし，この方法ではテキスト内にあるだけで全く関係ない文脈の可能性がある，

そこで二つめのアプローチとして，タレント名の係り受け関係にある形容詞を抽出する．自然言語処理を用いてタレント名とタレント名が係り受ける名詞，その名詞を評価する形容詞の (s_i, n_i, a_i) を3つ1組にし，抽出する．

<大泉洋>

【名前との共起】
ハケンの品格 坊ママ レイトン教授 ゲゲゲの鬼太郎 北海道テレビ 救命病棟24時 オトン 水曜 オカン onちゃん レイトン教授登 本日のスープカレー 悪魔の舞 東京タワー サンサン フジテレビ系 演劇研究会 雑野 劇団 実写映画 テレビ朝日 フジテレビ 小早川伸木の恋 おにぎり ロッキー 所属事務所ねずみ
【Wikiでの別名エンティティ】
洋ちゃん
【ブログデータから抽出してきた結果】
良い ドラマ 好き 上戸彩 情報 小栗旬 坊ママ 映画 視聴率 自分 2008年 ファン 面白い 動画 人気 大泉 多い 女優 テレビ 嵐 CM 武尻エリカ 話題 感じ DVD 水曜どうでしょう 俳優 最後 詳しい SP 高い 最終回 大好き

Fig. 1: 対象の関連抽出結果

それを s_n ごとにまとめ，そこから n_m と a_k を使って評価軸に評価していく．n_i は，日本語係り受け解析機で直接 s_i がかかっているもの，s_i にかかっているものを抽出する．また，文をまたいで係り受け関係にある名詞もとってこられるようにする予定である．a_i は n_i がかかっている形容詞となる．

4 まとめ

今回は，タレントの名前を使って，Weblogからその人の評判やイメージをとってくる手法を説明した．今後は対象の近辺にある言葉に注目して，係り受け解析などを強化していきたい．また，Weblogは日々顔文字や新しい言葉を生み出している．それらに対してのアプローチも考えていきたい．別名に関しては，外間ら [4] の提案する手法として，Web上からフルネームと隣接する文字列を抽出し，対象人物の呼称候補を抽出する方法もあるのでそちらも考慮していきたい．

参考文献

[1] Peter D. Turney: "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.", In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002),pp. 417-424, July 2002.

[2] Bo Pang, Lillian Lee (2004): "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts." In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL), pp. 271-278.

[3] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan: "Thumbs up? sentiment classification using machine learning techniques,", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002),pp.76-86,2002.

[4] 外間 智子, 北川 博之: "Webデータを用いた人物の呼称抽出", DBSJ Letters Vol.5, No.2