2G2-1

Weblogからのタレントに関する好感度情報抽出

Abstraction of Goodwill Information for Performer from Weblog

大根 千明[*1] 松尾 豊[*2] 木戸 冬子[*1] 勝 芳邦[*4] 石塚 満[*1]
Chiaki One Yutaka Matsuo Huyuko Kido Masayoshi Katsu Mitsuru Ishizuka

[*1]東京大学情報理工学系研究科電子情報学専攻
Graduate School of Information Science acd Technology, The University Tokyo

[*2]東京大学大学院工学系研究科
School of Engineering, The University Tokyo

[*3]東京大学情報理工学系研究科
Graduate School of Information Science acd Technology, The University Tokyo

[*4]ヤフー株式会社
Yahoo Japan Corporation

In these days, people can easily disseminate the information including their personal evaluation, opinions for some products and services on the Internet. The massive amount of their information is beneficial for both product companies and users who are planning to purchase and use them. Since the evaluative opinions of the people are mainly presented in textual forms such as comments on bulletin board systems, it is mainly researchers in the research field of natural language processing who devote themselves to developing techniques for exploring, extracting, mining, and aggregating the opinions and sentiments. This sort of techniques are commonly called "sentiment analysis". In this paper, we research personality's reputation by sentiment analysis with Weblog.

1. はじめに

近年、HTML や Perl などといった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、Social Network Service (SNS) などのサービスが大変普及している．これらのサービスが普及することにより，PC 初心者でも 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，多くの人の意見を収集するという目的を達成する手段として，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．上記のような一般の人々がインターネット上のメディアに各自公開している「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．本報告では,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．またタレントの好感度自体を出すことが可能になると，対象を電気製品や書籍などに変えて出すことも可能なほか，各ページで述べられている，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述べ、第 3 節で Weblog からの情報抽出方法を説明する。

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．このような評判情報抽出技術は，宣伝とは異なる消費者の生の声として，マーケティング支援や商品購入支援として有効であると期待されている。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。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．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．

連絡先: 大根千明，東京大学情報理工学部，chiaki@mi.ci.i.u-tokyo.ac.jp

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，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，というように多値にわたる．ただし，今まで「良い・良くない」の 1 次元であった評価軸を，多値に増やすことによって何軸にするかは，検討するべき問題である（例えばかっこいいの対極はかっこ悪いと取るべきか，その中にダサいという形容詞を含めていいものか）．

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．また，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る．
- タレント名と評価語となりうる形容詞が共起する回数で評価優位をつける
- 対象名の近場の形容詞を取り出す．
- タレント名の係り受け関係にある形容詞を抽出する．

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．

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．これらがさ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．Wikipedia 内のプロフィール部分には，生年月日や血液型などがまとめられた表が多くのタレントに存在し，そのタレントに別名がある場合には.「別名」という項目もある．これを利用して，別名を再び検索にかけることで，対象に対する関連語の幅が，より広がる．

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．

タレントの名前に対して，電通バズリサーチの収集された Weblog のデータを使って，評判に関する語を抜き出す．電通バズリサーチとは消費者によるネット上の書き込みをモニタリング・分析するシステムであり，特有の消費者の本音や生の声をリアルタイムに把握が可能で，マーケティング活動に利用することが出来る．この API サービスは，ブログ検索システムにおける各種サービスへの問い合わせを行い，その結果を XML 文書として出力するものである．データ抽出方法は，wget で api アドレスを呼び出し，キーワード部分に対象名を utf-8 で代入する．タレントの氏名をキーワードにして Weblog 検索エンジンが XML を出力するので，対象を形容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．

＜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　レイトン教授役　本日のスープカレー　悪魔の箱　東京タワー　サンサン　フジテレビ系　演劇研究会　嬉野　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　沢尻エリカ　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

図 1: 対象の関連抽出結果

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くとともに,「A を昨日見た。超かっこいい」といったような，文をまたいでいる評価に対しても対象自身のこととしてとってこれるようにしたい．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．

参考文献

[1] 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, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02, 2004.

[2] 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.

[3] 藤村滋, 豊田正史, 喜連川優, “電子掲示板からの評価表現および評判情報の抽出”, 第 18 回人工知能学会全国大会, 3F1-03, 2004.6

[4] 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.

[5] 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.

[6] Anindya Ghose, Panagiotis G. Ipeirotis and Arun Sundararajan: Opinion Mining using Econometrics: "A Case Study on Reputation Systems", Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007), 2007.

[7] Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten de Rijke.: "Using WordNet to Measure Semantic Orientations of Adjectives.", 4th International Conference on Language Resources and Evaluation (LREC), 2004.

[8] Osgood, C. E., Suci, G. J., and Tannen-. baum, P. H.: "The Measurement of Meaning,", University. of Illinois Press (1957)

[9] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi (NEC), and Toshikazu Fukushima (NEC); "Collecting Evaluative Expressions for Opinion Extraction", In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04), pp. 584-589

[10] .F. Sebastiani. 2002. "Machine Learning Automated Text Categorization.", ACM Computing Surves, Vol.34 No.1, pp.1-47.

[11] EE, Vasileios Hatzivassiloglou, Janyce Wiebe: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. COLING 2000: 299-305.