言語資源と分類タグ付与技術を利用したWebテキスト分析システムの開発

金丸 敏幸 *1　村田 真樹 *1　井佐原 均 *1
独立行政法人 情報通信研究機構 *1
({kanamaru, murata, isahara}@nict.go.jp)

1　はじめに

本稿では，われわれが開発した，Web 上の利用者生成データを分類，分析する二つのシステムについて紹介する．

近年，インターネット上では，ユーザが意見を述べることによって内容を拡充していく，ブログや掲示板，SNS といった，コンシューマ・ジェネレイテッド・メディア (Consumer Generated Media: CGM) が注目を集めている．これら CGM による記事や文書は，日々，膨大な量が生産されており，例え一部であっても，人手によって全体の内容を把握することは困難となっている．

言語処理の分野では，CGM の文書内容を解析し，どのような動向があるかといった傾向を分析する研究や，評判情報の抽出や分類に関する研究が多く行われている (乾・奥村 2006)．評判情報に関する研究では，特定の製品やサービスなどへの意見や評価などを，肯定的 (positive) か否定的 (negative) かといった評価極性によって分類するものが多く見られる (Pang et al. 2002; Turney 2002)．

確かに，その製品が肯定的なものであるかどうかは，製品の購入などを考える際には重要な観点となる．しかし，評価極性だけではユーザの意見を十分に処理できたとは言えない．例えば，同じ否定的評価であっても，その評価の生じた原因が，製品に対する怒りから発生したものか，失望から発生したものかによって，対処する側の取るべき行動は異なるものとなる．つまり，評価極性による分類だけでは，意見情報の分析としては不十分な可能性があり，より詳細な意見の内容や性質に関する処理が求められる．

そこで，われわれは，意見分類タグを付与したコーパスを利用して，書き手の評価や意見に関する種類と内容を自動分類する意見情報分類システムを構築した．

本システムでは，意見分類タグを付与したコーパスを学習データとして，対象となる文書に類似した性質を持つ文書を検索し，検索された文書に多く共通する意見の種類や内容を対象文書の意見の種類や内容として出力する．これにより，意見文書をさまざまな観点から分類し，分析することが可能となった．

CGM の発展は，それまで情報の受け手であったユーザに情報発信の機会を提供することで，利用者にとって多くのメリットをもたらしたが，その一方で，すでに多くの人により指摘されているように，インターネット上におけるユーザの口コミの力は，負の側面でもますます無視できないものになってきている．

その一つにインターネット上における風評被害の発生が挙げられる．風評被害とは，災害や事故，事件などの虚偽や不十分な情報 (風評) により，消費の減退や選択の回避が行われることによって，金銭的な被害を受けたり，ブランドイメージに悪影響を受けることをいう．CGM が発展するまでは，風評の原因は，主にマスコミによるものであった．しかし，CGM の発展に伴って，インターネット上の個人により，風評被害が発生することも珍しくなくなっている．

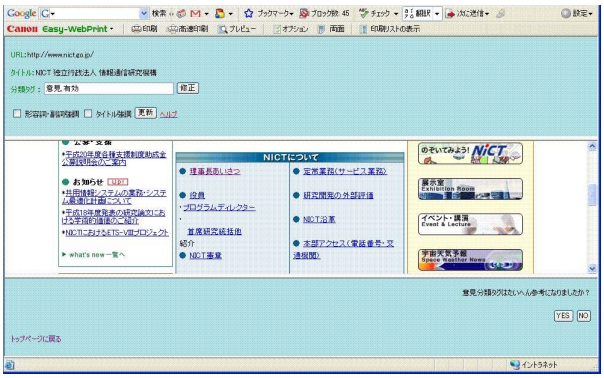
そこで，われわれは，Web 上の文章に対し，それが災害や事故，事件などの虚偽の報道等により生じた情報 (風評情報) に基づくものかを，機械学習を利用して判断する風評情報分析システムを構築した．対象となるページが風評情報に基づくものであるかどうかを判定し，風評情報に基づくものである場合には，風評被害の対象への悪意がある情報なのか，単なる誤解に基づく情報なのかを，評価付き副詞・形容詞辞書を利用して判断する．

2　意見情報分類システム

われわれは，意見情報分類システムの評価用に，Web 上で利用できるインターフェイスを構築した．ユーザが，分析を行いたいページの URL を指定すると，システムはそのページを取得し，分類タグとともに対象ページを出力する．

分析対象のページを取得したシステムは，対象ページから文書を抽出し，意見分類タグ付きコーパスから類似文書を検索するためのクエリを生成する．生成されたクエリを用いて，類似する文書の検索を行い，類似度による文書のランク付けを行う．

図 1　意見情報分類システムの出力



ランク付けされた文書に付与されたタグに対し，各タグのスコア計算を行い，閾値以上のスコアを得たタグを，対象ページの意見情報の分類タグとして出力する．分類タグ付け付与システムについては，2.2 で説明する．

システムが出力した分類タグが間違っていた場合，そのタグを人手で修正することができる．これらの修正データは，データベースに保存される．われわれは，この修正データをシステムの分類精度を向上させるために利用したり，新たなタグ付きコーパスとして利用したりする予定である．

また，出力ページでは，種々の強調表示 (村田ほか 2007) を組み込んだ．そのため，意見分類だけでなく，評価付き形容詞・副詞の強調表示やタイトル分の強調表示を行って，対象ページの詳細な分析を行うこともできるようになっている．出力ページのイメージを図 1 に示す．

2.1　意見分類タグ付きコーパス

意見情報分類システムが，意見分類のタグを付与するための学習データとして，国立情報学研究所が提供している「Yahoo! 知恵袋」(ヤフー株式会社 2007) を利用した．「Yahoo! 知恵袋」データから，「家電，AV 機器」と「コスメ，美容」に関する質問を選び，それらの質問に対するベストアンサーの中から，ランダムで抽出した文書を意見文書ファイルとした．

抽出の対象とした文書は，質問がなされた期間が，回答数の安定し始めた 2004 年 11 月から 2005 年 10 月までの 1 年間に行われたものとした．また，対象とした質問に対して，質問に対する回答数が全体の中間値である 3 よりも多かったものという制限を設けた．この制限は，回答数が少ない場合，回答の質が低くてもベストアンサーとなってしまうため，そのような文書を取り除くために行った．さらに，抽出したベストアンサーは，ある程度の長さを持った文書とするため，7 文以上あるものに限定した．

図 2　意見分類タグ付与対象のサンプル

11789147　3029370　2005-02-17 15:11:17
メイクでは、一部のしみ等なら、ベースカバーやコンシーラーで、簡単に隠す事はできます。しかし、赤面症を隠すのにはどうかと思います。むしろ、大した事ではない、恥ずかしく思う事ではない…という、ポジティブな発想で対処する方法が、一番なのではないでしょか。緊張した時や興奮した時に顔が赤くなる事は、誰にでもありえることですょ。「顔、赤いよ」って言われる事はあっても、周りの人に、悪影響を及ぼす事は無いはずです。自分の体質の一つくらいに考えてみるのは、いかがでしょう。

これによって抽出された文書の数は，二つのカテゴリーを合計して 388 文書であった．抽出した文書の例を図 2 に示す．文書には，管理のための情報があらかじめ付与されている．それぞれ，タブで区切られており，最初の数字列が「回答番号」，二番目が「質問番号」，三つ目が「回答の日付」である．分析，分類の対象となる文書は四番目に記述されている．

われわれは，抽出した文書に意見分類タグを付与して，意見分類タグ付きコーパスを作成した．

意見分類タグは，文章の種類 (メタ)，書き手の態度，書き手の立場 (視点)，談話構造，書き手の観点，モダリティの 6 種類に分類されており，初期のタグとして全部で 58 種類用意した．以下に，用意したタグを挙げる．

- メタ：意見，事実，伝聞，引用，感想，疑問
- 態度：肯定，否定，快，不快，喜び，怒り，哀しみ，楽しみ，驚き，嫌悪，恐れ，希望，落胆，苦痛，羨望，諦め，共感，不満，侮蔑，安堵，励まし，困惑，不安，感嘆，焦り，恥，恐縮
- 視点：当事者，関係者，外部
- 談話構造：原因，条件，目的，方法，逆接，並列，例示
- 観点：主観，非主観，意志，判断，当為，疑問，婉曲
- モダリティ：意志，申し出，勧誘，命令，禁止，依頼，願望

一つの文書に対して，複数の分類タグが付与された．また，一つの文書に付与する意見分類タグの上限は特

に設けなかった．

一つの文書内において内容が大きく異なる部分があり，複数の段落に分割した方がよいと判断された場合は，その文書を分割し，独立の文書として扱った．この場合，分類タグは，分割した部分ごとに独立に付与した．

2.2 意見分類タグの付与

意見分類システムでは，k 近傍法を用いて，意見分類用のタグを付与している．あらかじめ意見分類用のタグが付与されたコーパスから類似する文書を検索し，それらの文書が持つタグのスコアを計算し，閾値以上のタグを出力する．類似文書の検索には，ruby-ir (内山 2005) を使用した．類似文書の検索における類似文書の重み付けには，SMART 重み付け法 (Singhal et al. 1996; Singhal 1997) を用いた．

また，出力するタグを決定する際には，Murata et al. (2007) が用いた k 近傍法の変形式を用いた．この方法は，類似する k 個の文書のスコアをタグごとに加算する方法である．類似する k 個の文書が持つそれぞれのタグに，その文書のスコアの分だけ加算し，タグごとのスコアを計算する．閾値以上のスコアを得たタグがシステムから出力される．

このようにして得られたタグを，対象の文書が持つシステムが分類した意見タグとした．

3 風評情報分析システム

われわれは，風評情報分類システムの評価用のために，Web 上で利用できるインターフェイスを構築した．ユーザが，分析を行いたいページの URL を指定すると，システムはそのページを取得し，対象ページが風評情報に基づくものであるかどうかを判定し，対象ページとともに出力する．

ユーザは，システムの判定が間違っていた場合，その判定を修正することができる．これらの修正データは，データベースに保存される．われわれは，この修正データを今後の精度向上のために利用する予定である．

また，出力ページでも，意見情報分類システムと同様，種々の強調表示を組み込んだ．

3.1 風評情報分析用タグ付きコーパス

われわれは，風評情報分析システムの学習用データとして，風評に関する情報のタグを付与したコーパスを作成した．

国立情報学研究所から提供された「Yahoo! 知恵袋」データの 2004 年 04 月から 2005 年 10 月までのベストアンサーの中から，風評に関係の深いと思われる文書を抽出するために，「聞いた話」という文字列を含む回答を抽出した．

また，ある程度の長さを持った文書を対象とするため，抽出したベストアンサーは，7 文以上あるものに限定した．その結果，抽出された文書の数は，620 文書であった．

これらの文書に人手で，風評かどうかの情報と風評の種類のタグを付与していった．風評の種類のタグは，以下の基準に従った．

- 風評情報ではなく、単なる意見であると判断したもの .....意見
- 風評情報でも、意見でもなく、事実が書かれていると判断したもの .....事実
- 風評被害の対象への悪意によって書かれたと判断したもの .....悪意
- 風評被害の対象への悪意がなく、不確定な伝聞や誤解によると判断したもの .....誤解
- 以上のどれでもないもの .....その他

まず，各文書に対し，これらの文書が意見情報か，風評情報かの判断を行った．風評情報と判断した文章は，その文章が風評被害の対象への悪意から書かれたのか，誤解に基づいて書かれたのかどうか，を判断した．ひとつの文章の中に，悪意と好意のような相反する意見が混在する場合でも，悪意がある場合には「悪意」のタグを付与することにした．

付与したデータを機械学習の学習データとして用いた．「悪意」と「誤解」のタグが付与された文書を風評に基づく情報として扱い，それ以外の文書は，風評とは関係ない文章として学習を行った．機械学習には，TinySVM(工藤 2000) を利用し，システムはこの学習結果に基づいて，入力文書を判定し，タグを出力する．

3.2 形容詞・副詞辞書

風評情報分析システムでは，コーパスを用いて機械学習を行い，その学習結果に基づいて，入力された文書が風評かどうかを自動で判別する．風評情報であると判別された文書は，文書の内容が，風評被害の対象に対して，肯定的なものか，否定的なものかを，形容詞・副詞辞書を用いて分類される．

この時用いられる形容詞・副詞辞書は，次のように作成した．まず，ipadic(浅原・松本 2003) と Unidic(伝ほか 2007)，JUMAN(黒橋・河原 2005) の各辞書から

形容詞・副詞を抽出した．次に，それらの語に対し，肯定的に使用されるか，否定的に使用されるかを人手で判別し，タグ付けを行った．タグは，肯定的，否定的の他，どちらでもない，どちらにも使用されるの四種類を使用した．

風評情報分析システムは，文書に使われている形容詞・副詞の肯定度・否定度を辞書から計算し，肯定的であれば，誤解に基づいた文書として，否定的であれば，悪意に基づいた文書として判定する．

4 おわりに

本稿では，われわれが開発した，Web 上の利用者生成データを分類，分析する二つのシステムについて紹介した．

一つは，意見分類タグを付与したコーパスを利用して，書き手の評価や意見に関する種類と内容を自動分類する意見情報分類システムである．本システムでは，意見分類タグを付与したコーパスを学習データとして，対象となる文書に類似した性質を持つ文書を検索し，検索された文書に多く共通する意見の種類や内容を対象文書の意見の種類や内容として出力する k-近傍法を採用した．これにより，意見文書をさまざまな観点から分類し，分析することが可能となった．

もう一つは，Web 上の文章に対し，それが事故や事件などの虚偽の報道等により生じた情報 (風評情報) かを，機械学習を利用して判断する風評情報分析システムである．対象となるページが風評情報であるかどうかを判定し，風評情報である場合には，風評被害の対象への悪意がある情報なのか，単なる誤解に基づく情報なのかを，評価付き副詞・形容詞辞書を利用して判断するものである．

謝辞

本研究は，独立行政法人 科学技術振興機構 (JST) の地域イノベーション創出総合支援事業「シーズ発掘試験」の助成を受けたものです．また，本研究の実施にあたっては，ヤフー株式会社が国立情報学研究所に提供した Yahoo! 知恵袋データを利用しました．ここに記すとともに，データを提供くださいましたヤフー株式会社ならびに，国立情報学研究所に感謝いたします．

参考文献

浅原正幸・松本裕治，2003,『ipadic version 2.7.0 ユーザーズマニュアル』.

伝康晴・山田篤・峯松信明・内元清貴・小磯花絵・小木曽智信，2007,「多様な目的に適した形態素解析システム用電子化辞書の開発」「日本語コーパス」全体会議電子化辞書班報告.

乾孝司・奥村学，2006,「テキストを対象とした評価情報の分析に関する研究動向」『自然言語処理』13(3): 201–42.

工藤 拓，2000, “TinySVM: Support Vector Machines,” http://cl.aist-nara.ac.jp/taku-ku/ software/TinySVM/index.html.

黒橋禎夫・河原大輔，2005,『日本語形態素解析システム JUMAN 使用説明書.version.5.1』．東京大学大学院情報理工学系研究科.

Murata, M., T. Kanamaru, T. Shirado, & H. Isahara, 2007, “Automatic F-term Classification of Japanese Patent Documents Using the k-Nearest Neighborhood Method and the SMART Weighting,” *Journal of Natural Language Processing,* 14(1): 163–90.

村田真樹・金丸敏幸・白土保・馬青・井佐原均，2007,「種々の重要表現強調表示ツールバーの開発」『言語処理学会第１３回年次大会』.

Pang, B., L. Lee, & S. Vaithyanathan, 2002, “Thumbs up?: sentiment classification using machine learning techniques,” *EMNLP ’02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing,* . Association for Computational Linguistics.

Singhal, A., 1997, “AT&T at TREC-6,” *TREC-6,* .

Singhal, A., C. Buckley, & M. Mitra, 1996, “Pivoted Document Length Normalization,” *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* .

Turney, P. D., 2002, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Vol. 40.

内山将夫，2005, “Information Retrieval Module for Ruby,” http://www2.nict.go.jp/x/x161/ members/mutiyama/software.html.

ヤフー株式会社，2007,『Yahoo!知恵袋 研究機関提供用データデータ仕様書，国立情報学研究所 (NII) 提供版 ver1.0』．国立情報学研究所.