Webコミュニティの信頼度判定に関する研究
Research for analyzing reliability of Web community

溝口達也† 池辺正典†
Tatsuya Mizoguchi† Masanori Ikebe†
†文教大学　情報学部
†Faculty of Information and Communication, Bunkyo Univ.

要旨
近年，インターネットでは，掲示板や Weblog，SNS，SBM など，参加者が意見を掲載する情報媒体が多数提供されている．しかし，インターネットの情報は，信頼性に疑問のある情報が含まれるなどの問題点がある．このため，本研究では，特定の情報媒体について，参加者による評価傾向の分析を行うことで，信頼性判定の指標となる情報抽出を行うことを目的とする．

1. はじめに

近年，インターネットでは，掲示板や Weblog，SNS（Social Networking Service），SBM（Social Bookmark）といった利用者が意見を掲載することが可能である CGM(Consumer Generated Media)と呼ばれる情報媒体が普及している．しかし，CGM の情報は，不特定多数の利用者が，情報源の有無にかかわらず情報を発信しているため，信頼性に疑問のある情報が含まれるという問題点がある．そのため，CGM から情報を得るには，その情報の信頼性を判断する必要がある．こうした背景から，様々な手法を用いて，膨大な情報の中から，必要な情報を収集する研究が行われている．また，CGM の特徴としては，掲載情報に対するコメントが付加できる点である．このため，利用者は，信頼性判定の指標として，コメントを取り扱う．このため，CGM の信頼性判定の解析では，コメントの評価が必須であると考えられる．本稿では，CGM の信頼性の解析のために，従来研究で一般的な手法である評価辞書を用いる方法に加え，コメントの情報量から信頼性判定の参考となる情報抽出方法についての分析を行った．

2. 本研究の概要

SBM とは，任意の Web ページに対して，閲覧者がお気に入りの Web ページとして登録を行った上で，自由にコメントを付加することができる仕組みである．このため，コメントを解析することで，Web ページに対する閲覧者の評価傾向を確認することができる．本稿では，SBM から Web ページとコメントの抽出を行い，抽出した情報を分析することで，信頼度判定の指標となる情報の抽出を行う．本稿における信頼度とは，評価の傾向に加え，評価情報の情報量にも着目する．

2.1. 評価辞書を活用した評価情報の取得

評価辞書の活用では，評価辞書の評価数値を用いた評価判定方式[4]が提案されているため，本稿においても同様に，SBM から抽出した Web ページとコメントに対して，評価辞書を用いた判定を行い，評価値として数値化した評価情報の取得を行う．具体的な処理の手順としては，「Web ページに対するコメントの形態素解析」と「形態素解析結果と評価辞書とのマッチングによる評価数値の取得」から構成される．本稿では，形態素解析器として，Chasen を利用し，評価辞書は，東京大学が公開を行っている「Polar Phrase Dictionary」を利用した．また，評価数値の取得時には，形態素解析結果と評価辞書を比較して，マッチングが可能であった単語についての評価数値の総和を取得した．

2.2. 構文解析による情報量の判定

一般的な文章において，主題となっているキーワードに対しての係り受け関係が意味に及ぼす影

響は大きい．このため，本稿における情報量の判定においては，単純な単語数のみではなく，係り受けの階層数も評価対象とする．具体的な処理としては， SBM から抽出したニュース記事及びコメントに対して構文解析を行い，その文節数を取得する．本稿では，構文解析器として CaboCha を利用した．その上で，先の評価辞書による評価情報と文節数を組み合わせた評価値を作成した．

3. 実証実験

本稿での提案内容が信頼性判定に有効な情報であることを検証するために，任意の Web ページに対する SBM を取得し，提案手法による解析を行った．本実験は，任意の Web ページに対するコメント 465 件を判定対象とした．そして，提案手法による解析結果と，目視によるコメントの評価を「-10～10」の範囲で行い，これを正解データとして比較を行い，相関係数の算出を行った．ここでの相関係数の算出には，絶対値を利用している．絶対値を用いた理由としては，肯定的な評価と否定的な評価における文体裁が同様と考えられるためである．

表 1：実験の結果

	取得件数	取得漏れ	誤検出	相関係数
評価辞書のみ	60 件	95 件	0 件	0.571
文節数のみ	155 件	1 件	5 件	0.579
評価辞書＋文節数	155 件	1 件	5 件	0.707

最初に，評価辞書とのマッチングで評価情報が取得可能であったコメントの件数は，60 件であり，正解データの 155 件と比較すると取得率は，38.7%であった．また，これらの 2 値の絶対値における相関係数は 0.571 であり，弱い相関があると考えられる． 次に，情報量の判定として，コメントから文節が取得された件数は 155 件であり，このうち正解データと比較した際の誤検出は 5 件であった．また，正解データの取得漏れは 1 件であり，取得率は，96.1%である．このことから，評価要素となる単語の検出には，文節数が及ぼす影響は高いと考えられる．また，2 値の相関係数は，0.579 であり弱い相関があると考える．さらに，評価辞書と文節数を組み合わせた本提案手法では, 相関係数が 0.707 となり, 他の 2 手法と比較し, 正解データとの相関の高い結果が得られた．

4. まとめ

本稿においては，SBM に付けられたコメントを解析することで，ニュース記事の発信者に対する評価情報や情報量を分析することで信頼度判定のために参考となる情報抽出を行った. 評価情報の取得に情報量を取り入れることによって, 信頼性判定のための新たな観点を提案した．今後の発展としては，肯定表現と否定表現の特定を行いたいと考える．

参考文献

[1] 和多太樹，関降宏，田中省作，廣川佐千夫，単語の出現頻度に着目した病院評判情報の分析，情報処理学会研究報告，Vol.2005 No.50，2005.
[2] 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抽出，情報処理学会研究報告，NL-176, 2006, pp.121-128.
[3] 廣嶋伸章，山田節夫，古瀬蔵，片岡良治，評判検索におけるクエリ依存型の評価極性付与，NL-176, 2006, pp.129-134.
[4] 杉木健二，松原茂樹，意見文からの評判情報抽出に基づく自然言語検索，情報処理学会研究報告，NL-176, 2006, pp.135-141.