グループウェアと
ネットワークサービス 45－12
（２００２．１０．２５）

インターネット上の口コミ解析手法の提案

斉藤典明、菅野元之、佐野直美
ＮＴＴ情報流通プラットフォーム研究所
〒180-8585 東京都武蔵野市緑町 3-9-11

コンシューマの中で広がりつつある”口コミ情報”を解析する手法として、インターネット上の BBS を対象に、単語（名詞・評価語）、発言者、発言時間に着目し、”口コミ情報”が広まる過程を抽出する手法について提案する。また、実際のインターネット上の BBS の記事を用いて実験した結果、提案方式の有効性といくつかの”口コミ情報”の伝播の過程を観測することができたので報告する。

Mining Method for “Word-of-Mouth” Information via Internet

SAITO Noriaki, SUGANO Motoyuki, SANO Naomi
NTT Information Sharing Platform Laboratories
3-9-11 Midori-cho Musashino-shi, TOKYO 180-8585 JAPAN

We developed a mining method of "Word-of-Mouth-Information" which spread among consumer via the Internet. The feature of our method enables to observe process of spreading of "Word-of-Mouth-Information" by analysing based on words, the contribution date and the contribution user included in messages in BBS. And we examined this method by using BBS's messages which were collected from the Internet. In this paper this method and the examination are described.

１.はじめに

インターネットの発展によって、普通の人々が広く雄弁に語れるようになり、個人的なホームページの公開にとどまらずメーリングリストやチャット、掲示板などを使って様々な情報が”口コミ”で伝わってゆくようになった。最近は、このような普通の人たちの間で広がる”口コミ情報”を商品選択に活用する試み、企業における自社製品の評判のチェックに活用する試みが増加している。　インターネット上で収集した情報の中からコンシューマの中で広がりつつある”口コミ情報”を解析する手法として、形態素解析などを用いテキスト文面中の名詞に着目し名詞の出現頻度によって分析する方法がある[1-3]。さらには、文面中の形容詞に着目し、ポジティブ評価なのかネガティブ評価なのかを分

析する方法の検討が進んでいる。そこで、ここでは文面だけではなくさらに発言者に着目し、今は小さな変化だが将来大きく広がるであろう”口コミ情報”を抽出する手法について検討したので報告する。

2.研究の背景

インターネット上からの”口コミ情報”を収集するためには、様々なコミュニケーション媒体とその特徴を考慮する必要がある。

例えば、インターネット上の情報には、個人のWebページのように一方的に見解を述べるものと、掲示板（以下BBS）やメーリングリスト（以下ML）のように参加者同士のコミュニケーションをベースにしているものと大別できる。このうち、多くの人が目にしたWeb上の情報と、ほとんどの人が参照しないWebページ上の情報とでは”口コミ情報”の発信源の効果としては異なる。しかしながら、ほとんどの人が参照しないWebページ上であっても、多くの人が同じようなことを記述していれば、その話題は浸透している、と考えることができる。つまり、”口コミ情報”の媒体として効果していないWebページであっても、その情報がどのくらい浸透したのかを計る指標としては有効となる。ただし、どのようにしてその話題が広がっていったのかはWebページを参照しているだけではわからない。おそらく、インターネット上では掲示板や電子メール、インスタントメッセージ（以下IM）などのように双方向で対話することの可能な媒体や、あるいはインターネット以外の現実世界での情報交換やマスコミの効果によって広がっていったと考えられる。）このことから、Web上に記された話題を抽出することによって”口コミ情報”が広がっている過程を観測することはある程度可能であると考えられる。

このようにインターネット上の”口コミ情報”を測定するには、媒体の特徴を考慮して”口コミ”のどのような効果を測定するのかを考慮する必要がある。そこで本検討では、双方向性のある媒体を対象に、広がってしまった”口コミ情報”を評価するのでなく、口コミ情報が発生し次第に広がり浸透してゆく過程について検討を行う。

表1.”口コミ”チャンネルとその特徴

電子メール系	1対1メール、ML、メルマガ 比較的クローズドな情報交換。
Web系	個人ページ、法人ページ、BBS 公開用に用いられる。
IM系	1対1、チャット 一過性の情報交換。
NetNews系	NNTPによる掲示板 ユーザ層が偏っている(？)。

”口コミ情報”が参加者同士のコミュニケーションをベースに浸透してゆく過程を測定する場合、以前の検討において、MLを対象に参加者間の情報のやり取りで用いられている単語に着目して分析し、一旦出現した単語が次第にその参会者間で広まり常識化してゆく過程を観測した[4]。しかしながら、MLは比較的クローズドなグループ内での情報交換であるが、本研究の目的である”口コミ情報”がインターネット上を広く浸透してゆく過程を測定するには、更なる検討が必要となった。

インターネット上で扱う情報は、意図的に多く流された噂話や、嘘や勘違いの情報、世の中の平均的な感覚とは大きくかけ離れた評価なども多く存在する。そこで情報の信憑性を考慮する必要がある。このような信憑性にかける情報を排除するためには、メッセージのテキスト本文を解析しただけでは判別することはできない。そこで、このような信憑性にかける情報を排除し有益な”口コミ情報”を収集し活用する方法として、これまで ⑴口コミサイトのように良質な口コミ情報を収集する場を設けるという方法と、 ⑵良質な情報源から文章を収集しその単語を解析することによって”口コミ”を活用する方法があった。

しかしながら本研究では、”口コミ情報”としては根拠のない風評なども考慮する必要があると

考える。つまり、このような情報を排除して口コミ情報を測定するのは妥当ではなく、そのような根拠のない情報が”口コミ”として広まってゆくか・淘汰されるのかは不明であり、これらも明らかにする必要があると考えている。

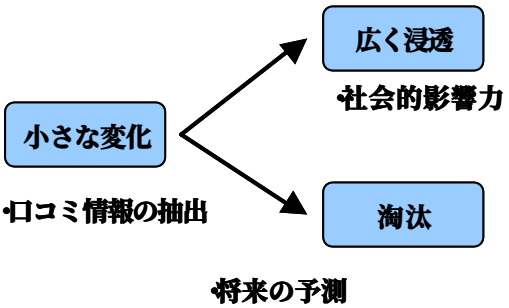

図1.”口コミ情報”の解析の目標

そこで、本検討の対象とする”口コミ情報”の解析では、広く浸透した情報だけでなく、流行または風評として将来大きく発展するであろう（現在の）小さな変化を見つけ出すことを目標とする。

今は小さな話題だが将来大きな話題になるような”口コミ情報”については、広く集めた情報の中から、話題が発生し次第に広まってゆく過程を抽出し、その特徴を明らかにする必要がある。そこで、インターネット上の情報をこれまでの単語単位で評価するだけでなく、情報を流している人物を識別し、その人物の行動を参考に”口コミ情報”の伝播を評価することとした。そのために、まずインターネット上の匿名であっても人物を識別できるようなBBSを対象にテキスト文書の情報を収集し、文章中の単語、投稿者の人物名、投稿日時に着目し、どのように言葉が使いまわされてゆくかについて検討し、これによって”口コミ情報”の伝播を判定する手法を実現した[6]。

3.BBSにおける口コミ情報の伝播

ここでは、インターネット上のBBSを大きく分けて次の3つに分類する。

(1) 単体サイト(2chのような独立タイプ)
(2) レンタルBBS　個々のページは別々だが集合体としては同一のサイト)
(3) 個人CGI　BBSソフトを自前で使用しているタイプ)

本検討では、これら共通に使えるような手法としてクローラを用いて閲覧者として収集し（つまりBBSには手を入れない）、その中に含まれる投稿記事、投稿者の識別情報、投稿日時を抽出して分析することとした。

分析手法として、収集した情報を人物・時間・単語の3つの軸の組み合わせで解析する。人物のみ、時間のみ、単語のみの解析を一次元解析とし、人物×時間、単語×時間、人物×単語の組み合わせの解析を二次元解析、人物×単語×時間の組み合わせの解析を三次元解析とした。

それぞれの解析によって表2のような動向がわかる。特に3次元解析のパターンによって”口コミ情報”がどのように広がってゆくかがわかる。

4.方式確認実験

本方式の有効性を確認するための実験を行った。ここでは、ある単語をキーとしてクローラを使って収集されたBBSの情報を用いて前述の解析を行った。基本データとして表3および図2～図5のような値である。

用いた記事は、インターネット上の掲示板から収集した2002年の1月1日から25週間の日付のついた投稿記事422件であり、その中の投稿者数は272人（一意性が保証された人数ではないため厳密に重複がないわけではない）で

表2.　解析軸と解析項目

一次元解析（基本データ）	
利用者	利用者数
単語	単語出現回数
時間	期間
二次元解析	
利用者×単語（名詞）	ある単語の認知度
利用者×単語（評価語）	誰がどんな評判を流したか
単語×時間	いつどんな単語が流行したか
利用者×時間	誰がいつ発言したか
三次元解析	
利用者×単語×時間	単語の浸透度

表3.　一次解析結果

利用者数	272人
投稿件数	422件
出現単語（名詞）	2653個
出現単語（評価語）	26個
期間	約175日（25週）間

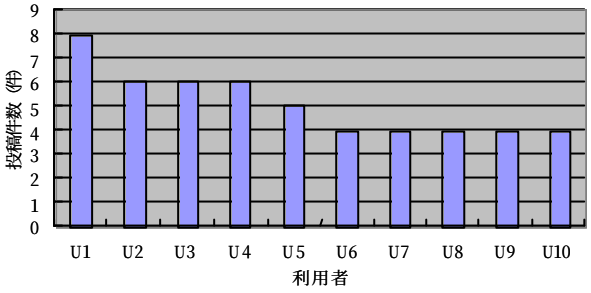

図2.　利用者ごとの投稿件数

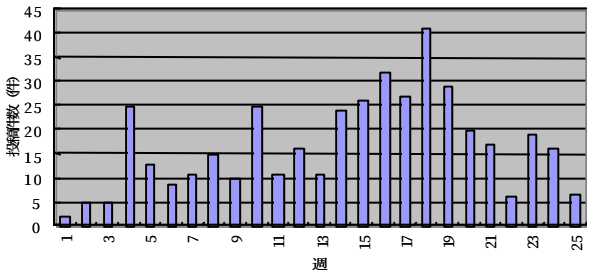

図3.　投稿の分布

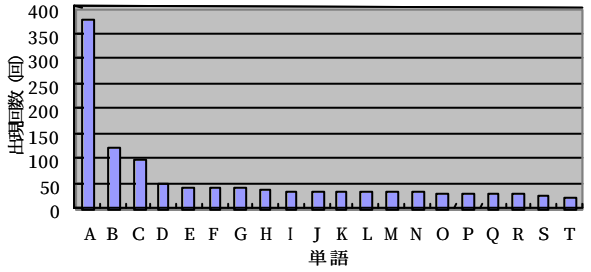

図4.　単語（名詞）ごとの出現頻度

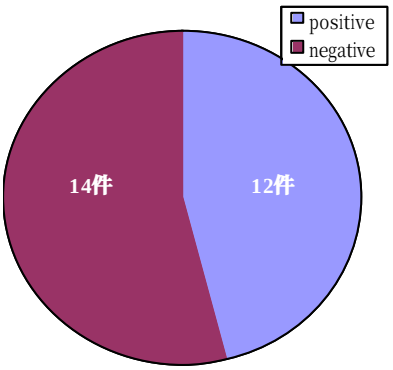

図5.　評価語抽出結果

あった。投稿記事数の多い順の人物リストは図2のとおりである。また、25週間の投稿記事数の分布は図3のとおりである。このような投稿記事に対して、形態素解析（JTAGを使用）を行なった結果、名詞として抽出できた単語が2655個であり、出現頻度の高い順の単語リストは図4のとおりである。さらに、形容詞と言い回しに着目し次のような基本ステップで評価語を抽出した。

(1)　名詞＋格助詞＋（名詞, 形容詞, 動詞）」のパターンを抽出

(2) 評価語辞書（速」、安」など約40語使用）による絞込み

(3) 否定語を考慮してポジティブ / ネガティブ判定

その結果、26語の評価を抽出でき、その内訳は図5のとおりである。なお、評価語の抽出率は、人手によって総数を確認した結果、約50%であった。

このような情報に対して、2次元の解析を行った。その結果は図6～図8のとおりである。

図6は、単語の認知度」である。これは単語の出現頻度のうち同一利用者のものは1回としてカウントしたものであり、単純な出現頻度ではなく何人の人がその単語を口にしたかを知るためのものである。この表のおける「A、B、C・・・」は図4の単語の出現頻度の「A、B、C・・・」と同じである。図4と図6を比較してわかるとおり、出現頻度の高い単語が必ずしも多くの人が口にしているわけではないことがわかる。（図4は「D、E、F」の順に少なくなっているが、図6では「F、D、E」の順になる。）

図7は、単語の流行語」を調べるための図である。流行とは一時的な盛り上がりであり、図7の単語「A」は定常的に出現していることから流行性の単語ではなく定常的な単語であることがわかる。一方「B～F」は一時的な期間に大きな出現頻度の山を観測できた。図7の場合は5月と6月に出現頻度の山があり、この時期になんらかの流行原因を持つ語であったと考えられる。

図8は、「評判」を知るための図である。これは発言者ごとの活動期間のグラフに図5で抽出したポジティブ発言なら「○」、ネガティブ発言なら「×」形式でマーキングしたものである。ここでわかることは、「○」または「×」が連続して現れないことから意図的に評判情報を流そうとしている人はいない、ということがわかる。

さらに、ここでは単語と利用者と時間のデータを組み合わせて3次元の解析を行った。ここでは、どのように単語が浸透していったのか（「浸透度」）がわかる。

ただし、すべての単語については解析を行っていないため、今回の検討で抽出できた代表的な2つのパターンについて説明する。その結果は図9と図10のとおりである。

図9および図10の右側のグラフは着目した単語を含む記事の投稿数の時間軸のデータである。左側のグラフは、着目した単語の出現頻度を発言者数で割ったものを時間軸で並べたものであり、これを話題の浸透度とする。

図9は、記事数は一時期非常に多かったが、話題の浸透度は一定値を示す傾向である。このことから記事数のデータを見ただけでは広く浸透しているように見えるが、人数あたりの浸透度を測定すると実は多くの人には浸透していないことが観測された例である。言い換えると、話題が盛り上がっているように思われるが、実は同じ人たちの間、あるいは普段と同じペースで口コミ情報が広がっていたという例である。）

一方、図10は、投稿記事数も増加した時期があり、浸透度を測定した結果でも多くの人がその単語を口にしたことが観測された。そのため話題が盛り上がった時期に多くの人に浸透していった過程が観測された例である。

5. 考察

以上みてきたように、ここでは、単語、発言者、発言日時を考慮することによって"口コミ情報"の広がりを観測する手法を提案し、実験データを用いて有効性を検証した。その結果、抽出した単語が多くの人に広がってゆく過程を

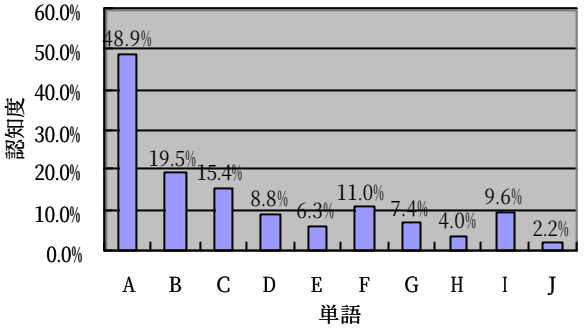

図6. 二次解析結果（認知度）

期間 / 単語	1月	2月	3月	4月	5月	6月
A	←		2～25週			→
B					↔	16～19週
C					↔	17～19週
D				↔	↔	15～18週
E				↔	↔	15～17週
F				↔	↔	16～18週

図7. 二次解析結果（流行語）

期間 / 利用者	1月	2月	3月	4月	5月	6月	7月
U1	←	2～8週（1.4件/週）	→				
U2		×negative		8～21週(0.46件/週)	→		
U3	←		4～24週(0.29件/週)			→	
U4				←	16～25週(0.6件/週)	→	
U5				←	16～24週(0.55件/週)	→	
U6	←	4～17週(0.29件/週)	○positive	→			
U7			↔	9～11週(1.33件/週)			
U8		↔	8～10週(1.33件/週)				
U9				↔	16～19週(1.0件/週)		
U10					↔	18～19週(2.0件/週)	

図8. 二次解析結果（評判）

測定することが可能となった。しかしながら、"口コミ情報"が広く伝播してゆく過程は様々なパターンがあると考えられることから、今後はさらに様々なパターンを抽出し、本方式の有効性について検討する必要がある。

6. まとめ

インターネット上で収集した情報の中からコンシューマの中で広がりつつある"口コミ情報"を

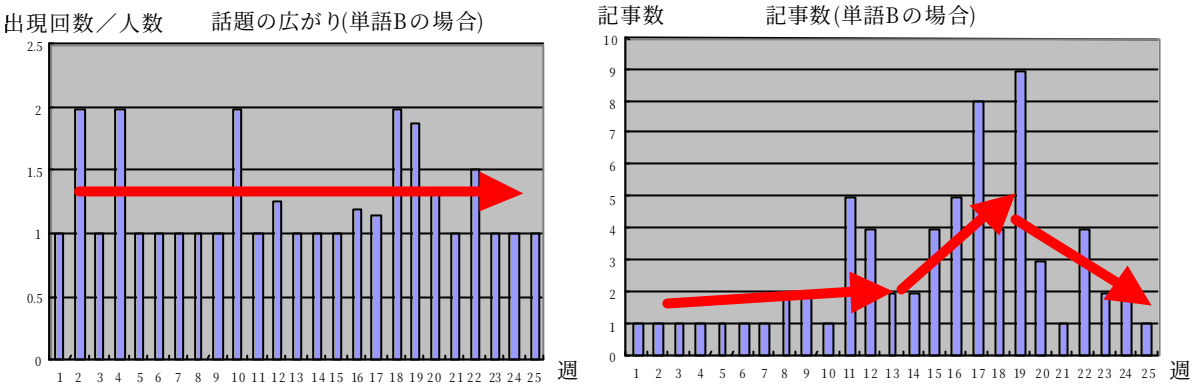

図9. 三次解析結果 (A)

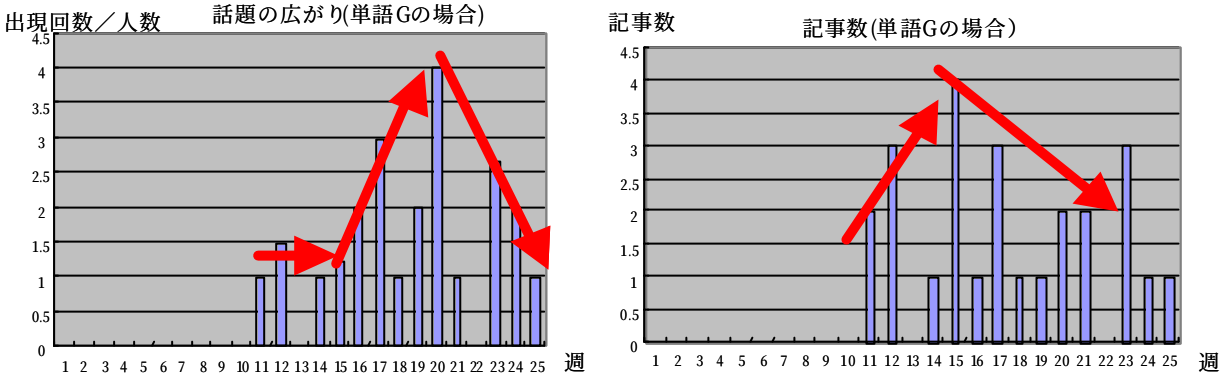

図１０. 三次解析結果 (B)

解析する手法として、BBSの中で流通する記事を対象に、単語（名詞・評価語）、発言者、発言時間に着目し、 今は小さな変化だが将来大きく広がるであろう”口コミ情報”を抽出する手法について提案した。 しかしながら”口コミ情報”はBBSだけでなく、MLやIM、TV、雑誌、電話、対面コミュニケーションなど様々なメディアを介して広がり、また様々なメディア同士の影響も考慮する必要がある。 今後は、 このような複雑な”口コミ情報”を効率的にかつ定量的に捕らえ、様々な”口コミ情報”の伝播の過程を定式化を検討することが重要である。

[参考文献]

[1] 立石ほか, "インターネットからの評判情報検索", 情報処理学会研究会報告, NL-144-11, pp.75-92, 2001.
[2] 二本木ほか, "文の構造化による口コミ評価の分析・検索", インタラクショ2002 論文集, pp.175-176, 2002.
[3] 立石ほか, "Web 上の自動意見分析 - 情報抽出とテキストマイニングの融合-", 情報処理学会第64回全国大会, pp.3-19 - 3-20, 2002.
[4]斉藤ほか,"話題の自動抽出による電子メールの情報組織化手法" 情報処理学会論文誌,Vol39,No.10,pp.2907-2913,Oct.1998.
[5]菅野ほか,"口コミ情報解析による情報遷移の把握に関する一考察",FIT2002 M-39,第四分冊 P.111-112,2002.