M-128

口コミ・マイニングにおける話題の浸透度の測定

The Measurement Method for Infiltration of Topic at "the Viral Mining"

斉藤典明　　浅倉剛　　佐野直美

SAITO Noriaki,　　ASAKURA Takeshi,　　SANO Naomi

１．はじめに

インターネットの発展に伴い誰でも自由に発言することが可能になった。この自由な発言は、広く市民運動として広がり社会を動かすことさえある。企業においてもこのような動向を、潜在ニーズの発掘や風評対策として取り込むことが注目されている。

そこで、インターネット上でどのように話題が伝播してゆくのかを検知する手法として”口コミ・マイニング”を提案した[1]。今回、”口コミ・マイニング”システムを開発し、約４ヶ月間の運用実験を行い、特にその間に出現した新出単語の社会への広がりを計測し、今後の可能性を確認したので報告する。

２．研究の背景

”口コミ・マイニング”の技術は、広くインターネット上から様々な情報を集め、その中から伝播してゆく情報を見つけ出すことを目標としている。

この目標に対し、世の中に”口コミ”で広がる情報がインターネットから取れるのか？という原理的な問題がある。これに対して、ここでは、インターネット上に公開されている掲示板などの文字を介した”口コミ”情報を対象にしている。また、インターネットの普及率が向上するにしたがって、インターネット上でサンプリングし集計した結果が社会全体の指標になりうると考えられる。特に、総務省発表の情報通信白書によれば２００２年の日本のインターネットの家庭への普及率は５０％を超えた[2]。インターネットが社会の調査対象として有効なメディアになっていると考える。

上記目標を達成するために以下の手順で進める。

(Step1)素データの収集
(Step2)情報の伝播の測定
(Step3)伝播する口コミ情報の抽出
(Step4)口コミ情報からのニーズの抽出

３．”口コミ・マイニング”技術の概要

先の４つのステップに関して、今回はStep1とStep2までの検討を行なったので、以下でこれらについて説明する。

（１）素データ収集

素データ収集においては、技術的なポイントは３つあり、「情報源分類」、「内容／ユーザ／時間情報の抽出」、「日々の差分収集」である。

NTT 情報流通プラットフォーム研究所

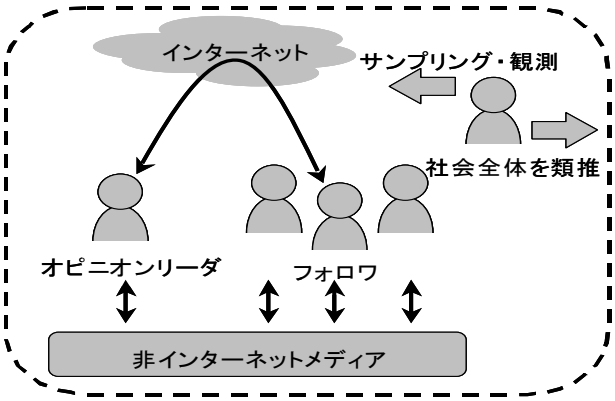

図１．”口コミ・マイニング”の位置付け

・情報源分類手法：インターネットを介したテキスト系”口コミ”情報は、メール、インスタントメッセージ、掲示板、個人ページの中にあると考えられる。ここでは特に公開され誰もが目にできる部分から抽出することを前提とする。また、非常に個人的な”口コミ”情報は対象とせず、社会全体への浸透を対象にした”口コミ”情報を対象とする。そのため、マスコミ報道や企業発表なども考慮し、”口コミ”情報の情報源として、マスコミ・企業系サイト、コンシューマ系ページ（個人ページ、ＢＢＳ）の２つに分類してデータを収集する。収集にあたってはクローラによる方法を前提とする。

・内容／ユーザ／時間情報の抽出手法：”口コミ”情報の社会的な浸透を測定するためには、クローラから収集した素データから、内容情報、ユーザ情報、時間情報を抽出する必要がある。Ｗｅｂページにおいては、このようなメタとなる情報の付与があいまいである。そこで、ここでは２分類した情報源ごとに、情報の特徴にあわせてメタとなる内容／ユーザ／時間情報を抽出する。内容情報の抽出にあたっては形態素解析を用い名詞句のみを抽出した。ユーザ／時間情報の抽出は情報源ごとにＷｅｂページのＵＲＬとＨＴＴＰのヘッダ情報から抽出する場合、Ｗｅｂページ内の文面ごとに抽出する場合を切り分けて行なった。またユーザ情報抽出においては、特にＢＢＳの場合は個別の投稿記事ごとに一意なユーザの判定・計数を行なった。

・差分収集手法：蓄積をベースとした通常のサーチエンジン用の情報収集とは異なり、日々の変化を調べるためには古いデータは必要なく、当日情報から抽出した素データのみを収集する。

（２）情報の伝播測定

情報の伝播測定では、技術的なポイントは２つあり「伝播分析」と「伝播分析指標」である。

・伝播分析手法：どのように伝播しているのかを、一人の人が多く口にしたのではなく、より多くの人が口にしていることを計測することとし、また、日々の人物の入れ替わりを考慮し「観測日または一定期間の観測期間における”その話題を口にした人数の２乗／その話題の記事数”」を基本形とし変形した値を「浸透度」として定義し測定した。また過去すべての情報を活用するのではなく一定期間（ここでは３５日間）における浸透度の動向を測定し話題の伝播を測定する手法を特徴とする。

・伝播分析指標：上記の浸透度を様々な話題に対して計測し、一定の指標を作成する。ここで、話題とは、話題を象徴するキーワードを含む記事とし、話題伝播では記事数、口にした人数、公開された時間を数値化して分析する。今回は実験的にいくつかの話題について話題伝播を測定した。統計的な有意な指標を出すことは今後の検討課題である。

４．口コミ・マイニング・システムによる運用実験

口コミ・マイニング・システムを試作し３月から６月まで運用実験を行なった。ここでは、マスコミ・企業系から２サイト、コンシューマ系としてあるポータルサイトに登録されている配下の個人ページ、ＢＢＳ系２サイトを測定対象とし、コンシューマ系は時事系、コンピュータ／ネットワーク系の話題に登録されている記事から素データを収集した。

（１）全体傾向

３月～６月の１２２日間における調査対象となった記事件数は約３０万件、一意なユーザ数は約１０万人、名詞句は約６０万語（種類）であった。

（２）伝播計測

口コミ情報の伝播の測定として、３章で述べた「浸透度」によって計測した。図２には、あるネットワーク系のブランド名５つについての記事の出現数のグラフを示す。これに対して、「浸透度」の計算を当てはめ比較したものが図３である。

表１．観測された話題伝播の特徴

	既出	新出
時事系	コンシューマの関心の強弱を検知可能。	話題の認知の広がりが検知可能。
商品名	マスコミ・企業系とあまり連動せずコンシューマ主導で話題が展開。	企業発表に同期してコンシューマ主導で話題が展開。

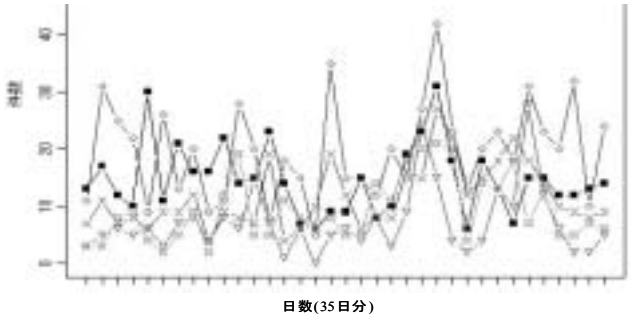

図２．話題に関する記事の出現頻度の推移(例)

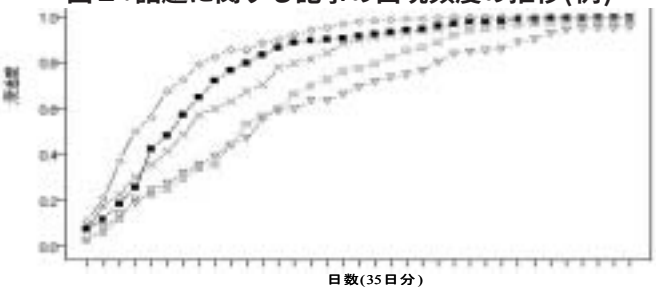

図３．話題の浸透度の推移（例）

伝播計測では、世間一般で認知されている・され始めている様々な話題（単語）に対して、図２と図３のグラフを作成し、時事系の話題、商品名の話題、既出の話題、新出の話題について特徴の違いを比較し、傾向を分析した。

分析結果の定性的な評価を表１に示す。

５．考察

図２を用いることにより、見かけ上の話題の多さを観測することができる。一方、出現記事数は多くても多くの人が口にしている話題、特定の人が口にしている話題があり、図３の「浸透度」により、より実際の話題性の強さや報道発表などの反響を知ることが可能になる。

今後は事例を収集し「浸透度」の標準的な指標つくりが課題である。

６．まとめ

話題が伝播してゆく過程を検知する手法として口コミ・マイニング・システムを開発し、４ヶ月間の運用実験により実インターネット上で情報伝播を測定しその結果を報告した。

[参　考　文　献]

[1]斉藤ほか，”インターネット上の口コミ解析手法の提案”，情報処理学会研究会報告 GN45-12，p.65-p.70，2002.

[2]”我が国におけるインターネットの着実な普及”，平成14年版情報通信白書，総務省．
http://www.johotsusintokei.soumu.go.jp/whitepaper/ja/h14/html/E1011000.html