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－

２Ｘ－０４

立石健二　森永聡　山西健司　福島俊一
ＮＥＣインターネットシステム研究所

１．　はじめに

企業活動における新製品開発や商品評価のためには、消費者からの意見を収集･分析することが重要であり、そのために市場調査や顧客満足度調査がしばしば実施されている。その方法は、まずアンケートを実施して消費者から意見を集め、得られた回答文を調査員が一つずつ確認して消費者の商品に対するイメージや商品のトレンドを分析することが一般的である。

しかし、このようなアンケートによる意見の収集は、アンケート実施者が調査票作成にかかるコストや、良質の回答モニターを選別しアンケートを依頼するコストが多く必要であり、そのように労力をかけて調査を行っても十分な数の自由回答文の意見が得られない場合があるという問題があった。また、収集した人の意見の人手による分析はスケーラビリティに欠け、分析過程での見落としや誤りが生じる可能性も否めない。

そこで筆者らは今回、インターネット上で様々な形で発信された意見を抽出し、テキストマイニングの手法で分析する新しいフレームワークとして「意見収集分析システム」を提案する。インターネットは各個人が自由に情報を発信できる場であり、そこには多くの人の意見が存在すると期待できる。本システムは、着目する商品に関する様々な人の意見を Web から自動抽出し、その意見をテキストマイニングツール SurveyAnalyzer[1]用いて分析することができ、新しいマーケットリサーチツールとしての可能性を持つ。

２．　意見収集分析システム

図１に意見収集分析システムの構成を示す。システムは大きく分けて、Web ページ収集部、意見抽出部、意見分析部の、3 つの部分から構成される。ユーザが収集条件として商品名を入力すると、Web ページ収集部はその条件に合致したページを収集する。収集条件として、ユーザは例えば「モバイルギア」「カシオペア」「シグマリオン」のような同分野（この場合 PDA）の商品名を複数入力する。次に、意見抽出部は、収集した Web ページから収集条件の商品名に関する意見に該当する文字列を抽出する。意見分析部は、意見抽出部の出力を、ユーザが指定した分析条件で分析し出力する。例えば、ユーザが「モバイルギア」を分析条件とすれば、システムはその特徴や傾向を最初に入力した３つの商品の意見全体と比較分析する。

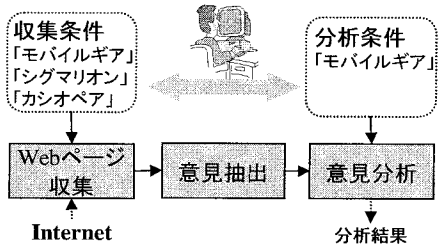

図１　意見収集分析システムの構成

上記のような意見収集分析システムの要件は以下の点である。

(1) **様々な Web サイトから意見を収集できること。**インターネットには、掲示板サイト、個人サイト、ニュースサイト等の様々な Web サイトが存在するが、各 Web サイトにはそれぞれの文化が存在し、それぞれには情報の偏りが存在すると考えている。そのため、本システムでは、収集条件に合致する多くのサイトから意見を抽出することが必要である。

(2) **様々な分析条件が指定できること。**分析条件として「商品名＝モバイルギア ＆ 評価＝肯定」のように指定できれば、「モバイルギア」に関する意見の中で「肯定」の評価を持つ意見の特徴を分析するというような、より詳細な分析が可能になる。

(3) **様々な分析方法を選択できること。**言葉の出現頻度による分析だけでなく、分析条件を満たす意見に特徴的に現れている言葉の分析や、ポジショニングマップによって商品間の関係を視覚的に分析ができることが必要である。

３．　システムの実装

前節で説明したようなシステムの要件を満たすためには、Web ページ収集部、意見抽出部、意見分析部をどのように実装するかが要点となる。この内、Web ページ収集部に関しては、今回、既存の検索エンジン「Google」で収集条件の商品名をキーワードとして入力した検索結果の上位 1000 件を用いた。以下、意見抽出部、意見分析部の実装方法について述べる。

３．１　意見抽出部

意見抽出部は、収集条件の商品名に関する人の意見を、収集した Web ページの集合から抽出し、それらに

Analyzing Opinions on the Web
--A Framework for Combining Information Extraction with Text Mining --
Kenji Tateishi, Satoshi Morinaga, Kenji Yamanishi, Toshikazu Fukushima
Internet Systems Research Laboratories, NEC Corp.

	意見	商品名	評価	表現	URL
1	モバイルギアMC-CS13の方が小さくていい	モバイルギア	肯定	いい	http://...
2	モバイルギアはすこし重い	モバイルギア	否定	重い	http://...
3	シグマリオンは..	シグマリオン	...	...	...

表１　意見収集部の出力例

カテゴリ毎のラベルを付与する。つまり、Web ページの集合を表１のような形式に整形する。

意見は、「モバイルギアは良い」のように商品名と評価表現という２つの単語の関係に着目し、次のような２つのステップで抽出する。

１．商品名と評価表現が一定の距離内に存在する場合は、両者を含む文書を意見候補とする。

２．意見候補の構文的な意見らしさを定量的に測定し、閾値以上の候補を意見とみなす。

ここで、評価表現とは、物事に対する評価を示す表現であり、「良い」「好き」「美味しい」といった人の感情や感覚を示す表現と、「速い」「重い」といった物の性質や特徴を示す表現を辞書として登録している。なお、この意見抽出方法に関しては[2]で詳しく説明しており、抽出精度約 87%を実現している。

次に、抽出した意見に対する各カテゴリのラベルを付与する。この中で、評価のラベルについては、評価表現にあらかじめ「肯定」又は「否定」の基本評価を付与しておき、その近傍に否定表現が出現する場合は評価を反転する方法を採用している。

３．２　意見分析部

意見分析部では、意見抽出部により抽出された意見をユーザが指定した分析条件に従って分析する。すなわち、意見抽出部の出力は表１のように、「意見」というテキストデータに、「商品名」や「評価」といったラベルがついたものの集合であるが、これらに対してテキストマイニングを行う。特に、ここでは意見分析という目的に鑑みて、自由記述アンケート分析手法として[1]に提案されているものを採用する。

例えば、最も基本的な意見分析は「モバイルギアに関する意見における特徴語」といった、ユーザが指定したラベル値（モバイルギア）を持つデータ（意見）を特徴づける単語（例えばモノクロ）を抽出するということである。この分析は「その単語（例えばモノクロ）を含むか否かでどれだけラベル値を予測できるか」を確率的コンプレキシティ[1]等で測定し特徴語としての適切性とみなすと言った方法で実現する。これによって単純に該当データ群における高頻度語を抽出するだけでは不可能な高品質の分析結果を得られる。

さらに、この特徴語抽出の拡張として、複数のラベル値（例えばモバイルギア、シグマリオン、、）に対する特徴語抽出の結果を視覚的にわかりやすい方法で提

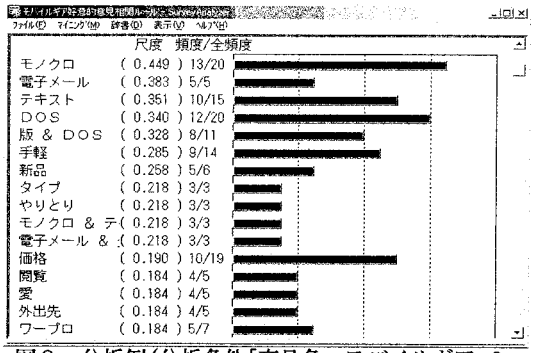

図２　分析例(分析条件「商品名＝モバイルギア & 意見の評価＝肯定」)

示する事（ポジショニングマップ）や、ユーザが指定した単語と特に強く結びついている単語の抽出（共起語分析）、特定ラベル値を持つデータの典型例を抽出（代表度分析）などの分析を行う[1]。

３．３　分析例

図２に、収集条件を「モバイルギア」等の４つの PDA の商品名とし、分析条件として「対象物名＝モバイルギア & 意見の評価＝肯定」を指定した場合の特徴後分析による分析結果を示す。比較対象は、分析条件で指定された分析対象以外の全てとした。分析結果からは、モバイルギアに関する肯定的な意見として文章入力について（例、「テキスト」「ワープロ」「タイプ」「電子メール」）や、旧モデルについて（例、「DOS 版」「モノクロ」）について特徴があることがわかる。

４．　おわりに

本稿では、着目する商品に関する様々な人の意見を Web から自動抽出し、その分析結果を出力する意見収集分析システムについて述べた。本システムの課題は、意見抽出部が出力する誤りの意見に含まれる言葉が特徴語として分析結果に表示される場合がある点である。この誤りの意見とは、例えば「カシオペア」のように PDA 以外にも多くの意味を持つ単語が対象物名である場合に多く見られる。また、Web ページ収集部は、今回簡易的な方式で実装しているが、Focused crawler 等により収集条件に関連するページを選択的に収集する方法が必要である。今後は、これらの問題を解決するとともに、Web 上の情報の特性を生かした分析手法を開発していきたい。

参考文献

[1] H. Li, K. Yamanishi, "Mining from open answers in questionnaire data", KDD2001, ACM Press, pp.443-449, 2001.

[2] 立石健二　石黒義英　福島俊一, "インターネットからの評判情報検索", 情報処理学会研究報告, NL-144-11, pp.75-82, 2001.