

報道内容の印象分析のためのテキストマイニング手法

須藤 一弘[†] 長尾 光悦[†] 大内 東[‡]

北海道情報大学[†] 北海商科大学[‡]

1. はじめに

今日の情報化社会において、消費者は報道メディアから発信された情報の良し悪し、すなわち印象に影響された行動をとる。多くの産業では、このような消費者の行動変化に迅速に対応する必要がある。したがって、報道内容が消費者に与える印象を分析可能にすることで、消費傾向の変化に対する適応性が実現可能になる。

本稿では、報道内容の印象分析のためのテキストマイニング手法を提案する。ここでは、形態素解析、特徴語の抽出、順位相関係数、係り受け解析などを行い、時間経過に伴う情報の変遷と与える印象の変化を分析する。この二つの結果を用いて報道内容の印象の分析を行う。提案手法の妥当性を実際の報道事例に適用することで検証を行う。

2. 報道内容の印象分析

報道メディアから発信される情報は、内容の一喜一憂が消費者行動に変化をもたらす。例えば、フードファディズムや風評被害などは、報道メディアが過剰に特定の情報を伝えることにより発生する。

すなわち、ある事柄に関する情報が消費者に対してどのような印象を与えるものか、また、どのような変遷をしているのかを分析可能にすることで、報道メディアに起因する消費行動の変化に素早く対応したサービスを提供することができる。

3. 提案手法

3.1 アルゴリズムの概要

図 1 に印象分析のためのテキストマイニング手法の概要を示す。先ず、ある話題に関するテキストデータに対して形態素解析を適用し、品詞分解をする。これにより、テキストデータ中の基本的な情報を得る。次に、相関ルールマイ

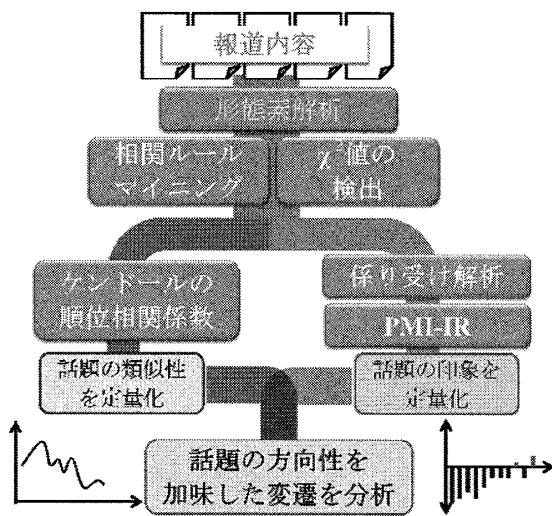


図 1. 提案手法の概要

ニングにより頻出語を抽出することで、文書の傾向を得る。また、文中の語の共起の偏りを文書の特徴として抽出するために χ^2 値を用いる。得られた文書の傾向・特徴について、ケンドールの順位相関係数を用いて類似性を測り、これを話題の遷移度合として分析する。更に、傾向・特徴を表す語がどのような用い方をされているのかを係り受け解析を用いて決定し、PMI-IR によって、その良し悪しを判定する。これらの結果を総合的に判断して、話題の印象を分析する。

3.2 形態素解析

本稿では、MeCab を用いて品詞分解を行う。この中から名詞を抽出し、分析する語として不適切な語を除外する。抽出した単語の出現文の数を計測し、出現語リストを作成する。

3.3 相関ルールマイニング

ここでは、アприオリアルゴリズムにより、頻出の単語、単語の組み合わせを抽出する。支持度と信頼度を計算し、それぞれ設定した最小値を満たす語を頻出語リストに加える。支持度では、日ごとに異なる出現語数と出現語種数の比率を加味する。

A Text Mining Method for Impression Analysis of Media Reports

[†] K. Suto, M. Nagao · Hokkaido Information University

[‡] A. Ohuchi · Hokkai School of Commerce

3.4 χ^2 値の検出

語の共起の偏りを特徴として抽出するために χ^2 値を用いる。抽出にあたり、特定の 1 語とだけ多く共起する語は、値が高くなるが、語に付随する語である場合が多いため、これを除外する。

3.5 ケンドールの順位相関係数

日ごとに抽出された傾向と特徴をケンドールの順位相関係数で比較し、その類似性を算出する。得られた傾向・特徴をもとに、ある二点間の話題の類似度を算出する。

3.6 係り受け解析

傾向・特徴として抽出された語 i を含む文を係り受け解析ソフトである Cabocha を用いて解析し、その語を係り受けている分節中の名詞、動詞、形容詞、副詞、接続助詞を抽出する。この組み合わせをテキスト中の語の意味の方向とする。

3.7 PMI-IR

係り受けによって定められた方向が、良い印象なのか悪い印象なのかを計測する。文中から抽出された語の組み合わせ $phrase$ と良い意味を表す語、悪い意味を表す語をそれぞれ合せて検索エンジンで検索し、そのヒット件数から相互情報量によって $so(phrase)$ を求める[1]。

また、PMI-IR の値は 2 語以上の単語検索において語間が短いものだけを結果とする NEAR 検索を用いることが望ましい。Google はこれに近い結果であるため、Google を用いる。

傾向・特徴として抽出された語 i の意味の方向性 $SO(i)$ は、 $so(phrase)$ の平均値とする。最終的に、テキストデータの意味の方向性 SO_a を傾向と特徴の意味の方向性から求める。

得られた話題の類似度と話題の印象の二つの観点から、総合的に報道メディアの話題内容の変遷に伴う印象の分析を行う。

4. 提案アルゴリズムの実例への適用

提案アルゴリズムを実例へ適用し、その妥当性を検証した。本稿では、適用事例として 2007 年に発生した新潟県中越沖地震の際に Niigata-Nippo, MNS 産経, z YOMIURI ONLINE 各サイトで発信された情報を用いる。この地震は、風評被害により観光産業へ多大な被害を与えた。

図 2 には話題の類似度の変遷を示し、表 1 には印象の定量化結果を示す。図 2 においては、ヒューリスティックな記事の分類による分析結果と近似した結果が見られた[2]。一方、印象の定量化の結果は、印象の値を良いならば正に、

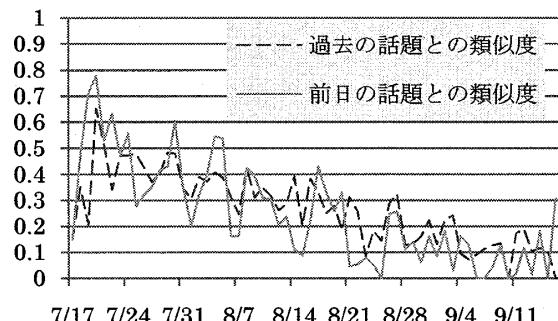


図 2. 話題の類似度の変遷

表 1. 印象の定量化結果

7月 16 日	7月 27 日	8月 16 日
地震 -1.214	地震 -1.001	地震 -0.646
柏崎市 -0.69	東電 -1.019	復興 -1.116
震度 6 強 -1.032	必要 -0.442	仮設住宅 -1.348
観測 -1.309	義援金 -0.96	新潟明訓 -0.827
...
$SO_a = -1.977$	$SO_a = -1.542$	$SO_a = -1.654$

悪いならば負として表しているが、ほとんどの値がマイナスになり、人間による分析結果とも異なる結果が見られた。原因としては、係り受け解析における方向性の決定のための語の不足、PMI-IR の算出における良い意味と悪い意味の基準とする語の選択が挙げられる。

5. おわりに

本稿では、報道内容の印象分析のためのテキストマイニング手法を提案した。提案アルゴリズムを実装し、実例への適用によりその妥当性を検証した。

参考文献

- [1]Peter D. Tuney : "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 (ACL'02), pp417-424(2002)
- [2]須藤一弘, 長尾光悦, 大内東 : "メディアの情報遷移を把握するための話題分析アルゴリズムの開発", FIT2009 講演論文集, pp247-248(2009)