D-048

テキスト、リンク、時系列分析を用いたソーシャルメディアにおける話題分析
Extracting Topics on Social Media with Link, Text, and Time Series Analysis

石田 和成†

Kazunari Ishida

1． まえがき

本研究ではブログ更新データから、キーワードの時系列データを抽出し、関心の高い話題を分析する。時系列データの抽出には、単純なキーワード頻度に加え、ブログ内のテキスト、リンク情報を併用した指標を用いる。これらの指標を用いて、ブログに自動挿入されるニュースやテキスト広告の影響と、ブロガーの持つ深い関心を区別する。さらに、抽出した時系列データに対して、独立成分分析（Independent Component Analysis, ICA）[1]を用いて相互に独立な成分を抽出し、主要トピックを分析する。キーワードは文脈によって様々な意味を持つため、単純なキーワード間の共起関係では、キーワードの意味の多義性を区別できない。この区別を行うため、ICA による主要トピックの波形を文脈として用い、その波形と有意な相関を持つキーワード群を用いて、その文脈の解釈を行う。

2．キーワードの関心度

ブログやそれに付随するページ群は、様々な主体が更新するソーシャルメディアを形成している。個人のもつ深い関心事項に関するブログが存在する一方、日々日記のようにテーマを定めず更新するブログが非常に多い。また、マスメディアが配信する最新ニュースタイトルを自動挿入するブログサイトも多い。さらに、広告収入を目的としたブロガーは、更新回数を増やしてアクセス数を増やすために、ニュースのコピーなどでブログを更新している。これらの情報を区別するための指標を、2.1,2.2,2.3 で定義する。

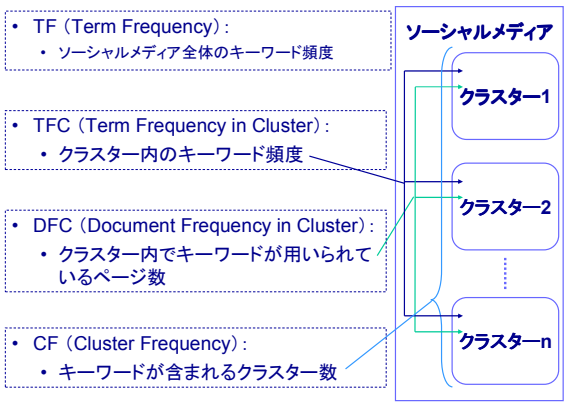

図 1. ソーシャルメディアの特徴量

それら指標の定義には、ソーシャルメディアから抽出できる 4 つの特徴量を用いる。それらの特徴量は、ブログや参照ページに含まれるキーワード頻度(Term Frequency, TF)、クラスター内キーワード頻度（Term Frequency in Cluster, TFC）、1 つのクラスター内で注目するキーワードが用いられるブログのページ数（Document Frequency in Cluster, DFC）、あるキーワードが含まれるクラスター数（Cluster Frequency, CF）である（図 1）。クラスター抽出には SI アルゴリズム[2]を用い、共参照リンクの多いブロガーのクラスターに分ける。それら 4 つのソーシャルメディアの特徴量にもとづき、3 つの指標を定義し、ブロガーの関心やマスメディアの影響を分析する[3]。

2.1 ソーシャルメディアの話題

1 つ目の指標は、キーワード頻度（TF）である。これは更新ブログとそれに参照されるページすべてのキーワードを抽出したものである。この指標には、ブロガーの意図したキーワードに加え、自動更新されるニュース記事のヘッドラインなどの意図しないキーワードも含まれるため、TF を「ソーシャルメディアの話題」と呼ぶ。

2.2 一般ブロガーの関心

2 つ目の指標は、クラスター内キーワード頻度(TFC)である。この指標はクラスター内に含まれるキーワードを全てカウントしているため、ブロガーが興味を持った一般的な流行キーワードが多く含まれるため、TFC を「一般ブロガーの関心」と呼ぶ。

2.3 特定ブロガーの関心

3 つ目の指標は、TFC、DFC、CF にもとづくランキングのトップ 10 キーワードに注目キーワードを含むクラスター数（Cluster Frequency based on Term Ranking, CFTR）である。各クラスターにおけるキーワードのランキングは、クラスター内のキーワード頻度（TFC）や、ブログのページ頻度（DFC）が高く、複数のクラスターに出現する頻度（CF）の低いキーワードが上位となるように、「TermRank ＝ TFC * DFC / CF^3」という式を用いる。このランキングにおける上位キーワードは、クラスターの特徴をよく表すものと考えられるため、CFTR を「特定ブロガーの関心」と呼ぶ。

3. 主要トピックの抽出

キーワードは文脈によって様々な意味を持つ。このキーワードの多義性により、1 つのキーワード数の変化には複数の社会事象の発生が影響を及ぼしている。本研究では、複数のキーワード数の波形から、相互に独立な社会事象の波形を得るために、独立成分分析（ICA）を用いる[1]。ICA は複数の音源から生じる音を、いくつかのマイクで録音したデータから、元の音源の音を分離する、カクテルパーティー問題を解く手法として知られている。ここでは、個々の音源を現実社会で観察される事象として、マイクを各キーワードとして、ブログの話題分析に ICA を適用する。

†島根県立大学, k-ishida@u-shimane.ac.jp

キーワードデータとして、継続的に収集しているブログ更新データから抽出した特徴量にもとづく 3 つの指標 TF,TFC,CFTR を用いる。分析したデータは 2006 年 3 月 19 日から､2007 年 5 月 20 日の間に､主要ブログサイトや PING サーバから収集した更新ブログである。このデータを 1 週間ごとに分け、クラスター、キーワードの抽出を行った。日本を取り巻く社会環境に関する主要トピックを調べるため、「独島、竹島、尖閣、北方領土、中国、北朝鮮、韓国、ロシア、アメリカ、自衛隊、米軍、拉致、自民党、民主党、共産党、公明党、ミサイル、核実験、慰安婦」という１９のキーワードを用いた。

3.1 キーワード間の相関

３つの指標、TF、TFC、CFTR、それぞれにおいて、統計的に有意なキーワード間の相関を調べるために、無相関の検定を行った。その結果、同一の島を示す「竹島」「独島」は３つの指標全てにおいて高い相関が見られた（竹島問題[５]）。ソーシャルメディアの話題(TF)と、一般ブロガーの関(TFC)については、国際政治的主体「北朝鮮」と、その駆け引きの手段である、「ミサイル」「核実験」の間でそれぞれ高い相関が見られた（ミサイル問題[６]と核実験問題[７]）。特定ブロガーの関心(CFTR)については、「竹島」「独島」それぞれに「自民党」「民主党」といった政党との高い相関が見られた。また、CFTR については有意な相関関係の密度は低く、話題の解釈は比較的容易であった。

しかし、キーワード間の有意な相関関係の密度が高い場合、具体的な話題の解釈が困難になる。例えば TF、TFC では、「竹島と独島」、「北朝鮮とミサイル、核実験」以外のキーワードは、相互に高い相関を持ち、具体的な話題の推測が困難であった。特に「中国、韓国、ロシア、アメリカ、自衛隊、米軍、拉致、自民党、民主党、共産党、公明党、慰安婦」は全てのキーワード間で有意な相関が見られた。

3.2 ICA による主要トピックの抽出

キーワードは多義的な意味を持つため、用いられる文脈は様々である。この文脈を取り出すために、本研究では独立成分分析(ICA)を用いる。この ICA の前処理として、データの次元圧縮のために、主成分分析(Principal Component Analysis, PCA)が用いられる。ここでは CRAN の fastICA パッケージ[４]を用いて、１９のキーワードの時系列データから、PCA を用いて８つの波形に圧縮し、さらに ICA を用いて 8 つの独立な波形に変換した。それら抽出された PCA、ICA それぞれの波形と、キーワードの波形との相関を、無相関の検定により調べた。

PCA の波形では、第 1 軸の波形は非常に多くのキーワード波形との間で有意な相関があるが、それ以降の軸については有意な相関が少なく、波形の解釈が難しい。それに対して ICA の波形の多くが複数のキーワードと高い相関を示しているため、波形の解釈を容易に行うことができる。

PCA と ICA それぞれで得られた波形が、元のキーワードともつ相関の高さを調べるために、8 つの波形それぞれについて、もっとも高い相関の絶対値を調べ、それらの平均を調べた(表 1)。相関の絶対値を用いた理由は、ICA により得られた波形における符号の不確定性のためである[1]。表 1 によると、PCA に比べ、ICA で得られた波形は、キーワードデータと高い相関を示しており、波形の示す文脈が解釈しやすいことが分かる。

表１． 抽出波形の相関の絶対値による最大値の平均

	PCA	ICA
TF	０．５８５５	０．７７１４
TFC	０．６０２２	０．８２１３
CFTR	０．５１５０	０．６７７６

ICA により抽出した波形とキーワードデータとの相関を調べたところ、ソーシャルメディアの話題(TF)、一般ブロガーの関心(TFC)について、3.1 で述べたキーワード間の関連性が同様に見られた。それらに加え、領土問題と解釈できる「北方領土、尖閣、竹島」の有意な相関が新たに見られた。また、3.1 で見られた、たくさんのキーワード間の密な相関関係は依然として見られたが､TF において｢慰安婦、共産党、民主党、尖閣」といった少数のキーワード間での有意な相関が見られた。これはキーワードの多義性のため区別が困難な複数の文脈の分離に役立つ可能性を示している。

特定ブロガーの関心(CFTR)については､3.1 のキーワード間の相関では見られなかった、「ミサイル、北朝鮮」（ミサイル問題）や、「自衛隊、アメリカ、韓国、北朝鮮」の間に有意な相関が見られた。

4. まとめ

本研究では継続的に収集しているブログ更新データから、キーワードの時系列データを抽出し、関心の高い話題を抽出した。時系列データの抽出には、単純なキーワード頻度に加え、ブログ内のテキスト、リンク情報を併用した指標を用いた。それら時系列データに ICA を適用し、主要なトピックの抽出を行った。今後の課題は、キーワード指標の改善や、ICA で抽出する波形数、また、ICA における符号の不確定性についての取り扱いを検討する。

参考文献

1. Hyvarinen, A., J. Karhunen, and E. Oja (根元幾、川勝真喜 訳), 「詳説　独立成分分析-信号解析の新しい世界」, 東京電機大学出版, 2001 (2005 訳).
2. 石田和成, ″ライングラフと局所的類似度にもとづくマルチクラスタリングアルゴリズム″, データベースシステム研究会および情報学基礎研究会による合同研究会, 情報処理学会研究報告, Vol.2006, No.59, pp. 69-76, 2006.
3. 石田和成, ″社会環境の変化と CGM ″, 情報システムと社会環境研究会, 情報処理学会研究報告, Vol.2006、No.92, pp. 37-44, 2006.
4. Marchini, J.L., C. Heaton, and B. D. Ripley, Implementation of FastICA algorithm to perform Independent Component Analysis (ICA) and Projection Pursuit., http://cran.r-project.org/src/contrib/Descriptions/fastICA.html
5. 山陰中央新報,”竹島資料室が臨時開設”, 2007/2/22.
6. 東京新聞, “北朝鮮ミサイル発射　横須賀基地は平静”, 2006/07/06,
7. 中央日報, “北朝鮮核実験場所、花台郡舞水端里山の地下と推定”, 2006/10/9.