様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月２０日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500207
研究課題名（和文）　ネットニュースにおける話題分析アルゴリズムの開発と自然災害風評被害への適用
研究課題名（英文）　Development of Topic Analysis Algorithm for Online News and Application of Algorithm to Harmful Rumor caused by Natural Disaster
研究代表者　大内　東（OHUCHI AZUMA）
北海道大学・大学院情報科学研究科・教授
研究者番号：50002308

研究成果の概要：
本研究では，ネットニュースから中心的に扱っている話題を語の共起情報や出現頻度に基づいて抽出し，その時系列的な変化を提示するアルゴリズムの提案した．また，提案したアルゴリズムを2007年に発生した新潟県中越沖地震発生時のネットニュースに適用し，当時のネットニュースでの報道状況を解析した．これにより，提案アルゴリズムの検証を行うと共に，ネット上の情報にバイアスがかかり歪められていく過程を明らかにした．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007年度	1,800,000	540,000	2,340,000
2008年度	1,700,000	510,000	2,210,000
年度			
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：複雑系工学，システム工学，観光情報学
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：話題分析，ネットニュース，自然災害風評被害，情報バイアス

１．研究開始当初の背景

情報化社会の到来により，近年では Web 上から様々なネットニュースを入手することが可能となった．特に，Blog 等を用いて個人でも情報発信の主体となれることから，以前には無い，多種多様なネットニュースが提供されることとなった．しかしながら，Web 上の情報には様々なバイアスがかかっており，ネットニュースが必ずしも現実の正しい姿を反映しているとは限らない．時には，誇張された情報や全くの流言飛語が流れてしまい，観光客が途絶えたり，食品の売り上げが激減したりするなどの不利益が生じてしまう場合がある．このような現象は風評被害と呼ばれ，深刻な社会問題として認識されている．

風評被害は「ある事件・事故・環境汚染・災害が大々的に報道されることによって，本来「安全」とされる食品・商品・土地を人々が危険視し，消費や観光を取りやめることによって引き起こされる経済的被害」と定義される．この風評被害の問題に対して，従来の研究では，社会心理学的な分析を加えた上で，「不確かな情報に踊らされないように関係者を啓蒙していくべき」という結論に留まっているのが現状である．しかしながら，特に

自然災害の場合には関係者の範囲が広いため啓蒙活動を周知徹底することは困難な状況であることもあり，風評被害は依然として深刻な問題となっている．更には，近年のインターネット発のニュースソースの増加に伴って，バイアスがかかった情報の弊害が顕著化している状況であり，風評被害を代表とする情報バイアスによる問題の解決は強い社会的ニーズになっている．

２．研究の目的

本研究では，情報が歪められた事例として風評被害の問題を取り上げ，ネットニュースと現実との乖離を時系列で調査することによって，情報にバイアスがかかっていく過程を明らかにすることを目的とする．

具体的には，ネットニュースから中心的に扱っている話題を語の共起情報や出現頻度に基づいて抽出し，その時系列的な変化を提示するアルゴリズムの提案を行う．更に，提案アルゴリズムを 2007 年に発生した新潟県中越沖地震の事例に適用し，当時のネットニュースでの報道状況を解析することによって，ネット上の情報にバイアスがかかり歪められていく過程を明らかにする．

３．研究の方法

本研究では，ネットニュースが現実とどのように乖離していくかを明らかにするため，対象のニュースの話題を抽出すると共に，その時系列での変遷を提示するアルゴリズムを開発することを目的としている．

これを開発するための前段階として，自然災害の発生に伴い風評被害が発生した際のネットニュースにおける報道状況の分析を実施する．ここでは，風評被害の発生事例として，2007 年に発生した新潟県中越沖地震を事例として取り上げる．これら地震災害発生時の情報量の調査，及び，風評被害へつながると考えられる内容の報道であるか否かの分類を実施し，これらの時系列分析を実施する．

この分析により得られた知見に基づき，ネットニュースにおける話題内容を時系列で分析可能なアルゴリズムを開発する．本研究における話題分析アルゴリズムを図１に示す．図１に示されるように，与えられた災害関連情報を，言語で意味を持つ最小単位である形態素へ分解する．抽出された形態素に対して相関ルールマイニングを適用することにより災害関連情報内で頻出する語を抽出する．更に，抽出された頻出語の共起頻度に基づき χ^2 値の算出を行い，これにより重要語を抽出する．これらの頻出語及び重要語リストに基づき順位相関係数を算出する．これら係数から

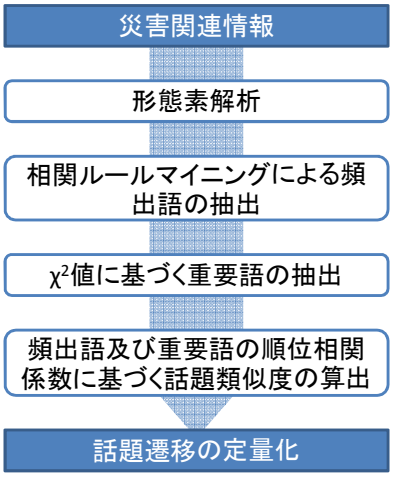

図１　話題分析アルゴリズム

ら話題の類似度を算出する．最終的に，この値の変遷を算出することによってネットニュースにおける話題遷移の定量化を行う．

本アルゴリズムでは，相関ルールマイニングによって，全体的な話題内容を把握可能とし，更に，χ^2 値に基づく重要語の抽出によって，全体的な話題の中での注視されている特定の話題を把握可能とし，この両面から話題内容を分析する．各ステップの詳細を以下に示す．

形態素解析

形態素解析は，京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジン Mecab を用いて行う．また，抽出された形態素情報から語の連結を行う．形態素解析においては，例えば，「新潟県中越沖地震」という単語は，「新潟-名詞」，「県-名詞」，「中越-名詞」，「沖-名詞」，「地震-名詞」と一つの単語が分解されて抽出される．このため，名詞が連続して出現する場合には，それらを連結し一つの名詞として扱うための前処理を適用する．次に，名詞以外の語，単体数字，記号，指示語，位置をさす語，複数形に関する語等をストップワードとして除去する．これら処理により，災害関連情報から名詞情報の抽出を行う．

相関ルールマイニングによる頻出語の抽出

1994 年に Agrawal によって提案された相関ルール抽出アルゴリズムであるアプリオリアルゴリズムにより頻出語の抽出を行う．このアルゴリズムは，ユーザが支持度と確信度の最小値を与えることにより，これら閾値以上の値を持つ相関ルールを抽出するものである．これにより頻出語の抽出を行う．また，ここでの語とは，単語

または複数の単語からなるフレーズである．

第一に，最小支持度を満たすサイズ 1 の語を求める．ここでは，支持度の算出において基準となる日と比較し，文書全体において語が出現する確率を考慮した重みづけを行う．以下に算出方法を示す．

$$s = \frac{P(X \mathrm{I}\ Y) \times I \times N_{std}}{N \times I_{std}}$$

N : 述べ出現語数
I : 出現語種数
N_{std} : 基準出現語数
I_{std} : 基準出現語種数

本研究では，支持度算出の基準日を地震発生 0 日目と設定し，最小支持度を 0.015 として設定する．これはヒューリスティックに決定したものである．

続いて，サイズ 1 の頻出語の共起行列からサイズ 2 の語を生成し，この支持度に基づく頻出語の抽出を行う．但し，サイズが 2 以上の場合には，出現回数が 1 回の語は除外される．語のサイズを増加させ，最小支持度を満たすものがなくなるまで処理を繰り返す．

第 2 ステップとして，抽出された頻出語に対して以下のように確信度 c の算出を行い，最小確信度を満たさない頻出語は除外される．ここでの最小確信度の設定についてもヒューリスティックにより 0.7 と設定する．これにより頻出語リストを作成する．

$$c = P(Y \mid X)$$

また本研究では，サイズの大きな頻出語抽出においては，部分集合となる語の支持度が同一の場合には，これを除去する操作を行う．これにより，頻出語における冗長性の解消を行う．

χ^2 値に基づく重要語の抽出

相関ルールマイニングにより抽出された頻出語リストから共起行列を生成する．頻出語の出現頻度に基づき各語の期待頻度を算出する．この値と語同士の共起頻度を観測値として χ^2 値を以下の式により算出する．また，ここでの値としては，ある一語とだけ特別に共起するものは，それらの語の間は付属語または付随語の関係があるものと仮定し，χ^2 値の最大値を取り除いたものを採用する．これにより重要語リストを作成する．

$$\chi^2(i) = \sum_{w \in W} \frac{(freq(i,w) - n_i pw)^2}{n_i p_w}$$

$$\chi^2(i)' = \chi^2(i) - \max_{w \in W} \left\{ \frac{(freq(i,w) - n_i p_w)^2}{n_i p_w} \right\}$$

$freq(i,w)$: 語 i と語 $w \in W$ の共起頻度
n_i : 語 i と頻出語群 W の共起総数
pw : 頻出語単独での生起確率

ケンドールの順位相関係数に基づく話題類似度の算出

頻出語リストおよび重要語リストは日単位でネットニュースから抽出される．これらリスト間においてケンドールの順位相関係数を算出し，全体的な話題内容とその中での注目されている話題の類似性を算出する．

$$r_k = \frac{\sum P_{ij} - \sum Q_{ij}}{\sqrt{\frac{n(n-1)}{2} - T_x}\sqrt{\frac{n(n-1)}{2} - T_y}}$$

$$T_x = \sum_{i=1}^{n_x} \frac{t_i(t_i - 1)}{2}$$

$$T_y = \sum_{j=1}^{n_y} \frac{t_j(t_j - 1)}{2}$$

P_{ij} : リスト x の語 i と語 j の順位関係 $i > j$ がリスト y でも満たされる組み合わせ
Q_{ij} : リスト x の語 i と語 j の順位関係 $i > j$ がリスト y では $i < j$ である組み合わせ
n : リストの長さ
t_i : リスト x における第 i 位の同順位数
t_j : リスト y における第 j 位の同順位数

これら二つの順位相関係数を用いて，全体としての話題類似度を算出する．

$$Sim(i,j) = r_{k\_freq}(i,j) \times \left(1 + \sqrt{(r_{k\_imp(i,j)})^2}\right)$$

但し，$i < j$
$Sim(i,j)$: i 日目と j 日目の話題類似度
$r_{k\_freq}(i,j)$: i 日目と j 日目の頻出語リスト間の順位相関係数
$r_{k\_imp}(i,j)$: i 日目と j 日目の重要語リスト間の順位相関係数

また，ここでは話題類似度の算出には順位相関係数の絶対値を採用している．これは，リスト間に負の相関がある場合は，全体的な話題内容は同一であるが，注目される話題の順位が入れ替わった状態である

と考えられるため，負の相関がある場合にも話題内容の類似性があると考えたためである．

更に，この類似度の算出方法に基づき $i-1$ 日目と i 日目の類似度，及び，0 日目から $i-1$ 日目までと i 日目の類似度を算出し，その平均値を検証する．

本アルゴリズムを新潟県中越沖地震におけるネットニュースに適用し，発生当時の話題変遷の分析を行う．

４．研究成果

本研究で実施したネットニュースにおける報道状況の分析結果を図２及び図３に示す．

図２は，新潟県中越地震でのネットニュースにおける情報量の分析結果である．これは，全国誌が運営するネットニュースサイトである，MSN 産経，読売 Online，及び，地方紙が運営する新潟日報 Online の３サイトにおいて情報量の調査を実施した結果である．ここで収集した災害関連情報としては，各サイトにおいて新潟県中越沖地震関連の記事が集約されたページにおける記事数を調査している．結果では，３サイトにおける記事数の平均値を示している．

図３は，収集された記事を風評被害へつながると考えられる内容の報道であるか否かで分類した各カテゴリの割合を示している．ここでの分類は，地震発生地域の近隣観光地への観光に対して忌避感情を抱かせるか否かに基づき分類を行った．死亡や被害状況のような，忌避感情を抱かせる内容であれば「－」，復興の進展や PR 活動などの忌述感情を緩和させる内容であれば「＋」，関連する政治経済の動静などの忌避感情に影響しない内容であれば「0」と分類した．これにより，メデイアが災害地の近隣観光地をどのような存在として取り上げたかが明確となる．

災害関連記事の情報量及び分類結果から，新潟県中越沖地震においては，地震発生直後から２週間程度は情報量が多く，また，被害などの近隣観光地への忌避感情を誘発する「-」カテゴリの記事割合が高い．すなわち，ネット上の情報にバイアスがかかっている状態となっている．その後，情報量が徐々に減少し，「-」カテゴリの記事割合の減少と共に，忌避感情の緩和につながる「+」カテゴリの記事が増加し，地震発生から１ヵ月程度でバイアス状態が解消されていると考えられる．

続いて，図４及び表１に本話題分析アルゴリズムに基づき新潟県中越沖地震の災害関連情報を分析した結果を示す．

図４は，話題分析アルゴリズムに基づき算出された話題の類似度の時系列データを示

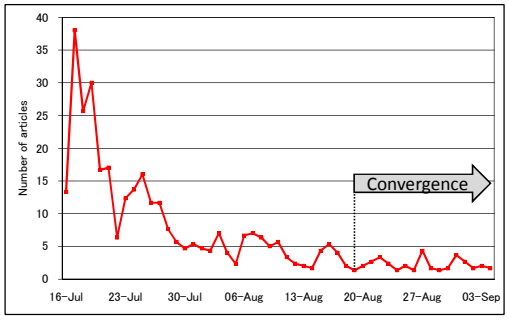

図２　災害関連情報の記事数

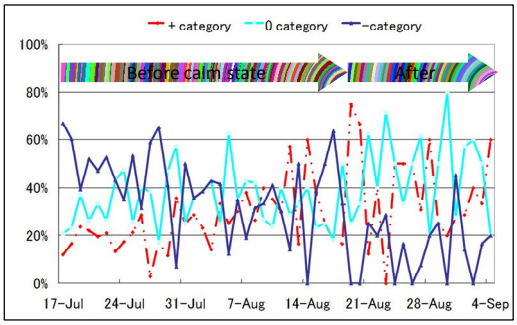

図３　災害関連記事の分類結果

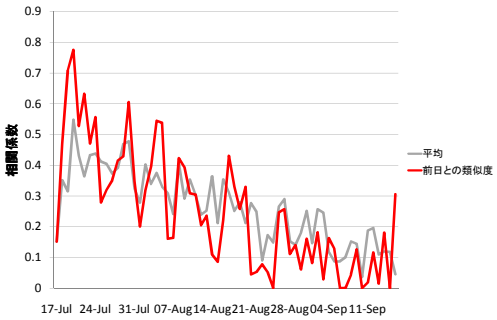

図４　話題分析アルゴリズムによる災害関連情報の分析結果

している．図から，地震発生から１カ月程度経過した８月 22 日当たりから平均値が減少し，これまでの話題と方向性の異なる話題に変化してきたことが確認された．更に，前日との類似度の値も低い値となり，話題内容にばらつきが発生していることが確認された．また，８月 22 日から３日間の頻出語及び重要語の上位５個を表１に示す．これらの語からも，この３日間では話題にばらつきがあることが確認できる．

すなわち，地震発生から１カ月程度で話題におけるバイアス状態が解消されたと考えられる．この結果は，前述の災害関連記事の忌避分類結果ともほぼ一致する結果である．これにより，本アルゴリズムによって，ネッ

表１　バイアス状態解消期間における頻出語と重要語

頻出語		
8月22日	8月23日	8月24日
夏休み	日本	地域
中越沖地震	再開	質問
予定	廃業	被災地
柏崎市	検討	参加
授業	新潟県中越沖地震	研修
重要語		
8月22日	8月23日	8月24日
揺れ	世界	人たち
入居	原発	地元
必要	危機感	感謝
刈羽村	東電	天候
今井町長	今回	心配

トニュースにおける話題変遷の定量化を実現可能であることが示された.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 須藤一弘、長尾光悦、大内　東、災害関連情報の比較分析に基づく風評被害対策方法の研究、観光情報学会誌、第５巻、33-44、2009年、査読有

〔学会発表〕（計12件）

① 長尾光悦、テキストマイニングに基づく災害関連情報の話題遷移分析、電子情報通信学会2009年総合大会、2009年３月18日、松山市・愛媛大学
② 長尾光悦、地震災害関連情報の分析に基づく風評被害対策に関する考察、情報処理学会第71回全国大会、2009年３月12日、草津市・立命館大学びわこ・くさつキャンパス
③ Mitsuyoshi Nagao, An Analysis of Media Information for Implementing Effective Countermeasure against Harmful Rumor, Applications of Physics in Financial Analysis 7th International Conference, 2009年３月３日，東京都・東京工業大学
④ 長尾光悦、地震発生時における災害関連情報の推移と被害特性の分析、情報処理北海道シンポジウム2008、2008年９月19日、稚内市・稚内北星学園大学
⑤ 大内　東、自然災害における近隣観光地の風評被害～モデル，被害度指標，災害ポータル～、日本オペレーションズ・リサーチ学会 2008 年秋季研究発表会、2008年９月11日、札幌市・札幌コンベンションセンター
⑥ 長尾光悦、メディアの動向を意識した風評被害対策に関する考察、日本オペレーションズ・リサーチ学会 2008 年秋季研究発表会、2008年９月11日、札幌市・札幌コンベンションセンター
⑦ 長尾光悦、風評被害対策に向けた話題分析に関する基礎研究、第７回情報科学技術フォーラム（FIT2008)、2008年９月２日、藤沢市・慶應義塾大学湘南藤沢キャンパス
⑧ 長尾光悦、風評被害の早期回復に向けたメディア情報転換期に基づくアプローチ方法の研究、第５回観光情報学会全国大会、2008年５月28日、旭川市・大雪クリスタルホール 国際会議場
⑨ 大内　東、大規模自然災害による近隣観光地の風評被害研究、第５回観光情報学会全国大会、2008年５月28日、旭川市・大雪クリスタルホール 国際会議場
⑩ 長尾光悦、風評被害の抑制・防止に向けた地震災害情報の分析、情報処理学会第70回全国大会、2008年３月15日、つくば市・筑波大学
⑪ 長尾光悦、風評被害対策に向けたメディア分析に関する研究、情報処理学会北海道シンポジウム 2007、2007 年 ９月 19日、札幌市・北海道工業大学
⑫ 長尾光悦、災能登半島地震におけるメディア分析～風評被害対策に向けて～、第４回観光情報学会全国大会、2007年６月19日、湯沢町・越後のお宿いなもと

６．研究組織
(1)研究代表者
大内　東（OHUCHI　AZUMA）
北海道大学・大学院情報科学研究科・教授
研究者番号：50002308

(2)研究分担者
川村　秀憲（KAWAMURA　HIDENORI）
北海道大学・大学院情報科学研究科・准教授
研究者番号：60322830
長尾　光悦（NAGAO　MITSUYOSHI）

北海道情報大学・経営情報学部システム情報学科・准教授
研究者番号：30343015

(3)連携研究者
なし

(4)研究協力者
須藤　一弘（SUTO　KAZUHIRO）
北海道情報大学・大学院経営情報学研究科・修士課程２年