E-001

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

Development of Topic Analysis Algorithm for Obtaining Media Information Transition

須藤 一弘† 長男 光悦† 大内 東‡
Kazuhiro Suto Mitsuyoshi Nagao Azuma Ouchi

1. はじめに

メディアから発信される情報は，消費者の行動に影響する．時間の経過に伴う情報の連続性を分析可能にすることにより，風評被害のような，メディアの情報提供に起因する事象について詳しく分析することが可能となる．

本稿では，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提案を行う．ここでは形態素解析，頻出語や重要語の抽出，順位相関係数を用いた話題分析アルゴリズムを開発し，ある話題における情報遷移を分析可能とする．適用事例として地震災害における風評被害を取り上げ，発生時の災害関連情報に適用し，アルゴリズムの妥当性について考察する．

2. メディアの情報遷移

メディアの情報は，消費者の心理に影響を与え，その行動に大きく反映される．例えば，風評被害のように，ある事物に対する偏った情報の発信によって，経済的被害を受ける場合もある．また，ある話題に関する報道は，時間の経過とともに内容も変化し，それに伴い，消費者の関心の方向性も変化していくと考えられる．時間の経過に伴う情報遷移の度合いを定量化することにより，話題の連続性が保存されているかを分析することが可能となり，将来的には，メディアの情報と消費者行動の関連について分析可能となることが期待できる．

3. 話題分析アルゴリズム

3.1 アルゴリズムの概要

図 1 に話題分析アルゴリズムの流れを示す．

まず，ある一定期間収集された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，単語に分解する．これにより，テキストデータに出現する語や行数といった，基本的な情報を取得する．

次に，相関ルールマイニングを行い，頻出語リストを作成する．単語単体だけでなく，いくつかの語から構成されるフレーズを抽出することにより，文書の量に応じ適切な範囲の意味を持つ情報を比較することができる．

作成された頻出語リストより，共起の特異性を基にした重要語リストを作成する．ある語とのみ特別に共起する語は，その文書の中で特別な意味をもった語であると考えられるためである．分布の偏りを検出する χ^2 検定の値を用い，この値を語の重要度とする[1].

最後に，時間経過に伴う情報遷移の度合いを，ケンドールの順位相関係数を用いて比較する．これをリストに適用

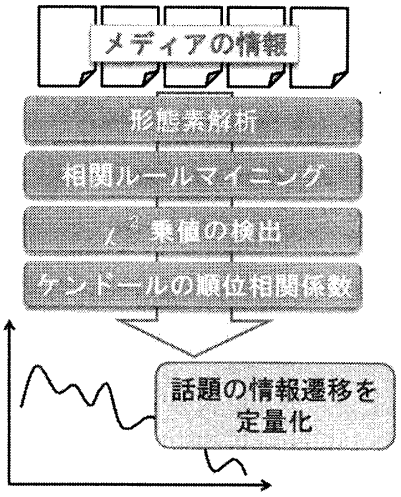

図 1. 話題分析アルゴリズムの流れ

することにより，リスト間の相関を比較することができ，情報の類似度を定量化することができる[2].

3.2 データの前処理

本稿ではオープンソース形態素解析エンジンである MeCab を用いて形態素解析を行い，文書を分かち書きする[3]．分割された文書において，名詞が連続して出現した場合は一つの名詞として扱う．この処理後，名詞のみを抽出し，分析する語として不適切な語（不要語）を除外する．各単語が出現した文の数を計算し出現単語のリストを作成する．

3.3 相関ルールマイニング

相関ルールマイニングでは，アプリオリアルゴリズムによりフレーズを作り，支持度と信頼度を計算し，設定した最小支持度，最小信頼度を満たした場合，頻出語リストに加える．支持度 *Sup* は以下の式により算出する．

$$Sup = P(X \cap Y) \cdot \frac{N}{I} \cdot \frac{I_{std}}{N_{std}}$$

$X:Y$ 以外の語（1 語の場合は $X = Y$）

支持度の計算において，出現回数の単位を行数とし，加えて単語の数と種類を加味することで，日ごとの情報の特徴を考慮する．このため，その日における語の種数 I と語の総数 N の比を分析初日の語の種数 I_{std} と語の総数 N_{std} の比とを比較した重みづけをする．

また，信頼度 *Con* は以下の式により算出する．

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

このときフレーズの信頼度は，Y に置く語により値が異なる．本稿では，最も低い値が最小信頼度以上ならば，条

† 北海道情報大学

‡ 北海商科大学

件を満たしたと判断する．これにより，語の繋がりが最も弱い部分であっても最小信頼度を満たすような，強力な組み合わせのみをフレーズとして扱う．すなわち，正確な文書の方向性を示す語を扱うこととなる．

3.4 χ^2 値の計算

重要度の指標となる χ^2 値の計算に重みづけをし，以下のように算出する．

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

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

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

特定の 1 語 w とだけ特別に共起している語 i は，χ^2 値は高くなるが，語 w に付随する語である場合が多いため，χ^2 値の最大の項を除いた値である $\chi^{2'}$ を用いる[1].

3.5 ケンドールの順位相関係数

頻出語リストの比較を以下の式により算出する．

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

$$T_x=\sum_{i=1}^{n}\frac{t_i(t_i-1)}{2}\qquad T_x=\sum_{j=1}^{n}\frac{t_j(t_j-1)}{2}$$

P_{ij}：リスト x の語 i と語 j の順位関係 $i>j$ がリスト y でも満たされている組み合わせ
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

また，本稿ではこの値の絶対値を用いる．これは，マイナスの値であっても，相関があることを意味しているためである．さらに，重要語リストを用いて重みづけをする．重要語においても相関があった場合，その二点間には，より強い相関があると考えられるためである．頻出語リストの相関 $R_{k\_freq}(i,j)$ と重要語リストの相関 $R_{k\_imp}(i,j)$ を用いた以下の式で話題類似度 $Sim(i,j)$ を算出する．

$$Sim(i,j)=\sqrt{(R_{k\_freq}(i,j))^2}\,(1+\sqrt{(R_{k\_imp}(i,j))^2})$$

分析では $i-1$ 日目と i 日目の類似度と，分析初日から $i-1$ 日目までと i 日目の類似度の中央値から検証する．

4. 災害関連情報への適用

本稿では，提案アルゴリズムを風評被害の発生した新潟県中越沖地震発生後のネットニュースにおける災害関連情報へ適用した．適用するにあたり，パラメータ設定を以下のようにした．

（1） 不要語は，数字単体，記号，指示語，位置を指す語，5W1H を指す語，複数形にする語とした．

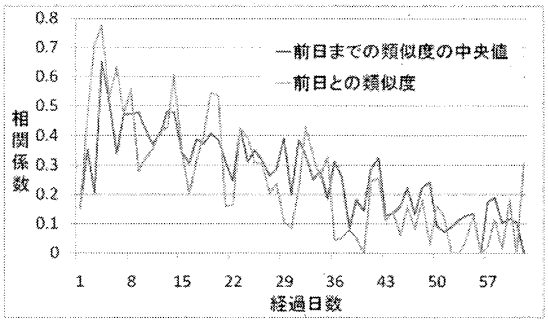

図 2．災害関連情報への適用結果

表 1．類似度の低い地点における頻出語

37 日目	38 日目	39 日目
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
新潟県中越沖地震	再建	新潟県中越沖地震
東電	売上げ	地域　減災
電力需要	原子力発電	自主防災組織

（2） 最小支持度は 0.015，最小信頼度は 0.7 とヒューリスティックに設定し，フレーズは出現回数 2 回以上の場合だけリストに加えた．

（3） ケンドールの順相関係数では，上位 20 語を比較した．

図 2 と表 1 は，災害関連情報への適用結果を示す．発生から 38 日目になると，前日との類似度，前日までとの類似度が共に極めて低い値になった．表 1 から分かるように，抽出されていた語を比較すると，内容が大きく違う語が抽出された．本アルゴリズムによって，情報の内容が明らかに異なる点を指摘することができた．この結果は，風評被害対策において重要であるメディアから発信された情報の連続性が失わる時期を分析することができると考えられる．

4. おわりに

本稿では，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提案をした．提案アルゴリズムを災害関連情報に適用し，適用結果からその妥当性について検討した．

参考文献

[1] 松尾豊，石塚満，”語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”，人工知能学会論文誌 Vol17，No3，pp217-223(2002)

[2] 大野成義，太田学，片山薫，石川博，”特徴間の類似性を考慮した特徴量集約手法の検討”，電子情報通信学会第 18 回データ工学ワークショップ(ISSN 1347-4413)，(2007)

[3] 京都大学情報学研究科―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，http://mecab.sourceforge.net/