4T-6

Weblog 上の話題伝播過程を利用した重要語抽出

Extracting Keywords using Topic Diffusion Process in Weblogs

古川　忠延†　　松尾　豊††　　大向　一輝‡
Tadanobu　Furukawa　　Yutaka　Matsuo　　Ikki　Ohmukai

内山　幸樹‡‡　　石塚　満†
Koki　Uchiyama　　Mitsuru　Ishizuka

1. はじめに

ウェブにおける情報発信の一形態として近年注目されているウェブログ（以下ブログ）では，その特徴として，記事が時系列に整理されていることや，一般にコメントやトラックバックの機能をもつなどの点が挙げられ，ブログ上では日常的に様々な新しい話題が生まれては，議論が広まっていく傾向がある．

ブログにおける話題は，その出現の仕方に様々なパターンがあり[1]，世間で流行として認識されているような大規模的なものばかりでなく，特定の嗜好を共有した小さなコミュニティ内でのみ伝播していく話題も存在していると考えられる．前者が時事を反映した突発的なものであるのに対し，後者は必ずしも突発的ではなく，徐々に広まっていくような話題である．本稿では，こうした普及の特性について，突発的に普及するタイプを「**瞬発性を持つ**」，徐々に広まるタイプを「**継続性を持つ**」話題と呼ぶ．これらはいずれもブロガの興味を惹きつけるものであり，本稿ではそれらを重要語として抽出していく．

本稿ではホスティングサービス Doblog[1)]のデータベースを用いることで，ユーザが記事を投稿する前に誰のブログを訪れたかという閲覧情報を使用した分析を行う．これにより，単に使用頻度が高いだけではない，「**広まりやすい話題**」を抽出できると考えている．

2. ブログにおける語の伝播

本稿における語の伝播は，「ある語を含む他者の記事を読んだブロガが，自身のブログにおいても**初めて**同じ語を含む記事を投稿すること」として定義する．

但し，分析に使用するデータでは，ブロガ単位での訪問関係は取得できるが，記事単位での取得はできないため，実際に読んで影響を受けたかどうかの判別は不可能である．そこで，ある制限期間 d を設け，ブロガ A が語 t を含む記事を投稿してから d 日以内に，別のブロガ B が A を訪問し，t を含む記事を自身のブログに投稿した場合に，A は B による対象の記事を読んで影響を受けたものとして扱い，「t が A から B へ伝播した」と定義する（図 1）．

ここで，トラックバックのデータを参考に d を定める．トラックバックは本稿で扱っている伝播の行動をまさに表すものであり，ある記事が投稿されてからトラックバックが行われるまでの日数が，d に対応するものであると言える．Doblog における全トラックバックについて，されるまでの日数とその件数をグラフで表したものが図 2 である．多くのトラックバックを網羅することが必要であるが，あまり日数が経ってからのトラックバックは内容的に伝播していると言い難いものがノイズとして含まれていたため，話題伝播として有効なトラックバックは，記事投稿から 3 日まで（全体の約 6 割）のものとした．すなわち，以降で述べる伝播では，***d=3***とする．

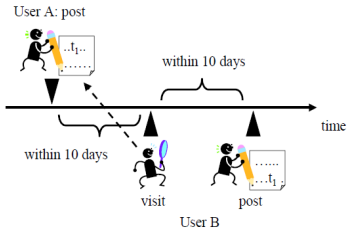

図 1　伝播の定義

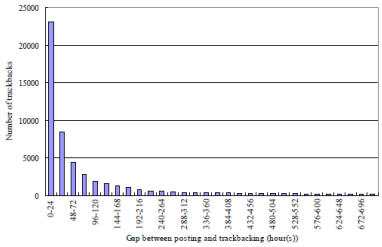

図 2　トラックバックされるまでの時間と件数の関係

3. 重要語抽出

本稿では仮説として，ブログ上における話題の伝播の仕方が，それぞれ静的に定まる「**語の影響力**」（語数の次元数を持つベクトル P）と「**人の影響力**」（人数の次元数を持つベクトル Q）の値によって説明できるとする．前述の通り，伝播情報は，「どの語が」「誰から」「誰に」伝播したかというデータで取得し，ここに各ブロガ・語に関する特徴を加味すると多様な形式で表現することが可能であるが，本稿では単純のため，「あるブロガが」「ある語を」伝播させた人数（＝語を誰が何人に伝播させたか）という二次元のデータ（行列 A）で表すこととする．以上を式で表すと下記の式のようになり，本稿で重要語を抽出することは，A から P（と Q）を計算し，高い影響力を持つ語を取得することに他ならない．

ここで仮説より，A は P と Q から定まる，すなわち人と語の静的な力によって行列の各成分が定まっていることを考慮すると，

$$A = P \cdot Q \cdot c \ (c \text{ は定数})$$

という積の形で表現できることになる．そこで本稿では，行列からベクトルを抽出する手法として特異値分解を用いる．

†東京大学大学院情報理工学研究科，
Graduate School of Information Science and Technology, University of Tokyo.
††産業技術総合研究所，スタンフォード大学，
National Institute of Advanced Industrial Science and Technology.: Stanford University.
‡国立情報学研究所，総合研究大学院大学，
National Institute of Informatics.: The Graduate University for Advanced Studies.
‡‡株式会社ホットリンク，
Hotto Link, Inc.

*1) http://www.doblog.com/.
データの期間は 2003 年 10 月から 2006 年 6 月まで

$$\vec{P} = (p_1, p_2, \cdots, p_m)$$
$$\vec{Q} = (q_1, q_2, \cdots, q_n)$$

$$A = \begin{array}{c} \\ usr_1 \\ usr_2 \\ \vdots \\ usr_n \end{array} \begin{array}{c} \begin{array}{cccc} tm_1 & tm_2 & \cdots & tm_m \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \end{array}$$

特異値分解は，最小二乗誤差に基づいて行列を 3 つの行列の積で近似する手法であり，行列データを圧縮する方法として因子分解法や Latent Semantic Indexing といった技術にも応用されている．図において，$k = 1$ とすることで情報損失を最小に抑えつつ U, V を一次元の行列（すなわちベクトル）にすることができ，M に情報の行列を代入することで，V を語の影響力 P，U を人の影響力 Q とみなすことができる．これにより定まる各語のスコアから語の順位付けを行うのが，本稿で提案する手法である．

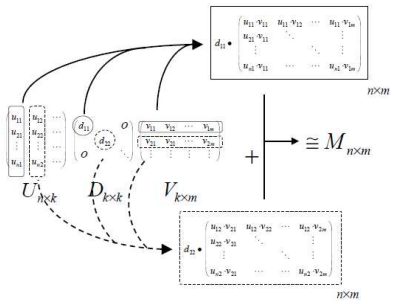

図 3 特異値分解

4. 評価実験

提案手法を Doblog データベースに適用し，その有効性の評価実験をおこなった．順位付け結果のうち，上位のものを表 1に示す．対象とする語には，本手法で抽出される語の性質を把握しやすいよう，**人気語**と**ランダム語**を用意した．前者は日常生活や検索において実際に話題に上った語であり，高順位に抽出されることが望ましいもので，2004･2005 年の Google 検索上位語およびユーキャン流行語大賞選出語から用意した．

4.1. ランダム語と人気語の比較

提案手法において，ランダム語と人気語ではスコアに違いがあるかどうかを調査するため，pairwise accuracy（PA）を計算した．これは対象語を高ランクになるべき語群 C_h とそうでない語群 C_l の 2 通りに分類した場合に，実験結果順位において C_h の語が C_l に属する語よりも上位となっている比率である．結果は約 7 割の正解率となり，人気語として用意した語，すなわち実際に話題に上っていた語は本手法において高ランクになりやすいことが分かる．

4.2. 既存手法との比較

語のランク付けを行う手法は多様に存在するが，本手法での狙いは，単に大規模的に話題になりやすい語だけではなく，出現状況だけでは重要性を観測しづらい語も抽出することである．そこで「出現頻度（記事数）」と「burst」[2]を比較対照とした．

記事数との比較では，上位 10 語のランキングを見ると，提案手法と大きな差は見られなかった．しかし，いずれにおいても上位はほぼ人気語が占めており，記事数によるランク付けがある程度信頼できることに依ると言える．一方で PA の値で比較すると，提案手法が約 9%優れていた．これはより下位での順位において差が出たためである．特に流行語は，ごく一時的に話題になっただけのものも少なくなく，そうした語は記事数では検出できなかったと考えられる．

順位	記事数	Burst	提案（$d = 3$）
1	ラーメン 検	台風 検	台風 検
2	台風 検	地震 検	ラーメン 検
3	地震 検	athens 検	地震 検
4	ガンダム 検	ハウルの動く城 検	切り替え ラ
5	切り替え ラ	震度 ラ	楽天 検
6	楽天 検	クールビズ 検	ガンダム 検
7	ライブドア 検	新規参入 流	ライブドア 検
8	衝動買い ラ	ごくせん 検	震度 ラ
9	自己責任 流	ツールバー ラ	衝動買い ラ
10	マクドナルド 検	愛知万博 検	自己責任 流
PA	58.7%	76.6%	67.6%

表 1 各手法でのランキング上位 10 語

Burst は，PA が優れており，流行語の「クールビズ」（提案手法で 60 位）や話題となったテレビドラマの「ごくせん」（提案手法で 68 位）などの語が高順位になっているのも特徴的である．記事数の少ない語では伝播が起こりづらく，提案手法では上位語として検出できず，差が現れたと考えられる．一方で，burst で 20 位未満，提案手法で 20 位以内に入った 10 語についてその累積記事数の推移を調べると，図 4（左）のようになった．いずれの語も特定の時期にのみ急激に上昇するということはなく，徐々に言及数が増えていくのが分かる．図4（右）で示されるように，burst ではその性質上，突発的な変化のない語を抽出することはできていない．対して提案手法では，そうした瞬発性のある語のほかに，継続的に使用され続けるような語も抽出できているのが特長である．

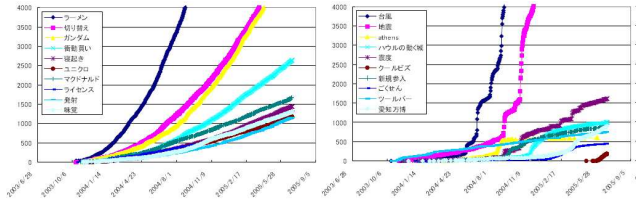

図 4 累積言及記事数の推移

5. まとめ

本稿ではブログにおける話題の伝播が，語の力とブロガの力によって説明できることを前提として，伝播の情報を表す行列を作成し，特異値分解を適用することで，語の影響力を計算して重要語を抽出する手法を提案した．

結果として話題になる語をある程度上位にランク付けすることができ，出現頻度の変化だけでは抽出しづらい語にも対応できたが，一方で絶対的な記事数による影響のためか，必ずしも高い精度は得られなかった．今後の課題として，伝播情報から行列を作成する際に，ブロガや語の特性を考慮することで精度を改善していくことが考えられる．

参考文献

[1] Fukuhara, T., Murayama, T., and Nishida, T.: Analyzing concerns of people using Weblog articles and real world temporal data, The 14th International World Wide Web Conference (2005)

[2] Kleinberg, J.: Bursty and hierarchical structure in streams, In Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1.25 (2002)