2F4-4

ブログ上での話題伝播に注目した重要語抽出

Extracting Key Phrases using Topic Diffusion Process in the Blogosphere

古川 忠延[*1] Tadanobu Furukawa　松尾 豊[*2] Yutaka Matsuo　大向 一輝[*3] Ikki Ohmukai　内山 幸樹[*4] Koki Uchiyama　石塚 満[*1] Mitsuru Ishizuka

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]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

[*3]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

[*4]株式会社ホットリンク
Hotto Link Inc.

In the blogosphere, users release lots of information in various topics. They also write an article about the topic they were attracted by other weblogs with a comment or a trackback. So arguments tends to be easy to propagate. This paper define the term that mentioned on a large scale or constantly as important topic. We assume that the process of diffusion among weblogs consist of the power of terms and the power of bloggers. Then, we propose a ranking algorithm for terms based on how easy it spreads. As a result, we can extract not only the bursty terms but also those that occur by degrees.

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ブログでは，その特徴として，記事が頻繁に更新され，それらが時系列に整理されていることや，コンテンツに自由にアクセスできる点などが挙げられる．そのため，ブログ上では日常的に様々な新しい話題が生まれては，コメントやトラックバック[*1]のつながりを介して多くのブログユーザ（ブロガー）に閲覧され，興味を持ったブロガーによってさらに議論が広まっていく傾向がある．

ブログ上で関心を惹く話題は，その出現の仕方に様々なパターンがある [Fukuhara 05]．そのため，ブログ上では世間で流行として広く認識されているような大規模的な話題ばかりでなく，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，コミュニティ内で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．瞬発性または継続性を持って広まる語（話題を代表する語）はブロガー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扱う．

文書中から重要語を抽出するには，多くの既存研究では語の出現状況に着目してきた．しかし，瞬発性の語を抽出するためには語の出現頻度や出現間隔に注目すればよいが，継続性の語を抽出するには各ブロガー間での実際の伝播に着目する必要がある．よりミクロな視点で観測することで，単に「使用されやすい」だけではない,「広まりやすい」語を取り出すことが可能であると考えられるためである．そこで本稿では，ブロガーが記事を書く前に誰のブログを見ているのかという閲覧情報を用いて語の重要度を計算する手法を提案する．人と語それぞれが影響力を持っていると仮定し，より多くの閲覧者に語を伝播させたブロガー，または多くの閲覧者に伝播した語はより大きな影響力を持っていると考える．そして大きな影響力を持った語を，重要語として抽出しようというものである．なお，実験には，ユーザ間の閲覧情報を扱うことができる，ブログホスティングサービス Doblog[*2] のデータベースを使用する．

以下，まず 2 章において本稿における伝播の定義を説明する．3 章にて提案する手法を説明し，4 章で評価実験と考察を行う．5 章でブログにおける話題抽出に関する既存研究に対する本稿の位置づけを述べ，最後に，6 章にて本稿をまとめる．

連絡先: 古川 忠延，東京大学大学院 情報理工学系研究科 創造情報学専攻，〒 101-0021 東京都千代田区外神田 1-18-13 秋葉原ダイビル 13F，furukawa@mi.ci.i.u-tokyo.ac.jp

*1　過去に他者が書いた記事と関連した内容で記事を書く場合に，引用元の記事に対して通知する機能．

2. ブログにおける語の伝播

本稿における「語 t の U_a から U_b への伝播」は以下のように定義する．

1. ブロガー U_a がある語 t を含む記事を自身のブログに投稿する．
2. U_b が U_a のブログを訪れ，さらに t を含む記事を初めて自身のブログに投稿する，という活動が，(1) からある日数 d 以内に行われる．

ここで，ある二者間での伝播を考えたときに，先に投稿して閲覧された側のブロガー A を先行投稿者と呼ぶこととする．伝播については，さらに以降で説明する条件を満たしているものとし，複数のブログ（先行投稿者）から影響を受けることも認める．例えば，U_b に対して U_a と同様に伝播の条件を満たすブログ U_c が存在していた場合,「U_a と U_c から U_b への伝播」として扱う．

日数 d を制限するのは，特定の記事を閲覧してその影響を受けているかどうかを正確に把握するのは不可能であり，訪問までの時間が長い場合には語を含む該当の記事を読んでいない可能性が考えられ，一方で投稿時期がかけ離れている場合には，同じ語が含まれていても閲覧した元の記事とは話題として異なっている可能性が考えられるためであるこの制限期間につ

*2　http://www.doblog.com/

いては場合分けをして実験を行うことで，ユーザ間での話題伝播が何日程度で起こっているのかを検証するものとする．

3. 重要度計算手法

本稿では前提として，ブログ上における語の伝播が

- 語の影響力
- ブロガーの影響力

の値によって説明できるとする．そして「語の影響力」と「ブロガーの影響力」はそれぞれ，語を使用するブロガー，対象とする語に依らず静的に定まるものとする．つまりこの二つの影響力によって，語の伝播の仕方が決定されるものと考える．そこで，伝播の定義に従って解析データベースから伝播情報を抽出して語の影響力を計算，これを語の重要度として扱うのが，本稿で提案する重要度計算手法である．

伝播情報としては,「どの語が」「誰から」「誰に」伝播したかというデータで取得することができる．これにさらに各ブロガー・語に関する特徴を加味することで，伝播情報は多様な形式で表現することが可能であるが，本稿では単純化のため,「あるブロガーが」「ある語を」伝播させた人数（＝語を誰が何人に伝播させたか）というデータで扱う．

前提として述べたとおり，語の伝播の挙動を決定するような，一意に定まる語・ブロガーの影響力の存在を仮定すると，対象とする語の数を m 個，ブロガーを n 人として，それらはそれぞれ 1, 2 で示すようなベクトル $\vec{P}$, $\vec{Q}$ で表現できる．p_i は各語の影響力，q_j は各ブロガーの持つ影響力である．

$$\vec{P} = (p_1, p_2, \cdots, p_m) \quad (1)$$

$$\vec{Q} = (q_1, q_2, \cdots, q_n) \quad (2)$$

一方,「各人が各語を何人に伝播させたか」の情報は，行を語に関する要素，列をブロガーに関する要素として，式 3 に示す行列 A として表現することができる．例えば，a_{12} は，語 tm_2 をブロガー blg_1 が伝播させた人数を表す．

$$A = \begin{array}{c} \\ blg_1 \\ blg_2 \\ \vdots \\ blg_n \end{array} \begin{array}{c} \begin{array}{cccc} tm_1 & tm_2 & \cdots & tm_m \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \end{array} \quad (3)$$

この行列 A が，語の影響力とブロガーの影響力によって定まるというのが本稿の仮定である．

A において，各行ベクトル $(a_{j1}, a_{j2}, \cdots, a_{jm})$ は伝播における語の振る舞いを，各列ベクトル $(a_{1i}, a_{2i}, \cdots, a_{ni})$ はブロガーの振る舞いを，それぞれ表すものである．本稿で提案する手法ではこの点に注目して，特に語の影響力 $\vec{P}$ を算出する．

3.1 特異値分解

行列からベクトルを抽出するメソッドとしては，本稿では特異値分解を用いる．特異値分解は行列を 3 つの行列の積で近似する手法であり，Latent Semantic Indexing [Deerwester 90] や因子分解法 [Tomasi 92] などにおいても，行列データのクラスタリング・圧縮の用途に用いられている．特異値分解は式 5 で表される．但し，$1 \leq k \leq min(x, y)$ であり，$\vec{u_i}$ は U の列ベクトル，$\vec{v_i}$ は V^t の行ベクトルである．

$$M_{x\times y} \simeq U_{x\times k} \cdot D_{k\times k} \cdot V^t_{k\times y} \quad (4)$$

$$= (\vec{u_1}, \cdots, \vec{u_k}) \cdot D_{k\times k} \cdot (\vec{v_1}, \cdots, \vec{v_k})^t \quad (5)$$

特異値分解では，分解後の行列のうち $i \leq k$ 次元までを用いた場合の積 $U_{x\times i} \cdot D_{i\times i} \cdot V_{i\times y}$ が，もとの行列に対して最小二乗誤差に基づいた近似となるように，分解を行う．式 5 において，行列 U の i 番目の列ベクトル $\vec{u_i}$ は，第 i 特異値 d_{ii} に対応する左特異ベクトルであり，分解前の行列 A の各列ベクトルが張る空間の正規直交ベクトルとなる．行列 V^t も同様の性質をもち，右特異ベクトル $\vec{v_i}$ は，A の各行ベクトルが張る空間の正規直交基底である．D は対角行列であり，左上から右下に向けて降順に特異値が並んでおり，各特異値 d_{ii} は，i 番目の左・右特異ベクトルの積が近似において寄与する度合いを表すものである．すなわち，i の値がより若い特異ベクトルほど，M における列・行の属性の性質をより強く表すようなベクトルとなっている．

3.2 伝播情報からの重要度計算

伝播情報を表す行列 A に特異値分解を適用することで，式 4 における U として，伝播におけるブロガーの影響力に関する特徴を表す行列（複数の列ベクトル群），V として語の影響力を表す行列（複数の行ベクトル群）を得る．特に特異値分解の定義から，$k = 1$ に対応する両特異ベクトル $\vec{u_1}, \vec{v_1}$ が，それぞれ語の影響力，ブロガーの影響力の特徴を最も強く表すベクトルとなる．すなわち，

$$\vec{Q} = \vec{u_1}$$

$$\vec{P} = \vec{v_1}$$

この $\vec{P}$ が，語の伝播における重要度を決定するものであり，本稿ではそのスコアが高い語を重要語として抽出する．

4. 評価実験

提案手法を Doblog データベース（データの期間は 2003 年 10 月〜2005 年 6 月）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 2 に示す[*3]．なお，対象データの詳細は以下の通りである．

- 語：本手法で抽出される語の性質を把握しやすいよう，1 で示すように，ランダム語と人気語（検索上位語と流行語）を用意した．人気語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，本稿で定義する瞬発性もしくは継続性を持った語である可能性が高い．
- ブロガー：上記の対象語のうちの，いずれか一つ以上の語の伝播に関わっている約 1,000 ユーザ．
- 伝播の条件における制限期間 d：話題の伝播として有効な期間を検証するため，1〜20 日の間でそれぞれ実験を行う．

4.1 ランダム語と人気語

提案手法において，ランダム語と人気語では順位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．PA は，対象語を高ランクになると仮定した語群 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

*3 本実験において提案手法によって各語に定まるポイント（語の影響力）は，実際には全て負の値（または 0）であった．しかし，対応する人の影響力もまた同様に負であり，本稿における各語の影響力を見る上では，符号を反転させても本質的な相違はないため，結果の表では正の値として示すこととする．

*9 http://www.google.com/intl/en/press/zeitgeist.html

*10 http://www.jiyu.co.jp/singo/

表 1: 評価実験に用いた語の種類．括弧内は語数．

種類		説明
ランダム語（69 語）		Doblog 内で使われている語からランダムに選んだ中程度頻度の語．
人気語（41 語）	検索人気語（28 語）	2004 年・2005 年月別の Google 検索頻度上位語[*4]．
	流行語（13 語）	2004 年・2005 年ユーキャン流行語大賞受賞上位語[*5]．

表 2: 各ランキング手法に基づくランキング上位 10 語と PA．語の右上の文字は各々，検：検索人気語，流：流行語，ラ：ランダム語であることを表す．手法「無作為」における PA は，語を無作為に順位付けした場合の理論値．

順位	提案 ($d=2$)	提案 ($d=20$)	記事数	Burst	無作為
1	台風 検	ラーメン 検	ラーメン 検	台風 検	
2	ラーメン 検	台風 検	台風 検	地震 検	
3	地震 検	地震 検	地震 検	athens 検	
4	切り替え ラ	切り替え ラ	ガンダム 検	ハウルの動く城 検	
5	楽天 検	楽天 検	切り替え ラ	震度 ラ	
6	震度 ラ	ガンダム 検	楽天 検	クールビズ 検	
7	ガンダム 検	ライブドア 検	ライブドア 検	新規参入 流	
8	ライブドア 検	衝動買い ラ	衝動買い ラ	ごくせん 検	
9	衝動買い ラ	震度 ラ	自己責任 流	ツールバー ラ	
10	自己責任 流	自己責任 流	マクドナルド 検	愛知万博 検	
PA	67.7%	66.5%	58.7%	76.6%	(50%)

順位において C_h の語が C_l の語よりも上位となっている比率を，精度として表す指標である．語 x について，手法でのスコアを $S(x)$（高順位ほど大），正解ランク（本来期待されるランキング）でのスコアを $T(x)$ としたとき，式 6 で表される [Richardson 06]．本実験では，ランダム語と人気語の間の違いを測定するため，ランダム語を C_l，人気語を C_h として PA を計算するものとする．

$$\begin{aligned} T_t &= \{x, y : T(X) > T(y)\}, \\ S_t &= \{x, y : S(X) > S(y)\} \\ \Rightarrow PA &= \frac{|T_t \cap S_t|}{|T_t|} \end{aligned} \tag{6}$$

結果の一部を表 3 に示す．d をいずれの値に設定した場合も 67%程度の正解率となり，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 話題の伝播の期間

では，本稿で定義する話題の伝播は，どの程度の期間内で起こるものだろうか．前述のとおり，PA を d ごとに比較した場合，大きな変動はないものの，$d = 2$ 付近で最大値をとり，$d = 8, 9$ で一度上昇する他は，時間とともにほぼ減少していく傾向があった．長いスパンで定義することによって，「伝播させた」として扱われるデータの量が増えるのに対して，人気語の順位が下がっていくことを表しており，このことは，少なくとも人気語の伝播は 2 日以内に起こるものであり，それ以上長いスパンで見た場合には，閲覧者が単に同じ語を含む違う話題で記事を投稿している可能性が高いことを示している．これは，「台風」や「地震」のような瞬間的な話題については明らかである．

また，本稿での伝播と類似した性質を持つトラックバックにおいても，投稿から 2 日目までに全体の約半数が行われていることが分かった．そして 2 日前後までに行われているトラックバックの大半は，ニュース系のサイトで見つけた話題について互いに議論しているものや，ウェブ上での性格診断のような占い，など，話題として連なっているものに対して使われていた．しかし時間が経つに連れて，検索で見つけた記事に対して「参考にさせてもらった」という意味合いで使われているケースや，もしくは自身のブログ内の関連のある記事へのものが多く見られるようになっていった．こうしたトラックバックは「他者の影響を受けて記事を書く」という伝播とは異なる行動であり，話題の伝播がより短いスパンで行われていることを指示するものである．

4.3 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．そのため，以下の二つの既存手法を比較対照とした．

4.31 出現頻度（記事数）との比較

語の出現状況を使用する場合，その頻度を数える方法が簡単であるが，ブログの各記事の長さは一様ではなく，その影響を受けてしまう可能性がある．そこで本稿では，一つの記事内では語が何度出現しても頻度は 1 としてカウントし，すなわち語を含む記事数を用いた．

記事数の多い語ほど高ランクとして順位付けを行った結果は表 2 に示すとおりである．まず，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．11 位以下の語のうちで両手法の間で 20 位以上の差がついた語は表 3 のとおりであり，一部の人気語で順位に大きな差が開いたためである．但し，このように提案手法では記事数の

表 3: 人気語の順位．灰色の行は，提案手法における順位の方が低い語．

語	提案 ($d = 2$)	記事数
ハウルの動く城	12	36
愛知万博	18	38
Winny	25	66
athens	29	52
六本木ヒルズ	42	14
ウォーターボーイズ	49	71
冬のソナタ	55	33
クールビズ	59	88
冬ソナ	61	23
チョー気持ちいい	66	103
中二階	79	105

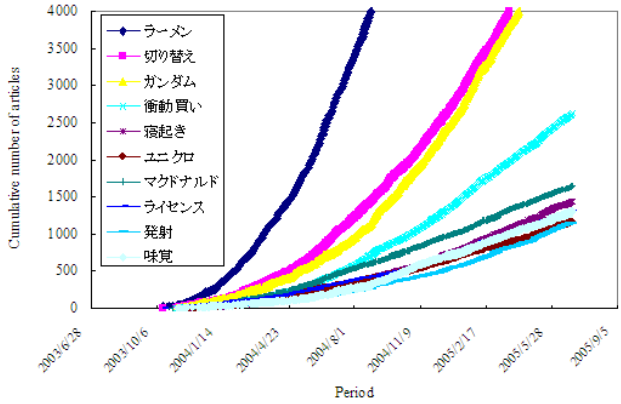

図 1: 提案手法で burst より上位の語の累積出現頻度の推移

あまりない語についても順位が改善されているが，必ずしも上位に抽出できているとは言えない．これは，絶対的に記事数が少ないために伝播件数が少なくなってしまうことに起因していると考えられ，今後手法を改善していく必要がある．

4.32　burst との比較

burst [Kleinberg 02] は時系列解析によって，語の流行を検出する手法であり，瞬間的な語の出現頻度の上昇度合い示すスコア（burst 度）を得ることができる [Fujiki 04]．本稿では各語における burst 度のうち最大の値を語の重要度と捉え，順位付けに利用した．burst によるランキングを表 2 に示す．

burst を利用したランキングでは，PA が提案手法・記事数のものと比較して大きく優れていた．流行語の「クールビズ」（提案手法で 60 位，記事数で 88 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，記事数で 63 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 1 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．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難しく，対して提案手法では，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るのが特長であると言える．

5.　まとめ

本稿ではブログにおける話題の伝播が，語の力とブロガーの力によって説明できることを前提として，伝播の情報から議論の連なりやすい語を重要語として抽出する手法を提案した．瞬発性や継続性を持つ語を重要語として定義することで，規模に依らず話題性のある語を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．しかし，絶対的な記事数による影響や，一般的に使用されやすい語を伝播として扱ってしまう問題点も含んでおり，精度においては改善の余地が見られた．

今後の課題として，ブログにおける話題普及の性質をより詳細に把握することが挙げられる．ブロガーや語の特性を考慮することで，伝播情報をより正確に表現することができれば，精度を改善できるはずである．

謝辞

Doblog データベースは株式会社 NTT データおよび株式会社ホットリンクよりご提供をいただきました．記してお礼申し上げます．

参考文献

[Deerwester 90] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A.: Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391–407 (1990)

[Fujiki 04] Fujiki, T., Nanno, T., Suzuki, Y., and Okumura, M.: Identification of Bursts in a Document Stream, in *Workshop on Knowledge Discovery in Data Streams* (2004)

[Fukuhara 05] Fukuhara, T., Murayama, T., and Nishida, T.: Analyzing concerns of people using Weblog articles and real world temporal data, in *2nd Annual Workshop on the Weblogging Ecosystem* (2005)

[Kleinberg 02] Kleinberg, J.: Bursty and hierarchical structure in streams, in *Proc. 8th ACM SIGKDD* (2002)

[Richardson 06] Richardson, M., Prakash, A., and Brill, E.: Beyond PageRank: machine learning for static ranking, in *Proc. WWW 2006* (2006)

[Tomasi 92] Tomasi, C. and Kanade, T.: Shape and Motion from Image Streams: a Factorization Method, Full Report on the Orthographic Case, Technical Report CMU-CS-92-104, Carnegie Mellon University (1992)