3D2-2

Weblog間の話題伝播過程に注目した重要トピックの抽出

LaTeX Style file for manuscripts of JSAI 20XX

古川 忠延[*1] Tadanobu Furukawa
松澤 智史[*2] Tomofumi Matsuzawa
松尾 豊[*3] Yutaka Matsuo
大向 一輝[*4] Ikki Ohmukai
内山 幸樹[*5] Koki Uchiyama
武田 正之[*2] Masayuki Takeda

[*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]東京理科大学 理工学部
Tokyo University of Science

[*3]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

[*4]国立情報学研究所
National Institute of Informatics

[*5]株式会社ホットリンク
hottolink,Inc.

On weblogs (blogs), users release lots of information in various topics everyday and write an article about same topic as one which they were attracted by reading it. So arguments tends to be easy to spread out. This paper extracts the terms which have big power as important topics on the assumption that the process of diffusion on blogs consist of the power of a term and the power of a blogger. After making a matrix including the information about existence of terms and users' browsing behavior, we treat that as a numerical model by applying the singular value decomposition and validate it.

1. はじめに

Web における情報発信の一形態として近年注目されている Weblog（以下 Blog）では，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などの機能をもつなどの点が挙げられる．このような特徴により，Blog 上では日常的に様々な新しい話題が生まれては，閲覧により他のユーザに伝播し，コメントやトラックバックも通じて議論が広まっていく傾向がある．こうした様子を分析すること，例えばトレンドやオピニオンリーダーを抽出することは，情報の効率的な伝達を実現する上で重要であり，マーケティングの分野において注目されている．

本稿では，Blog ホスティングサービス Doblog[*1] のデータベースを使用することで，話題の普及についての解析を行う．本稿においては話題を“ 語 ”で定義し，Blog のアクセス情報と記事内の語の出現状況から話題の伝播を抽出する．そして，伝播が「語の力」と「ユーザの力」によって成立しているという前提のもと，大きな力を持つ話題を重要トピック（語）として抽出する手法を提案する．

本稿の構成は以下の通りである．まず 2 章において，伝播に関する既存研究に対する本稿との位置づけを述べる．3 章にて提案する手法を説明し，4 章で実際の解析を行い，5 章において本手法の問題点に対する議論を行う．最後に，6 章にて本稿をまとめる．

2. 関連研究

Web における情報伝播に関する研究は，その時間的変化に着目する必要があることから限られたフィールドでの分析が多く行われてきた．例えば，松村らは，電子掲示板におけるコメントの依存関係からユーザの発言の内容や影響力などを抽出することによる普及モデルの提案 [松村 02] やユーザのプロファイリング [松村 03] などを行っている．

近年台頭してきた Blog においても，イノベーション普及モデルを利用による情報伝達の分析 [Gill 05] が行われている他，その時系列性・伝達性を利用することによる研究が数多く行われている．BlogWatcher[*2]は，Web 上をクローリングして Blog データを収集し，そこから話題の Burst 度 [Kleinberg 02] や評判情報などを抽出して提示するサービスである [Nanno 04]．BlogPulse[Glance 04] もまた同様に，トピックの人気度やトレンドの推移などを示すものである．

記事の内容や依存関係に注目した分析としては，Adar らは記事内で言及している内容のほか，参照している URL についてそのタイミングに着目することで，情報が Blog ネットワーク上でどのように推移しているのかを調べ，図示するといったものもある [Adar 05, Adar 04]．また，Blog 検索サービス BlogRanger[*3]に実装されている Blog の採点アルゴリズム EigenRumor[Fujimura 05] は，記事の参照関係を利用して“ 良いブロガー ”を定義している．

本研究は，Blog 記事内で使用されている語句に着目してユーザ間での伝播を解析するという点では，上記の研究と類似しているが，Doblog データベースを使用することによってユーザの閲覧情報を導入できるのが特長である．そのため，異なるブロガーが同様の記事を書いていた場合でも，伝播によるものとそうでないものを分けて分析することが可能である．そして結果として，伝播情報から伝わりやすい語を重要語として抽出していく．

3. 提案手法

本稿では前提として，Blog 上における話題の伝播が

- 語の伝播力
- 人の伝播力

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 http://www.doblog.com/，(株)hottolink，(株)NTT DATA

∗2 http://blogwatcher.pi.titech.ac.jp/

∗3 http://ranger.labs.goo.ne.jp/

の値によって説明できるとする．そしてこの 2 つの伝播力によって「誰が何人に伝播させるか」が決定されるものとする．

解析データである Doblog データベースでは，この伝播させた人数について扱うことができるため，本稿では，この情報から逆に語の伝播力（および人の伝播力）を抽出していく．本節ではその手法を説明する．

3.1 伝播の定義

本稿における“ 話題の伝播 ”は，以下のように定義する．但し，ユーザ U_a, U_b は，それぞれ Blog B_a, B_b の管理者とする．

1. U_a がある語 t を含む記事を自身の Blog に投稿する．
2. 1. から x 日以内に U_b が B_a を訪れる．
3. 2. から y 日以内に U_b が t を含む記事を**初めて**自身の Blog に投稿する．

以上の過程が成立しているとき，語 t は「B_a から B_b へ伝播した」と定義する．すなわち，U_b は U_a を読むことによって影響を受け，同じ話題で記事を書いたと見なすものである．また，複数の Blog から影響を受けることも認めることとし，例えば，B_b に対して B_a と同様に上記条件を満たす Blog B_c が存在していたとき,「B_a と B_c から B_b へ伝播した」と扱う．

また，条件に含まれる x・y についてはパラメータとして扱い，この値が本手法に与える影響については実験で検証を行っていく．このようなパラメータを定義したのは，x については，ユーザの訪問履歴は記事単位ではなく Blog 単位でしか抽出できず，何日以内にアクセスすればその語を含む記事を目にするのか分からないため，y については，記事を読んでから何日程度ユーザはその影響を受ける可能性があるのかが不明なためである．

3.2 伝播力の抽出

前提として述べたとおり，語の伝播力と人の伝播力が存在すると仮定すると，対象とする語の数を m 個，ユーザを n 人として，それぞれ以下のようなベクトル $\vec{P}, \vec{Q}$ で表現できる．

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

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

一方,「各人が各語を何人に伝播させたか」の情報は，行をトピックに関する要素，列をユーザに関する要素として，下に示す行列 A として表現することができる．

$$A = \begin{array}{c} \\ T_1 \\ T_2 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{cccc} U_1 & U_2 & \cdots & U_n \end{array} \\ \begin{pmatrix} k_{11} & k_{12} & \cdots & k_{1n} \\ k_{21} & k_{22} & \cdots & k_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m1} & k_{m2} & \cdots & k_{mn} \end{pmatrix} \end{array} \quad (3)$$

前提より，A は $\vec{P}$ と $\vec{Q}$ で表現できるので,

$$A = f(\vec{Q} \cdot \vec{P}^t) \quad (4)$$

すなわち，行列から 2 つのベクトルを抽出できれば，それらが語・人の伝播力を表すベクトルと捉えることができるはずである．

ここで，本稿ではこのベクトル抽出のメソッドとして特異値分解 [中川] を用いる．特異値分解は，ある行列を最小二乗誤差に基づいて 3 つの行列の積で近似する手法である．これを用いることによって，行列 A は以下のように変形できる．但し，$U^tU = I$, $V^tV = I$ であり，D は左上が最大で以降右下に向けて降順に値の並ぶ対角行列である．

$$A_{m\times n} \simeq U_{m\times k} \cdot D_{k\times k} \cdot V^t_{n\times k} \quad (5)$$

この式において $k = 1$ とすることで,

$$A_{m\times n} \simeq U_{m\times 1} \cdot D_{1\times 1} \cdot V^t_{1\times k} \quad (6)$$

$$= \vec{Q} \cdot x \cdot \vec{P}^t (x \text{ は定数}) \quad (7)$$

と，ベクトル $\vec{P}, \vec{Q}$ が抽出できる．

4. 解析

提案手法を Doblog の実データに適用して解析を行った．

4.1 パラメータ

各パラメータは以下のとおりである．

- 語数 m：100（個）
- ユーザ数 n：10651（人）
- 伝播定義のパラメータ x, y：3, 7, 10（日）

語は，全 Doblog ユーザの 5%にあたる 2,648 ユーザ[*4]をランダムに選び，その Blog 内全ての記事から，中頻度の語をランダムに選んだ．そしてユーザ数は，その 100 語のうちいずれか一つでも書いたことのあるユーザである．伝播定義のパラメータについては，x, y について 3 種類ずつ用意し，その組み合わせ 9 パターンについてどのような違いが現れるのか検証する．

4.2 結果

図 1 が，本稿が提案する手法に基づいて語の伝播力を求めたものである．凡例の“ 3x3 ”や“ 7x10 ”などは，伝播の定義におけるパラメータが，それぞれ $(x, y) = (3, 3), (7, 10)$ となっていることを示す. (“ 3x3w ”については後述.)

まず，(x, y) についての検証であるが，いずれの値においても，多少の値のずれはあるが，語間の大小関係にはほとんど違いは見られなかった．今回用意した 9 パターンのパラメータに関しては，厳密に吟味する必要はないのではないかと考えられる．

一方，語の伝播力については“ ツールバー ”(Doblog ツールバー．Blog 更新や RSS チェッカーなどの機能を持つソフトウェア）や“ 日本ハム ”(プロ野球）のような，実際に話題として伝播しているものも高い値を持っているが,“ 切り替え ”,“ 出没 ”,“ 不愉快 ”などといった特定のトピックとして伝播しているわけではない，一般的な語句も多く上位に来てしまっているという問題が見られる．

5. 議論

前節で示した結果では，トピックとして伝播していない語が上位に来てしまう，また，トピックとして伝播していても，特定の語句のみが非常に高い値となる一方で，他の語は期待通りの値を示していないなどといった問題点があり，語のランク付けという点からはあまり意味をなさない手法であると言わざるをえない．そこで，本節ではこの点について改良を試みる．

*4 データベースダンプ時点で 52,976 ユーザ

5.1 語に対する重み付け

前述のような問題が起こる理由として,「誰もが」「日常的に」使うような語は，内容として伝播していなくても，本稿での伝播の定義を満たしてしまうという点が挙げられる．例えば“ 不愉快 ”のような語は,「(政治のことで）不愉快に思う」と書かれた記事を見たユーザが，数日後,「(仕事上）不愉快なことがあった」と書いたという場合でも，伝播の定義を満たしてしまっていた．

このような場合に伝播力としての値が上昇することを防ぐため，語ごとに出現頻度の変化の度合いを利用した重み $w(t)$ を計算し，伝播人数を示す行列 A の各行に付加した．

$$w(t) = \frac{\sum_{i=1}^{9} |tf_i(t) - tf_{i-1}(t)|}{tf_{all}(t)} \quad (8)$$

$tf_n(t)$ $(0 \leq n \leq 9)$ は，Doblog 上で語 t が最初に出現した時点から最後に出現した時点までの期間を 10 等分し，各区間における出現数を表し，$tf_{all}(t)$ は全体での出現数を表す．伝播により他者の影響で語を使えば分子の値が大きくなり，また，局所的にのみ伝播している（出現頻度が少ない）場合にも対応するために分母を設けている．

結果は図 1 の“ 3x3w ”の項目である．(伝播定義のパラメータに関しては，ここでは x, y とも 3 を採用した.)“ 出没 ”,“ 切り替え ”,“ 不愉快 ”などの問題となっていた語の値を抑えることには成功している．しかし,“ ツールバー ”のような，実際に伝播している語の値も下がっており，その他の語も総じて 0 に近い値となってしまった．原因としては，重みの分母の項が効き過ぎてしまっていることが考えられ，今後より適切な重みを考える必要がある．

5.2 2 次元以上での検証

一方，これまでは，特異値分解の式 (5) において $k = 1$ として，すなわち，各語・各ユーザの伝播力を 1 次元の値として扱ってきたが，では $k = 2$ 以上で同様の分析を行うとどうなるであろうか．

仮に $k = 3$ とした場合，語・ユーザの伝播力 P, Q はそれぞれ $m \times 2, n \times 2$ の行列 (9),(10) で表されることになる．

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ p_{31} & p_{32} & \cdots & p_{3m} \end{pmatrix} \quad (9)$$

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ q_{31} & q_{32} & \cdots & q_{3n} \end{pmatrix} \quad (10)$$

これらの行列において，1 行目のみを伝播力ベクトルとして用いて元の行列 A を近似したものがこれまで述べてきた分析であるが，2 行目以降も用いることでより A を近似できるのが特異値分解の性質である．

そこで，$k = 2$ として 2 次元目まで扱った場合の結果が図 2 である．大抵の語では 1 次元目と 2 次元目の正負が逆転している程度であるが,“ メルセデス ”,“ 震度 ”,“ 日本ハム ”などでは 1 次元目・2 次元目ともに正となっている．特異値分解の性質より，1 次元目 >2 次元目 $> \cdots >$ k 次元目の順に元の行列の特性を表す因子となっていることから，1 次元目で行列 A の大まかな特徴を表す一方で，2 次元目ではより細かく近似するための要素が含まれていると考えられる．そのため，前述の 3 語には共通の特徴がある可能性があり，実際，この 3 語はいずれも話題として伝播している語であった．このことから，より高次元についても検証していくことも重要であると考えられ，またそれにより，語のクラスタリングのようなことも可能となるかもしれない．

6. まとめ

本稿では，Doblog データを用いることで Blog における話題の伝播を定義し，伝播しやすい語を重要トピックとして抽出することを試みた．話題の伝播が語の力とユーザの力によって説明できることを前提とすることで，伝播の情報を行列で表し特異値分解を適用することで，それぞれの力をベクトル形式で表現することができた．しかし，抽出された伝播力は，実際に伝播していないものもノイズとして含んでしまうような問題点から，必ずしも有効な値を示すものとはならなかった．伝播の定義の曖昧性や行列の作成方法など，手法において改善の余地は以前多く存在しており，今後も検証していく必要がある．

参考文献

[Adar 04] Adar, E., Zhang, L., Adamic, L. A., and Lukose, R. M.: Implicit Structure and the Dynamics of Blogspace, *The 13th International World Wide Web Conference* (2004)

[Adar 05] Adar, E. and Adamic, L. A.: Tracking Information Epidemic in Blogspace, *Web Intelligence 2005* (2005)

[Fujimura 05] Fujimura, K., Inoue, T., and Sugisaki, M.: The EigenRumor Algorithm for Ranking Blogs, *The 14th International World Wide Web Conference* (2005)

[Gill 05] Gill, K. E.: Blogging, RSS and the Information Landscape: A Look at Online News, *The 14th International World Wide Web Conference* (2005)

[Glance 04] Glance, N., Hurst, M., and Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs, *Workshop on the Weblogging Ecosystem, The 13th International World Wide Web Conference* (2004)

[Kleinberg 02] 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)

[松村 02] 松村 真宏, 大澤 幸生, 石塚 満：テキストによるコミュニケーションにおける影響の普及モデル, 人工知能学会論文誌, Vol. 17, No. 3, pp. 259–267 (2002)

[松村 03] 松村 真宏, 大澤 幸生, 石塚 満：影響の普及モデルに基づくオンラインコミュニティ参加者のプロファイリング, 人工知能学会論文誌, Vol. 18, No. 4, pp. 165–172 (2003)

[中川] 中川 裕志：Latent Semantic Indexing, `http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/infoDB/ir-lsi.pdf`

[Nanno 04] Nanno, T., Suzuki, Y., Fujiki, T., and Okumura, M.: Automatic Collection and Monitoring of Japanese Weblogs, *The 13th International World Wide Web Conference* (2004)

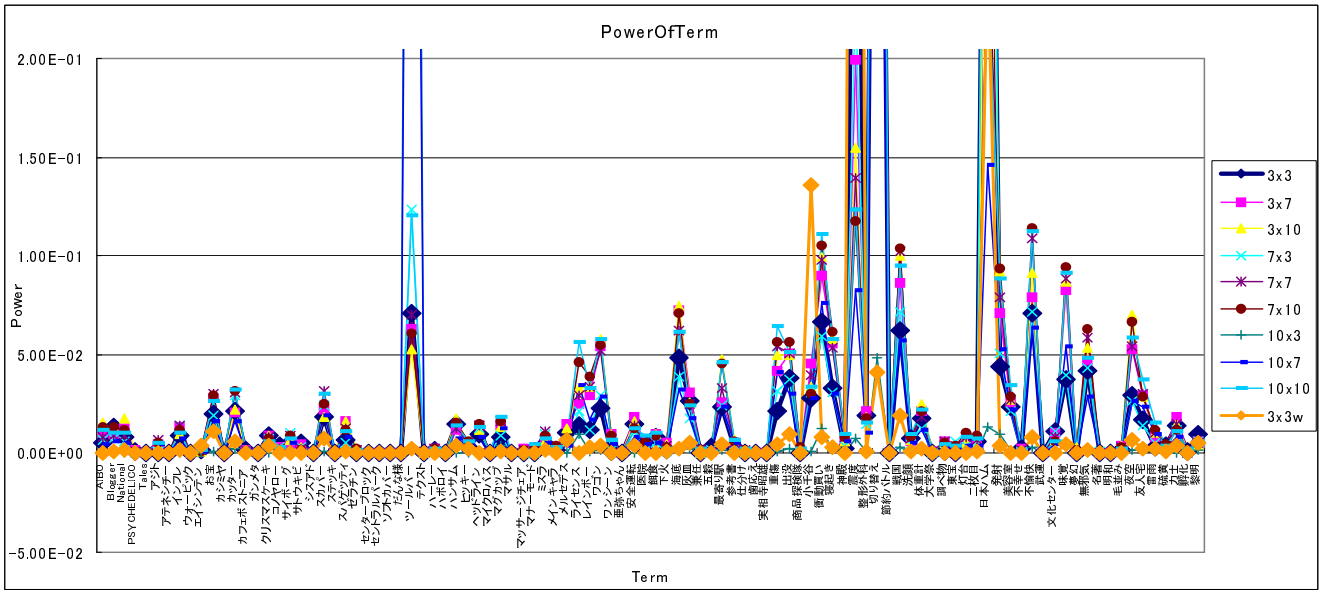

図 1: 語の伝播力.

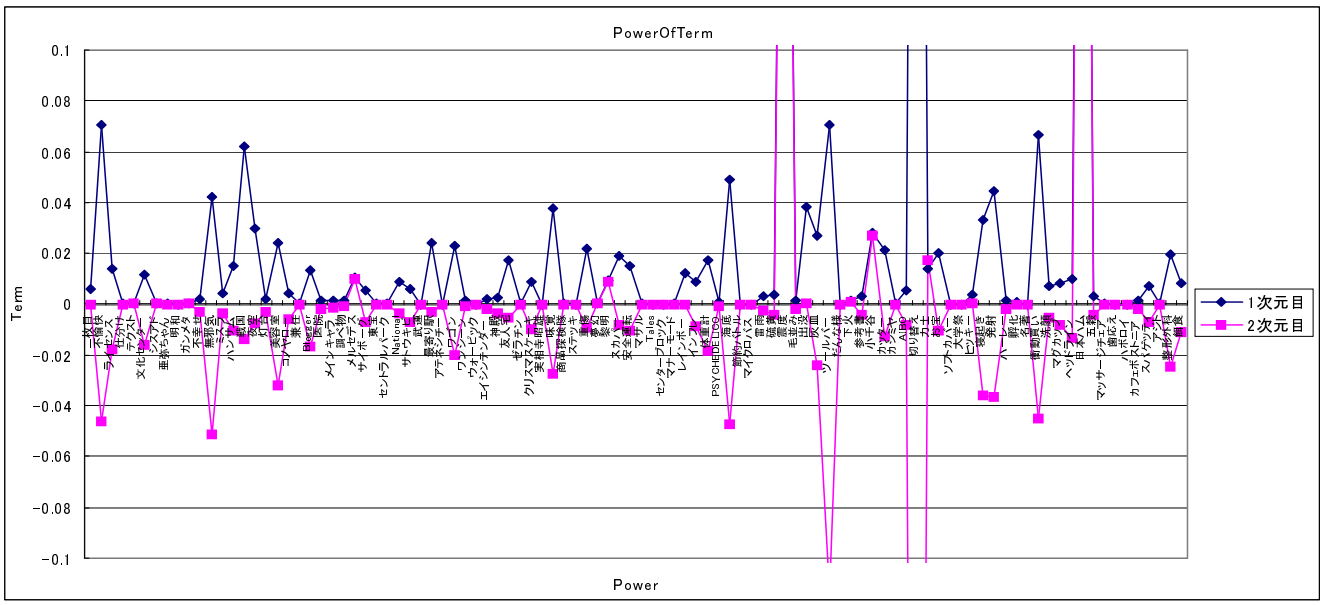

図 2: 複数次元での語の伝播力．但し，伝播のパラメータ x, y はいずれも 3.