IDM! 水玉潰し

IDM! Splash Back

松村 真宏*1
Naohiro Matsumura

*1大阪大学大学院経済学研究科
Graduate School of Economics, Osaka University

This paper defines *Gross Community Influence* (GCI) as an index of dynamic community activity. To predict the movement of topics in a community, promising messages, users, and terms that quasi-maximize GCI are explored. Proposed approach estimates incremental influence by extending IDM algorithm with a slack variable to realize loose term propagation. Preliminary experiments show expected results and future tasks.

1. はじめに

インターネットの利用者が 9000 万人を越えた現在 [総務省 09]，我々の生活の隅々にまでインターネットからの情報が浸透している．そのような氾濫する情報をより分ける手段の一つとして，コミュニティによって精錬された情報が広く利用されている．ソーシャルブックマークやクチコミサイト，サーチエンジンの検索結果などはその代表的な例であり，知らず知らずのうちにコミュニティから大きな影響を受けている．この影響の動向をいち早く予測することができれば，コミュニティという得体の知れない場のダイナミズムが見えてくる．

本研究では，影響伝播モデル IDM を用いてコミュニティが蓄えている総影響量 GCI (Gross Community Influence) を定義し，GCI を成長させるための道標を導くことに取り組む．具体的には，各種制約下（語の種類および量）のもとで GCI を準最大化する条件（どのメッセージにどの語を投下するか）を求める問題に取り組む．

2. IDM

IDM は，ネットワーク上を伝播する語に基づいて，ネットワークの構成要素（メッセージ，投稿者，語）の影響量を定量化するアルゴリズムである [松村 02, 松村 08a, 松村 08b, Matsumura08]．ネットワークのノードメッセージ，リンクはメッセージ間の返信関係やブログのリンク・トラックバックなどによって定義される．メッセージは投稿時刻順にソートされており，リンクは必ず過去のメッセージに向いているので，このネットワークは非循環有向グラフ (DAG; Directed Acyclic Graph) となる．以下で IDM について簡単に説明する．

まず，メッセージ x がメッセージ y に及ぼす影響量 $i_{x\to y}$ を定義する．これには様々な指標を用いることができるが，例えば以下のような指標が提案されている．

$$i_{x\to y} = \sum_{y=x}^{x+r} \frac{\|\bigcap_{n=x}^{y} m_n\|}{\|m_y\|} \tag{1}$$

$$i_{x\to y} = \sum_{y=x}^{x+r} \left\| \bigcap_{n=x}^{y} m_n \right\| \tag{2}$$

連絡先: 松村真宏，大阪大学大学院経済学研究科，〒 560-0043
豊中市待兼山町 1-7，matumura@econ.osaka-u.ac.jp

m_n はメッセージ n に含まれる語の集合，$\|\cdot\|$ は集合 $\cdot$ の要素数とする．式 (1) は伝播した語の割合を足し合わせる指標（[松村 02] を一部改変したもの），式 (2) は伝播した語を足し合わせる指標（[松村 08a]）である．語の伝播範囲は r で調整できる．本稿では式 (2) の定義を用いることにする．

この $i_{x\to y}$ を用いれば，メッセージ x の影響量 I_x は以下の式 (3) から求まる．

$$I_x = \sum_{y\in\{messages\ followed\ by\ x\}} i_{x\to y} \tag{3}$$

y は x を起点とするメッセージの連なり（メッセージチェーン）に含まれるメッセージである．

投稿者 s の影響量 J_s は，s の投稿したメッセージの影響量の総和になるので，以下の式 (4) から求まる．

$$J_s = \sum_{x\in\{messages\ sent\ by\ s\}} I_x \tag{4}$$

また，$i_{x\to y}$ を用いれば，メッセージ x に含まれる語 w の影響量 $k_{x,w}$ は以下の式 (5) から求まる．

$$k_{x,w} = \frac{I_x}{\|m_x\|} \tag{5}$$

$k_{x,w}$ を全てのメッセージについて足し合わせると，そのネットワークにおける語の影響量 K_w が求まる．

$$K_w = \sum_{x\in\{all\ messages\}} k_{x,w} \tag{6}$$

このような IDM の演算は，ネットワークの構造に沿って語を重み付けしながら足し合わせていることに等しいので，一種の畳み込み演算を行っていると見なすことができる．

なお，影響量の受けやすさを表す被影響量 [松村 08b]，伝播語の連鎖ネットワーク [Matsumura08]，影響量の期待値 [松村 08b] も求めることができるが，本稿では省略する．

3. GCI 最大化問題

2. ではメッセージ，投稿者，語のそれぞれの影響量を求めたが，それらの総和はいずれも等しい．本稿では，その総和をコミュニティの蓄えている総影響量 GCI と定義する．

$$GCI = \sum I_x = \sum J_x = \sum K_w \tag{7}$$

表 1: 影響量の変化．I, J, K は $\xi = 1$ のときの影響量，I', J', K' は $\xi = 2$ のときの影響量．

記事 ID	I	I'	∇I	投稿者（http://以降）	J	J'	∇J	語	K	K'	∇K
136617290	8	157	149	***.seesaa.net/	21	251	230	資生堂	357	502	145
159741437	57	175	118	www.blog-headline.jp/***/	68	186	118	ＣＭ	589	724	135
137707887	142	210	68	onlyone.air-nifty.com/***/	102	178	76	発売	11	51	40
127785764	36	96	60	***.seesaa.net/	154	224	70	シャンプー	63	97	34
136617285	8	63	55	***.seesaa.net/	19	80	61	女優	19	39	20
116355995	3	35	32	fine.ap.teacup.com/***/	45	105	60	拓哉	0	18	18
159521827	123	148	25	blog.livedoor.jp/***/	15	49	34	意味	7	24	17
146259965	13	37	24	***.seesaa.net/	135	160	25	ｌｉｖｅｄｏｏｒ	0	16	16
167278214	19	42	23	***.seesaa.net/	13	37	24	アイドル	0	16	16
158887514	68	90	22	***.seesaa.net/	19	42	23	花王	1	15	14
128271737	7	28	21	ameblo.jp/***/	68	90	22	ブログ記事	2	15	13
143177496	7	27	20	plaza.rakuten.co.jp/***/	7	28	21	ｙａｈｏｏ	1	13	12
126772208	1	20	19	www.blog-headline.jp/***/	1	20	19	荒川静香	22	34	12
152773185	65	84	19	***.livedoor.biz/	65	84	19	感じ	11	22	11
128679773	10	27	17	www.***.in/	72	89	17	既存	0	9	9
127735323	12	29	17	blog.so-net.ne.jp/***/	22	38	16	発行	0	9	9
159174859	72	89	17	d.hatena.ne.jp/***/	64	78	14	紹介	3	12	9
136617274	3	18	15	***.seesaa.net/	7	21	14	一息	0	9	9
130841837	2	16	14	plaza.rakuten.co.jp/***/	2	16	14	編集後記	0	9	9
132605629	4	16	12	***.at.webry.info/	4	18	14	商品	24	33	9

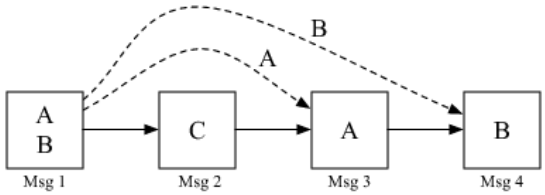

図 1: 語の伝播.

IDM のアルゴリズムから分かるように，ある語をあるメッセージに投下したときに語の伝播が起これば GCI は増加する．本稿で求めたいものは，GCI を最大化するための条件であるが，メッセージと語の全ての組合せを探索していては計算量が爆発する．そこで伝播条件を緩和することにより準最適解を求める方法を提案する．具体的には，スラック変数 ξ を導入して，以下の緩和した伝播条件で影響量を求める．

$$i_{x \to y} = \sum_{y=x}^{x+r} \|\check{m} \cap \hat{m} \cap m_y\| \quad (8)$$

$$\check{m} = \bigcap_{n=x}^{y-\xi-1} m_n \quad (9)$$

$$\hat{m} = \bigcup_{n=y-\xi}^{y-1} m_n \quad (10)$$

上式の意味を図 1 を用いて説明する．四角ノードがメッセージ，実線矢印がメッセージの返信関係（返信の向きは矢印の逆向き）を表しているとする．式 (2) の定義は途切れることなく伝播した語だけを後続するメッセージに影響を及ぼしたと見なすものなので，図 1 では 1 語も伝播していないことになる．これは，偶然による伝播を避けるために伝播の条件を厳しく設定しているためであるが，実際には伝播が途切れても語の影響力はしばらく持続していると考える方が自然である．そこで，伝播が多少途切れてもまた用いられたときには語が伝播したとするように式 (2) を拡張したものが式 (8,9,10) である．つまり，図 1 中で点線で表しているように，Msg1 から Msg3 に語 A，Msg1 から Msg4 に語 B が伝播していると判断できるように伝播条件を緩和しており，スラック変数 ξ は緩和する伝播範囲のマージンを表している．$\xi = 2$ のときは Msg1 から Msg3 へ語 A が，$\xi = 3$ のときは更に Msg1 から Msg4 へ語 B が伝播していることを認める．

このようにマージンを増やすことで，間欠的に伝播している語やその位置（メッセージや投稿者）を特定できる．これを利用することでコミュニティの動向の一歩先を掴もうというのが本稿の狙いである．

なお，$\xi = 1$ のときは式 (2) と等しくなるので，式 (8,9,10) は式 (2) を一般化したものと見なすことができる．また，$r = 0$ のときの語の影響力は語の文書頻度 (document frequency) と等しくなるので，式 (8,9,10) はネットワーク構造を反映するように文書頻度の定義を拡張したものと見なすこともできる．

この定式化により，伝播が一端途切れた後にまた繋がるところの影響量を計上できるようになる．したがって，式 (2) で求めた影響量 (I, J, K) と式 (8,9,10) で求めた影響量 (I', J', K') との差分 $(\nabla I, \nabla J, \nabla K)$ を見ることで，GCI が増えるポイントを間接的に見つけることができる．この方法で求まるポイントは必ずしも CGI を最大化するわけではないので，準最適解を求めていることになる．

4. 実験

4.1 全ての語を対象とした場合

2006 年 4 月に発売されたシャンプー「TSUBAKI」について 2006 年 3 月〜8 月の間に書かれたブログ記事 6464 件[*1]を

[*1] 自動生成されたアフィリエイト記事やアダルトな内容を含む記事は人手でチェックして除去している．

表 2: 特定の語についての影響力の変化．I, J, K は $\xi = 1$ のときの影響量，I', J', K' は $\xi = 2$ のときの影響量．

語	記事 ID ／投稿者	K	K'	∇K	I	I'	∇I	J	J'	∇J
資生堂		357	502	145						
	136617290				3	18	15			
	137707887				49	62	13			
	159521827				26	36	10			
	***.seesaa.net/							6	39	33
	onlyone.air-nifty.com/***/							0	20	20
	***.seesaa.net/							49	62	13
CM		589	724	135						
	116355995				3	35	32			
	136617290				0	15	15			
	137707887				85	95	10			
	***.seesaa.net/							11	54	43
	***.seesaa.net/							4	28	24
	***.seesaa.net/							96	106	10
発売		11	51	40						
	136617290				0	9	9			
	132605629				0	9	9			
	136617274				0	3	3			
	***.seesaa.net/							0	14	14
	plaza.rakuten.co.jp/***/							0	9	9
	**.at.webry.info/							0	4	4

対象にして実験を行った．なお，ブログ記事間に貼られたリンクは 453 本であった．

伝播範囲は $r = 5$ とし，スラック変数は $\xi = 1$（緩和なし），$\xi = 2$（緩和あり）に設定して実験を行った．影響量の変化の大きかったメッセージ，投稿者[*2]，語を表 1 に示す．∇I の上位のメッセージや投稿者は，伝播条件が一段階緩まるだけで大幅に影響力が増加することを示している．緩和なしのときの影響力（つまり I や J）が低いにも関わらず ∇I や ∇J が大きくなるメッセージや投稿者は，コミュニティにおいてはその影響力が顕在化していないが，少しのキッカケによって大きな影響を生み出す可能性がある．また，∇K の大きい語は励起状態にあり，刺激を与えることで雪崩のように伝播が引き起こされる可能性のある語だといえる．

4.2 特定の語を対象とした場合

次に，特定の語を投入したときの影響量の変化を調べてみる．「資生堂」「CM」「発売」を投入したときに影響量の増分（$\nabla I, \nabla J$）の大きいメッセージと投稿者の上位 3 件を表 2 に示す．ここで得られるメッセージや投稿者は，各語を広める可能性を持っているといえる．例えば，「資生堂」についての話題をもっと広めたいと思ったときには，記事 ID が 136617290, 137707887, 159521827 の周辺，もしくは投稿者 ***.seesaa.net/, onlyone.air-nifty.com/***/ の周辺が狙うべき対象となる．例えば，ある話題をコミュニティに効率的に広めたいときには，上記のブログ記事や投稿者のブログサイトに広告を載せるなどのアプローチを取ることが考えられる．

4.3 計算時間

4.1 節の実験において，MacBook (Mac OSX Tiger, 2GHz Intel Core 2 Duo, 2GB RAM) において ξ を 1 から 5 に変化させたときの解析時間（5 回平均）を図 2 に示す[*3]．ξ に対して解析時間は線形に増加していることが分かる．

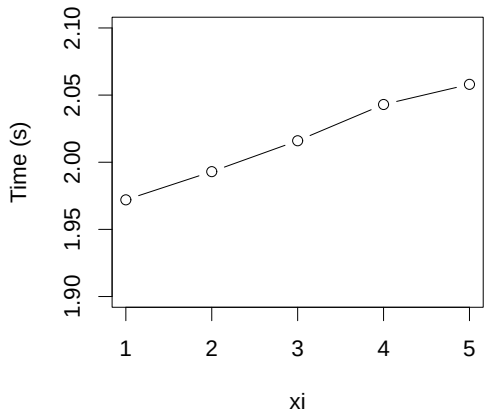

図 2: ξ と解析時間の関係．

5. まとめ

本稿では，コミュニティに蓄えられる総影響量 GCI を定義し，GCI を増やすためのアプローチとして IDM にスラック変数を導入する拡張を行った．ブログ記事のデータセットを用いて簡単な実験を行い，提案手法が狙い通りに機能していることと，計算量の増加をスラック変数の値に対し線形に抑えるこ

*2 ブログサイトの URL を投稿者としているが，投稿者を特定できないように一部を *** に置き換えている．

*3 形態素解析，ストップワードの処理，同義語処理の前処理は除く．

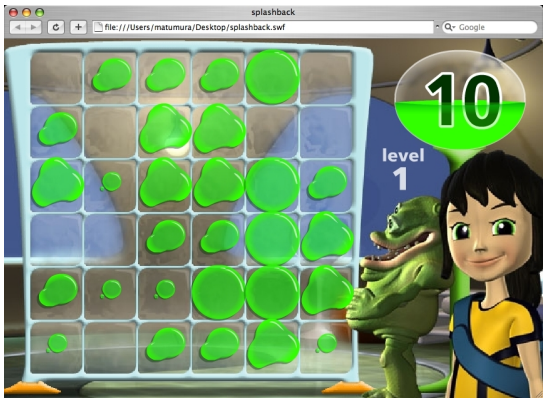

図 3: pya!水玉潰し.

とができることを確認した.

今後は，分析結果を詳細に検討して得られる知見を整理するとともに，より大規模なデータに適用する予定である．なお，今回は既存のネットワーク構造をそのまま利用したが，リンク数が少ないのでほとんどのブログ記事が孤立して，どこからも影響を受けていなかった．しかし，実際にはリンクとして顕在化していないが影響を及ぼしているであろう潜在的なリンクがあるはずである．そのような潜在的なリンクを考慮することは今後の課題の一つである．話題は伝播するにつれて変化していくものなので，そのように話題間の繋がりを考慮に入れて発展させるようなキッカケを見つけることも検討していきたい課題である．また，ある話題について肯定／否定しているという情報は，伝搬する情報の性質を理解するうえで非常に重要である．今回は考慮しなかったが，ブログ記事の極性を考慮した分析も今後の検討課題である．

謝辞

本稿で用いたブログ記事のデータセットはニフティ株式会社より提供を受けました．記して感謝いたします.

付録

最後に，本稿のアイデアの元になった「pya!水玉潰し」について紹介する．pya!水玉潰し [pya] は，図 3 に示すように「6 マス× 6 マス」のマス目と水玉を用いる Flash で作成されたゲームである．水玉は水を注入されると 4 レベルまで膨らみ，5 レベル目に到すると破裂して上下左右に 1 レベルの水滴が飛散する．飛散した水滴が他の水玉に当たると 1 レベル分の水が注入され，それによって水玉の破裂が連鎖的に起こりうる．全ての水玉を破裂させればゲームクリアとなり，注入 1 レベル分の水が補填される．手元の水が尽きればゲームオーバーとなるため，最小の水の注入量で全ての水玉を破裂させるために戦略を練る必要がある．1 手順異なるだけで全く異なる経路で連鎖が起こるところがこのゲームの面白いところである.

水玉が飛散して他の水玉に吸収され，それによって破裂の連鎖が起こるルールは IDM の影響伝搬のアイデアとよく似ており，本稿のアイデアも「pya!水玉潰し」で遊んでいるときに閃いたものである[*4]．ただし,「pya!水玉潰し」の伝播ネットワークは動的に変化する循環有向グラフ（DCG; Directed Cyclic Graph) となるため，残念ながら本稿で提案している手法で解くことはできない.

なお，このような影響伝播の連鎖反応は自然界でもよく見られる現象である．例えば，インフルエンザが人から人へ感染していく規模は他者への感染率と治療日数によって予測可能であり，大規模な感染はパンデミックと呼ばれている．森林火災についてもやはり原理的には同じであり，ある延焼のパターンでは木の密度 p がある値 (p=0.5928) を超えたところで森林火災が鎮火するまでの時間が急激に長くなる．このような現象は統計物理学の分野ではパーコレーションと呼ばれている [スタウハー 01].

*4 本稿のタイトルも「pya!水玉潰し」に敬意を示したものである.

参考文献

[スタウハー 01] D. スタウハー，A. アハロニー（著），小田垣考（訳）：パーコレーションの基本原理，吉岡書店 (2001)

[総務省 09] 総務省，平成 20 年通信利用動向調査，http://www.soumu.go.jp/main_content/000016027.pdf

[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 第 17 巻 3 号, pp. 259–267 (2002)

[松村 08a] 松村真宏，佐々木儀広：非営利組織における代替リーダーシップ行動の分析，情報処理学会論文誌, 49 巻 8 号, pp. 2783–2790 (2008)

[松村 08b] 松村真宏：影響普及モデル IDM の新しい影響量基準，第 22 回人工知能学会全国大会，1H2-08 (2008)

[Matsumura08] Naohiro Matsumura, Hikaru Yamamoto, Daisuke Tomozawa: Finding Influencers and Consumer Insights in the Blogosphere, International Conference on Weblogs and Social Media (ICWSM-08), Seattle, WA, March 31-April 2, pp. 76–83, 2008.

[pya] pya!水玉潰し, http://pya.cc/pyaimg/pimg.php?imgid=22468