1H2-08

影響普及モデルIDMの新しい影響量基準

New Influence Criteria on Influence Diffusion Model

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 proposes a new information criteria on Influence Diffusion Model (IDM). IDM is an algorithm to measure the influence of terms, messages, and participants by counting the number of propagating terms throughout message threads. The influence is designed to reflect the impact along with the context of topics of messages. However, IDM tends to estimate the influence of frequent terms higher than that of its actual impact, because frequent terms have more chance to propagate. To normalize such biased influence, I propose χ^2-influence, which is measured from IDM and its expected influence, as a new influence criteria of IDM.

1. はじめに

影響普及モデル（以降では Influence Diffusion Model を略して IDM と表記する）[松村 02, Matsumura 08] は，語の再帰的な伝播量に基づいて語，メッセージ，投稿者の影響量を求めるアルゴリズムである．これまで，電子掲示板 [松村 02, 松村 03b]，議論の書き起こしデータ [松村 03a]，メーリングリスト [佐々木 06]，ブログ記事 [Matsumura 08] など様々なデータの分析に用いられている．IDM ではメッセージのスレッド構造を利用して語の重み付けをするため，文脈において重要な語の影響量が高くなることが特徴である．

しかし，高頻度語は偶然に伝播する確率も高くなるため，影響量が本来持つべき量よりも高く計上されてしまう傾向がある．また，高頻度語は文脈と関係なく日常的によく使われる語であることが多いため，そのような語の影響量が高くなるのは好ましくない．日常的によく用いられる高頻度語がストップワードになっていたり，TFIDF 法が対数文書頻度の逆数を乗しているのも高頻度語の影響を減らすためである [徳永 99]．しかし，高頻度語でかつ影響量の高い語もあるため，IDM ではただ頻度の高い語の影響量を減らすのではなく，必然の伝播による影響量を正確に推計する枠組みが必要となる．

そこで本論文では，分析データと同様のメッセージ数，リンク数，語数を持つメッセージスレッドにおいて，語がランダムに用いられると仮定したときの影響量の期待値を求め，IDM による影響量とその期待値との乖離度を χ^2 影響量として求める新しい影響量基準を提案する．簡単な実験を行った結果，従来の影響量と比べて，文脈に関係して用いられる語の χ^2 影響量は高く，文脈と関係なく用いられる語の χ^2 影響量は低く算出されることが確認された．

2. IDM

IDM について簡単に説明する．なお，IDM のアルゴリズムは [佐々木 06] 以降に一部変更されており，本章でもその変更後のアルゴリズムについて述べる．

図 1 は 4 つのメッセージ（$\mathtt{Msg}_1, \mathtt{Msg}_2, \mathtt{Msg}_3, \mathtt{Msg}_4$）と，メッセージに含まれる語（A, B, C）を表している．また，実線矢印は返信関係（例えば $\mathtt{Msg}_2$ は $\mathtt{Msg}_1$ に返信している），点線矢印はメッセージ間を伝播する語（例えば $\mathtt{Msg}_1$ から $\mathtt{Msg}_2$ へは A が伝播している）を表している．

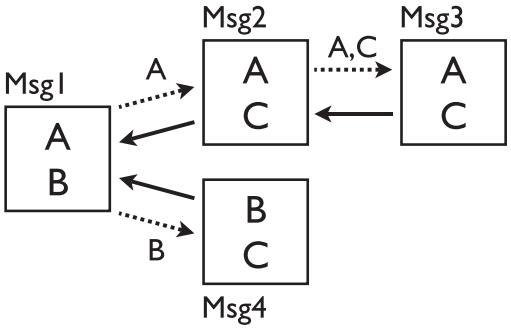

図 1: 4 つのメッセージの返信関係の例．A，B，C は各メッセージに含まれる語を表す．

ここで，$\mathtt{Msg}_1, \mathtt{Msg}_2, \mathtt{Msg}_3, \mathtt{Msg}_4$ に含まれる語の集合をそれぞれ w_1, w_2, w_3, w_4 とすると，$\mathtt{Msg}_x$ から $\mathtt{Msg}_y$ へ伝播する語数 $n_{x\to y}$ は式 (1) より算出される．

$$n_{x\to y} = |w_x \cap \cdots \cap w_y| \quad (1)$$

ここで $|w_x \cap \cdots \cap w_y|$ は $\mathtt{Msg}_x$ から $\mathtt{Msg}_y$ に至るメッセージチェーン上の全てのメッセージに共通して用いられる語の数を表している．伝播のとぎれた語はカウントしないようになっているが，これは伝播が途切れるとそこで文脈が変わることを仮定しているためである．

式 (1) に基づいて表 1 における伝播語数を求めると以下のようになる．

$$n_{1\to 2} = |w_1 \cap w_2| = 1$$
$$n_{1\to 3} = |w_1 \cap w_2 \cap w_3| = 1$$
$$n_{1\to 4} = |w_1 \cap w_4| = 1$$
$$n_{2\to 3} = |w_2 \cap w_3| = 2$$
$$n_{others} = 0$$

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

表 1: メッセージ間を伝播する語数

	$\mathtt{Msg}_1$	$\mathtt{Msg}_2$	$\mathtt{Msg}_3$	$\mathtt{Msg}_4$	影響量
$\mathtt{Msg}_1$	0	1	1	1	3
$\mathtt{Msg}_2$	0	0	2	0	2
$\mathtt{Msg}_3$	0	0	0	0	0
$\mathtt{Msg}_4$	0	0	0	0	0
被影響量	0	1	3	1	5

ここで，あるメッセージ $\mathtt{Msg_x}$ の影響量 i_x を他のメッセージに伝播した語の総数，つまり

$$i_x = \sum_{y \in all\_messages} n_{x \to y} \tag{2}$$

と定義すると，各メッセージの影響量 i_1, i_2, i_3, i_4 は以下のように表すことができる．

$$\begin{aligned} i_1 &= n_{1\to2} + n_{1\to3} + n_{1\to4} = 1+1+1 = 3 \\ i_2 &= n_{2\to3} = 2 \\ i_3 &= 0 \\ i_4 &= 0 \end{aligned}$$

また，あるメッセージ $\mathtt{Msg_x}$ の被影響量 j_x を他のメッセージから伝播してきた語の総数，つまり

$$j_x = \sum_{y \in all\_messages} n_{y \to x} \tag{3}$$

と定義すると，各メッセージの被影響量 j_1, j_2, j_3, j_4 は

$$\begin{aligned} j_1 &= 0 \\ j_2 &= n_{1\to2} = 1 \\ j_3 &= n_{1\to3} + n_{2\to3} = 1+2 = 3 \\ j_4 &= n_{1\to4} = 1 \end{aligned}$$

と表すことができる．

以上をまとめると，メッセージ間を伝播する語数，影響量，被影響量は表 1 のようになる．

ここで，$\mathtt{Msg}_1$ の投稿者を $\mathtt{S_a}$，$\mathtt{Msg}_2$ の投稿者を $\mathtt{S_b}$，$\mathtt{Msg}_3$ と $\mathtt{Msg}_4$ の投稿者を $\mathtt{S_c}$ とし，投稿者 $\mathtt{S_x}$ の影響量 I_x をその人が投稿したメッセージの影響量の和，つまり

$$I_x = \sum_{y \in all\_messages\_by\_x} i_y \tag{4}$$

と定義すると，投稿者 $\mathtt{S_a}, \mathtt{S_b}, \mathtt{S_c}$ の影響量 I_a, I_b, I_c は以下のようになる．

$$\begin{aligned} I_a &= i_1 = 3 \\ I_b &= i_2 = 2 \\ I_c &= i_3 + i_4 = 0 \end{aligned}$$

また，投稿者 $\mathtt{S_x}$ の被影響量 J_x をその人が投稿したメッセージの被影響量の和，つまり

$$J_x = \sum_{y \in all\_messages\_by\_x} j_y \tag{5}$$

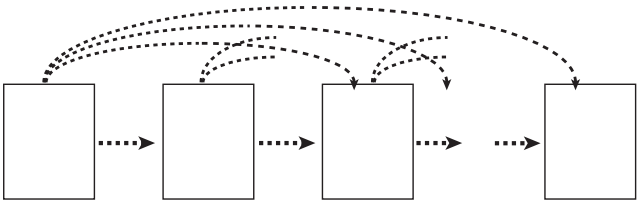

図 2: 仮定したメッセージスレッド

と定義すると，投稿者 $\mathtt{S_a}, \mathtt{S_b}, \mathtt{S_b}$ の被影響量 J_a, J_b, J_c は以下のようになる．

$$\begin{aligned} J_a &= j_1 = 0 \\ J_b &= j_2 = 1 \\ J_c &= j_3 + j_4 = 1+3 = 4 \end{aligned}$$

語の影響量についても同様に求めることができる．まず，関数 $\delta_{x\to y}$ を以下のように定義する．

$$\delta_{x\to y}(w) = \begin{cases} 1 & (\text{if } \{w_x \cap \cdots \cap w_y\} \text{ contains } w) \\ 0 & (\text{otherwise}) \end{cases} \tag{6}$$

この関数 $\delta_{x\to y}$ を用いて語 w の影響量 K_w を

$$K_w = \sum_{\{x,y|x<y\} \in all\_pairs\_of\_messages} \delta_{x\to y}(w) \tag{7}$$

と定義すると，語 $\mathtt{A}$, $\mathtt{B}$, $\mathtt{C}$ の影響量 K_A, K_B, K_C は以下のようになる．

$$\begin{aligned} K_A &= \delta_{1\to2}(A) + \delta_{1\to3}(A) + \delta_{2\to3}(A) = 3 \\ K_B &= \delta_{1\to4}(B) = 1 \\ K_C &= \delta_{2\to3}(C) = 1 \end{aligned}$$

このようにして，自分の発した語への興味の強さを影響量，他の人の発した語への興味の強さを被影響量として求めることができる．また，式 (1) に示したように，途切れることなく伝播した語だけが被/影響量としてカウントされる．したがって，IDM は文脈に関係のない語は被/影響量に計上されにくいアルゴリズムとなっている．しかし，語の頻出頻度が高くなれば偶然に伝播する可能性も高くなるため，高頻度語は本来の影響量より高く計上される傾向がある．

3. 新しい影響量基準

本章では，前章の最終段落で指摘した IDM の問題点を解決するために，IDM による影響量を影響量の期待値によって正規化する新しい影響量を定義する．

まず，メッセージを一列に並べてメッセージ間にリンクを張った図 2 の構造をもつスレッドを仮定する．図 2 では，見やすくするために語の伝播経路である点線矢印のみ示し，リンクを表す実線矢印は省略している．ここで，メッセージ数を N，リンク数を L，語 w の文書頻度を f_w とすると，メッセージに語 w が出現する割合 R_w は $R_w = f_w/N$ となる．また，メッセージに接続されているリンクの割合 R_L は $R_L = L/(N-1)$ となる．この時，語 w があるメッセージに出現するときに他のメッセージに伝播する割合は $R_w R_L$ ずつ減少していくと表すことができる．影響量に伝播する割合を掛けたものの総和が

表 2: 影響量と期待値と χ^2 影響量

	A	B	C	$\mathtt{Msg}_1$	$\mathtt{Msg}_2$	$\mathtt{Msg}_3$	$\mathtt{Msg}_4$	$\mathtt{S}_a$	$\mathtt{S}_b$	$\mathtt{S}_c$
影響量	3	1	1	3	2	0	0	3	2	0
期待値	1.88	0.5	1.88	1.06	1.06	1.06	1.06	1.06	1.06	2.13
χ^2 影響量	0.675	0.5	0.408	3.53	0.827	1.06	1.06	3.53	0.827	2.13

影響量の期待値となるので，$f_w \geq 2$ のときの語 w の影響量の期待値 E_w は以下の式で表される．

$$\begin{aligned} E_w &= \underbrace{\underbrace{\underbrace{R_w R_L}_{f_w=2\text{ のとき}} + 2(R_w R_L)^2}_{f_w=3\text{ のとき}} + 3(R_w R_L)^3 + \cdots}_{fw=4\text{ のとき}} \\ &= \sum_i^{f_w-1} i(R_w R_L)^i \\ &= \sum_i^{f_w-1} i\left(\frac{f_w}{N}\cdot\frac{L}{N-1}\right)^i \quad (\text{ただし } f_w \geq 2) \quad (8) \end{aligned}$$

$f_w \leq 1$ のときは語 w の伝播は起こりえないので $E(w) = 0$ となる．

IDM による影響量 K_w と影響量の期待値 E_w の差が大きければ伝播が偶然に起こっているわけではないことを示している．χ^2 統計量により IDM による影響量 K_w と影響量の期待値 E_w の乖離度 $K_w^{\chi^2}$ を求めると

$$K_w^{\chi^2} = \frac{(K_A - E_w)^2}{E_w} \quad (9)$$

となる．このように期待値により正規化された新しい影響量を χ^2 **影響量**と定義する．$f_w \leq 1$ のときは $E_w = 0$ となるが，このときは K_w も必ず 0 になるので $K_w^{\chi^2} = 0$ とする．

図 1 の場合に当てはめると，4 つのメッセージと 3 本のリンクからなるので $N = 4$，$L = 3$ となる．また，語 A, B, C の頻度はそれぞれ $f_a = 3, f_b = 2, f_c = 3$ なので，語 A, B, C の影響量の期待値 E_A, E_B, E_C は以下のように求まる．

$$\begin{aligned} E_A &= \frac{3}{4}\cdot\frac{3}{4-1} + 2\left(\frac{3}{4}\cdot\frac{3}{4-1}\right)^2 = \frac{15}{8} \\ E_B &= \frac{2}{4}\cdot\frac{3}{4-1} = \frac{1}{2} \\ E_C &= \frac{3}{4}\cdot\frac{3}{4-1} + 2\left(\frac{3}{4}\cdot\frac{3}{4-1}\right)^2 = \frac{15}{8} \end{aligned}$$

したがって，語 A, B, C の χ^2 影響量 $K_A^{\chi^2}, K_B^{\chi^2}, K_C^{\chi^2}$ は以下のようになる．

$$\begin{aligned} K_A^{\chi^2} &= \frac{(3-15/8)^2}{15/8} = \frac{27}{40} = 0.675 \\ K_B^{\chi^2} &= \frac{(1-1/2)^2}{1/2} = \frac{1}{2} = 0.5 \\ K_C^{\chi^2} &= \frac{(1-15/8)^2}{15/8} = \frac{49}{120} = 0.408 \end{aligned}$$

また，各メッセージの影響量の期待値 E_{msg} は，語の期待値の総和 E_{total} をメッセージ数で割った値になるので，

$$E_{msg} = \frac{E_A + E_B + E_C}{4} = \frac{17}{16} \quad (10)$$

となる．したがって，メッセージ $\mathtt{Msg}_1$, $\mathtt{Msg}_2$, $\mathtt{Msg}_3$, $\mathtt{Msg}_4$ の χ^2 影響量 $i_1^{\chi^2}, i_2^{\chi^2}, i_3^{\chi^2}, i_4^{\chi^2}$ は以下のようになる．

$$\begin{aligned} i_1^{\chi^2} &= \frac{(i_1 - E_{msg})^2}{E_{msg}} = \frac{(3-17/16)^2}{17/16} = 3.53 \\ i_2^{\chi^2} &= \frac{(i_2 - E_{msg})^2}{E_{msg}} = \frac{(2-17/16)^2}{17/16} = 0.827 \\ i_3^{\chi^2} &= \frac{(i_3 - E_{msg})^2}{E_{msg}} = \frac{(0-17/16)^2}{17/16} = 1.06 \\ i_4^{\chi^2} &= \frac{(i_4 - E_{msg})^2}{E_{msg}} = \frac{(0-17/16)^2}{17/16} = 1.06 \end{aligned}$$

また，投稿者の影響量の期待値は投稿数に比例するので，投稿者 $\mathtt{S}_a$, $\mathtt{S}_b$, $\mathtt{S}_c$ の影響量の期待値 $E_{S_a}, E_{S_b}, E_{S_c}$ は

$$\begin{aligned} E_{S_a} &= E_{msg} \times 1 = 17/16 \\ E_{S_b} &= E_{msg} \times 1 = 17/16 \\ E_{S_c} &= E_{msg} \times 2 = 17/8 \end{aligned}$$

となる．したがって，投稿者 $\mathtt{S}_a$, $\mathtt{S}_b$, $\mathtt{S}_c$ の χ^2 影響量 $I_a^{\chi^2}, I_b^{\chi^2}, I_c^{\chi^2}$ は以下のようになる．

$$\begin{aligned} I_a^{\chi^2} &= \frac{(I_a - E_{S_a})^2}{E_{S_a}} = \frac{(3-17/16)^2}{17/16} = 3.53 \\ I_b^{\chi^2} &= \frac{(I_b - E_{S_b})^2}{E_{S_b}} = \frac{(2-17/16)^2}{17/16} = 0.827 \\ I_c^{\chi^2} &= \frac{(I_c - E_{S_c})^2}{E_{S_c}} = \frac{(0-17/8)^2}{17/8} = 2.13 \end{aligned}$$

以上のようにして得られた影響量と期待値と χ^2 影響量を表 2 に示す．頻度の高い語 A, C と低い語 B の差が χ^2 影響量では少なくなったり，$\mathtt{Msg}_1$ や $\mathtt{S}_a$ の χ^2 影響量が相対的に高くなっていることなどが見て取れる．

なお，χ^2 被影響量についても χ^2 影響量と同様の手順で求めることができるので，本稿では省略する．

4. 分析事例

2 ちゃんねるの「大阪で最強のたこ焼き屋　その 4」スレッドを分析した．前処理として，引用符に加えて仮想的なリンクを 5 本/メッセージ張ることによりメッセージスレッドを構築し，形態素解析器 MeCab を用いて名詞，形容詞，副詞だけを残した．このデータに提案手法を適用した．

まず，頻度，影響力，χ^2 影響力による上位 10 語を表 3 に示す．このスレッドは 26 種のハンドルネームによる 362 投稿からなっており，スレッド名が示しているように大阪のたこ焼き屋について語り合っている．したがって，「大阪」「たこ焼き」といった語はスレッド全体に出現する頻出語となっている．また，「蛸次郎」についての荒らし投稿があったため「蛸次郎」も頻出語となっている．「レス」は返信することを指すジャーゴンであり，2 ちゃんねるでよく使われている．

表 3: 各指標による上位 10 語

順位	語（頻度）	語（影響量）	語（χ^2 影響量）
1	レス (133)	レス (50)	割高 (839)
2	たこ焼き (108)	たこ焼き (39)	馬鹿 (605)
3	大阪 (56)	馬鹿 (21)	無断 (504)
4	自分 (39)	大阪 (17)	寝屋川 (245)
5	寝屋川 (33)	寝屋川 (12)	迷惑 (167)
6	蛸次郎 (28)	自分 (10)	レス (159)
7	悪い (26)	割高 (9)	大阪 (147)
8	馬鹿 (25)	無断 (7)	敢えて (145)
9	必死 (24)	迷惑 (6)	評価 (145)
10	爆笑 (22)	悪い (4)	煽り (145)

表 4: 各指標による語の順位の変化

順位	語	頻度	影響量	χ^2 影響量
Up ↗	割高	19 位	7 位	1 位
	馬鹿	8 位	3 位	2 位
	無断	38 位	8 位	3 位
	迷惑	25 位	9 位	5 位
Down ↘	レス	1 位	1 位	6 位
	たこ焼き	2 位	2 位	28 位
	大阪	3 位	4 位	7 位
	蛸次郎	6 位	25 位	53 位

ここで，表 3 の語のうち，頻度，影響量，χ^2 影響量の順に順位が高くなっている語と，順位が低くなっている語の一部を表 4 に示す．表 4 より,「割高」「馬鹿」「無断」「評価」といったスレッドの主要な文脈に関わる話題に関しては順位が高まっているが,「レス」「たこ焼き」「大阪」「蛸次郎」といった発散的な話題に関しては順位が下がっている傾向が見て取れる．

次に，投稿者の投稿数，影響量，χ^2 影響量による上位 5 名を表 5 に示す．このスレッドは，ハンドルネームを記入せずに投稿すると自動的に「はふはふ名無しさん」がハンドルネームとして割り当てられる仕様になっている．2 ちゃんねるではハンドルネームを入れずに投稿するスタイルが主流となっているため，本スレッドにおいても 362 投稿中 322 投稿が「はふはふ名無しさん」によって投稿されていた．したがって，投稿数だけを見ると「はふはふ名無しさん」が主要なハンドルネームとして挙がってくるが実態はそうではない．影響量を用いても「はふはふ名無しさん」は第 1 位にランクインする．しかし，χ^2 影響量では，実質的な発言をする「寝屋川市民」，荒らしの一連のやり取りの中で一時的に使用された「12」，スレッドの始まりとなる投稿者の使った「1」といったハンドルネームが「はふはふ名無しさん」より上位になっており，実質的な発言をしている人をうまく取り出せていることが分かる．

5. まとめ

本稿では影響普及モデル IDM により求まる影響量を影響量の期待値で正規化した新しい影響量基準として χ^2 影響量を提案し，簡単な分析事例を示した．

語の連鎖を利用する J. Kleinberg の Burst アルゴリズム [Kleinberg 02] は語の伝播を扱う IDM と関連しているが，IDM では，メッセージのスレッド構造を利用している点，語・メッセージ・投稿者の影響量を同時に求めている点，期待値により正規化している点などが大きく異なっている．

スペースの都合もあり本稿には含めなかったが，χ^2 影響量はリンクの影響量にも同様に適用できる．したがって，これまでに取り組んできた人間関係ネットワーク [松村 02, 佐々木 06]，メッセージチェーン [松村 02]，投稿者のプロファイリング [松村 03b]，語と語の関係を表すワードチェーン [Matsumura 08] には χ^2 影響量を適用することもできる．今後は，これら従来研究との比較に加えて，χ^2 影響量の新しい領域への応用にも積極的に取り組んでいきたい．

参考文献

[Kleinberg 02] J. Kleinberg: Bursty and Hierarchical Structure in Streams, Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 91–101 (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 第 17 巻 3 号, pp. 259–267 (2002)

[松村 03a] 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 第 18 巻 4 号，pp. 165–172 (2003)

[松村 03b] 松村真宏，加藤優、大澤幸生，石塚満：議論構造の可視化による論点の発見と理解，知能と情報, Vol. 15, No. 5, pp. 554–564 (2003)

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

[佐々木 06] 佐々木儀広、松村真宏:NPO におけるリーダーシップ行動の発見，情報と知能，Vol.18, No.2, pp. 233-239 (2006)

[徳永 99] 徳永健伸 (著), 辻井 潤一 (編集) ：情報検索と言語処理，東京大学出版会 (1999)

表 5: 各指標によるハンドルネームの上位 5 名

順位	投稿者（投稿数）	投稿者（影響量）	投稿者（χ^2 影響量）
1	ぱふぱふ名無しさん (322)	ぱふぱふ名無しさん (127)	寝屋川市民 (2606)
2	寝屋川市民 (6)	寝屋川市民 (41)	12(1555)
3	12(5)	12(29)	1(598)
4	,,,,(3)	,,,,(13)	,,,,(515)
5	96(2)	1(8)	ぱふぱふ名無しさん (432)