3B4-1

# 影響伝播モデルIDMによるクチコミの感度分析

Sensitivity Analysis of Word-of-Mouth by 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

Influence Diffusion Model (IDM) is an algorithm to measure the influence of terms, messages, senders, etc based on the term propagation throughout message-threads. In this paper, we investigate the sensitivity of the influence in the following two conditions: (1) bias of term appearance and link structure, and (2) recursively/not-recursively term propagation. The results imply that IDM can be a sensitive detector of "burst" words.

## 1. はじめに

影響伝播モデル（以降ではIDMと略す）は，語の再帰的な伝播量に基づいて語やメッセージや投稿者の影響量を求めるアルゴリズムである．これまでIDMの線形代数表現による数理的背景などを求めてきたが [松村 2010]，本稿ではIDMによる影響量と語の出現箇所やネットワーク構造の偏りとの関係について検討する．IDMは人から人に情報が伝わる「クチコミは」の振る舞いをモデル化したものであるため，IDMの観点からクチコミの感度を分析することになる．

## 2. IDMによる語の影響量

IDMは，メッセージをノード，メッセージ間の参照関係をリンクとするネットワーク構造を対象とする．メッセージ間の参照関係とは，メールにおける返信関係，ブログにおけるリンクやトラックバックが相当する．このとき，リンクは必ず時間軸上の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．

ここでまず，メッセージ $x$ とメッセージ $y$ が同じメッセージスレッド上にあるときに（$x$ の投稿の後に $y$ が投稿されたとする），$x$ が $y$ 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，例えば [松村 2010] では以下の指標が用いられている．

$$i_{x\to y} \quad = \quad \beta^{step\#-1}\,|m_x \cap m_y| \tag{1}$$

$m_n$ 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，$\beta$ は減衰係数，$step\#$ は再帰伝播回数とする．式 (1) は伝播した語の数に減衰係数を乗じたものを影響量と見なす指標である．このとき，メッセージ $x$ に含まれる語 w の影響力 $k_{w,x}$ を $k_{w,x\to y} = i_{x\to y}/|m_x \cap m_y|$ とすると，語 $w$ の影響量 $K_w$ は以下のように求まる [松村 2010]．

$$k_w \quad = \quad \sum_{\substack{\forall x \in messages\ including\ w \\ \forall y \in messages\ followed\ by\ x}} k_{w,x\to y} \tag{2}$$

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

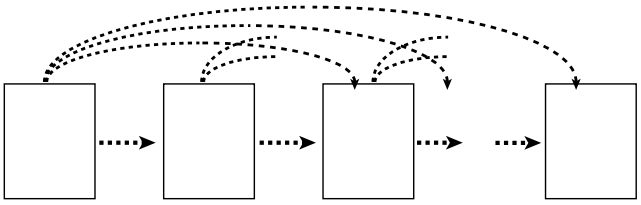

図 1: 仮想メッセージスレッド

## 3. 影響力の期待値

IDMではネットワーク上を伝播した語だけが影響量として計上されるので，文脈に関係のない語の影響を受けにくいアルゴリズムになっている．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．そこで，ネットワーク構造（ノード数とリンク数）と語の出現頻度から語の影響量の期待値を見積もることを試みる．

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 1 の構造をもつスレッドを仮定する．図 1 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 $f$ とすると，メッセージに語 $w$ が出現する割合 $R_w$ は $R_w = f/N$ 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．伝播機会数と伝播する割合を掛けたものの総和が影響量の期待値となるので，$R = R_w R_L$ とすると語 $w$ の影響量の期待値 $E_w$ は以下の式 (3) で表される*1．$\beta$ は減衰係数である．なお，$f \le 1$ のときは語 $w$ の伝播は起こりえないので $E_w = 0$ となる．

$$\begin{aligned} E_w \quad &= \quad \underbrace{\underbrace{\underbrace{(N-1)R}_{\text{伝播 1 回のとき}} + \beta(N-2)R^2}_{\text{伝播 2 回のとき}} + \beta^2(N-3)R^3}_{\text{伝播 3 回のとき}} + \cdots \\ &\simeq \quad \frac{NR}{1-\beta R} \quad (f \gg 1, R < 1 \text{ のとき}) \end{aligned} \tag{3}$$

*1 [松村 2010] とは異なっていることに注意．

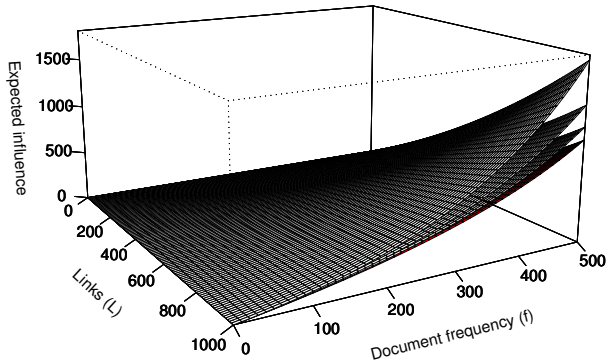


図 2: 語の出現箇所の偏り $\alpha_w$ とリンク構造の偏り $\alpha_L$ による影響量の期待値の違い．曲面は下から順に $\alpha_w = \alpha_L = 1$，$\alpha_w = \alpha_L = 0.8$，$\alpha_w = \alpha_L = 0.6$，$\alpha_w = \alpha_L = 0.4$ のときの影響量の期待値を表している．

## 4. 語の出現箇所とリンク構造の偏り

式 (3) は語の出現箇所とリンク構造がランダムの場合であったが，これらが偏っていれば語の影響量の期待値も異なる．そこで，**語の出現箇所のランダムネス**を表すパラメータ $\alpha_w (0 \leq \alpha_w \leq 1)$，**リンク構造のランダムネス**を表すパラメータ $\alpha_L (0 \leq \alpha_L \leq 1)$ を導入し，影響量の期待値を拡張する．$\alpha_w, \alpha_L$ が大きくなるほどランダムさが増すとすると，語の出現する可能性のあるノード数 $N_w$ とリンクの貼られる可能性のあるノード数 $N_L$ は以下のように表せる．

$$N_w = (N-f) * \alpha_w + f \quad (4)$$

$$N_L = (N-1-L) * \alpha_L + L \quad (5)$$

$N_w, N_L$ を用いると，メッセージに語 $w$ が出現する割合 $R'_w$ は $R'_w = f/N_w$，メッセージに接続されているリンクの割合 $R_L$ は $R'_L = L/N_L$ となる．このとき，$R' = R'_w R'_L$ と置くことで，式 (3) の影響量の期待値は以下のように拡張できる．

$$E_w \simeq \frac{N_w R'}{1-\beta R'} \quad (\beta R' < 1 \text{ のとき}) \quad (6)$$

$\beta R' = 1$ のときは $E_w = f(f-1)/2$ になる．

ここで，ランダムネスを変えたときの影響量の変化量を見るために，$N = 1000$，$\beta = 1$，$0 \leq f \leq 500$，$0 \leq L \leq 1000$ における影響量の期待値を図 2 に示す．図 2 中の 4 枚の曲面は，それぞれ下から順に $\alpha_w = \alpha_L = 1$，$\alpha_w = \alpha_L = 0.8$，$\alpha_w = \alpha_L = 0.6$，$\alpha_w = \alpha_L = 0.4$ のときの影響量の期待値を表している．これより，語の出現箇所やリンク構造が偏るほど影響量も大きくなることが分かる．また，その傾きは語の出現件数 $f$ やリンク数 $L$ が大きくなるほど急激に大きくなる．

## 5. 再帰

IDM の特徴は語の再帰的な伝播関係（2 ステップ以上の伝播）を影響量に含めるところにある．再帰的伝播がある場合の語の影響量の期待値 $E_w$ は式 (3) で表されるが，再帰的伝播がない（1 ステップの伝播だけ考慮する）場合の語の影響量 $e_w$ は $e_w = (N-1)R$ で求まる．ここで，再帰的伝播がある場合の影響量とない場合の影響量の違いを見るために，$N = 1000$，$\beta = 1$，$0 \leq f \leq 500$，$0 \leq L \leq 1000$ における影響量の期待値

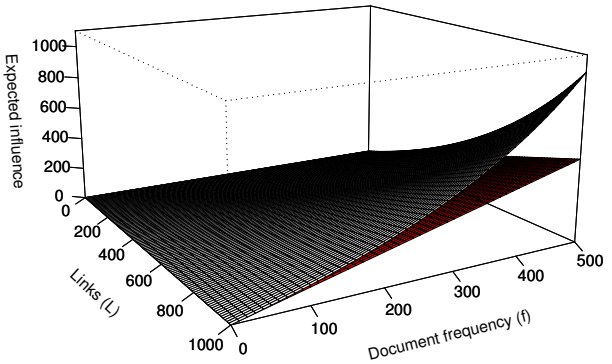


図 3: 再帰的伝播あり／なしにおける影響量の期待値の違い．下の曲面は再帰的伝播なし，上の曲面は再帰的伝播ありのときの影響量の期待値を表している．

を図 3 に示す．図 3 中の 2 枚の曲面のうち，下の曲面は再帰的伝播なし，上の曲面は再帰的伝播ありのときの影響量の期待値を表している．これより，再帰的伝播を考慮しないときには語の出現件数とリンク数に対して単調に影響量が増えるのに対し，再帰的伝播を考慮すると急激に増えることが分かる（図 2 の一番下の曲面と同じ）．

## 6. まとめ

本稿では語の影響量の感度分析を行い，同一語が偏って出現するほど影響量は急激に大きくなることを確認した．また，IDM の特徴である再帰的伝播についても検討し，再帰的伝播によって影響力が増幅されることを確認した．これらの結果は，IDM が，特定の場所もしくは特定のリンク構造の元において現れる語の高感度の検出器と成り得ることを示唆している．

## 参考文献

[松村 2010] 松村真宏：影響伝播モデル IDM の線形代数表現と Twitter 分析への応用，第 16 回 Web インテリジェンスとインタラクション研究会, IEICE SIG Notes WI2-2010-22, pp. 73–78 (2010)

## 付録

式 (3) を導出する．まず $E_w$ を以下のようにおく．

$$E_w = (N-1)R + \beta(N-2)R^2 + \cdots + \beta^{f-1}(N-f)R^f$$

$E_w$ の両辺に $\beta R$ をかけて $E_w$ から引いて整理すると

$$(1-\beta R)E_w = (N-1)R - \frac{\beta R^2(1-\beta^{f-2}R^{f-2})}{1-\beta R} - (N-f)\beta^f R^{f+1} \quad (7)$$

が得られる．ここで，$R < 1, f \gg 0$ のとき $R^f \simeq 0$ より，

$$E_w \simeq \frac{1}{1-\beta R}\left((N-1)R - \frac{\beta R^2}{1-\beta R}\right)$$

となり，さらに，$NR \gg \left|-\frac{\beta R^2}{1-\beta R} - R\right|$ より以下が導かれる．

$$E_w \simeq \frac{NR}{1-\beta R}$$