3M4-1

嗜好の類似性に着目したソーシャルネットワーク影響力分析

Social Influence Analysis Based on Taste Similarity

早矢仕　裕 *1　　親松　昌幸 *1　　廣井　和重 *1
Yu Hayashi　　Masayuki Oyamatsu　　Kazushige Hiroi

*1株式会社日立製作所 研究開発グループ
Hitachi, Ltd.

We propose the influence analysis method in social networks for the efficient information diffusion. In this method, we assume the information diffusion model in which the diffusion probabilities depend on users' taste similarlity. Under this assumption, we propose the method to select influencial nodes. We employ real twitter dataset to demonstrate to the validity of the proposed method about an efficiency of information diffusion.

1. 序論

近年, 消費者の間でブログ (blog;Weblog) や SNS(Social Networking Service) 等に代表されるソーシャルメディアの普及が進んでいる. 代表的なソーシャルメディアサービスである Twitter*1 や Facebook*2 は世界中で利用されており，数億人以上のユーザ数を保有している．ソーシャルメディアの普及に伴い, マーケティングにおいてもその影響は増大している. 消費者が商品及びサービスの購入時に参考にする情報として，従来影響力を持っていたマスメディアに代わり，口コミを中心にソーシャルメディアに関する要因が上位に挙がっている．

このような状況から，マーケティングにおいてソーシャルメディアを活用した手法が用いられている．手法の一例として，ユーザ同士の関係であるソーシャルネットワークに着目したバイラルマーケティングが挙げられる．バイラルマーケティングは，ソーシャルネットワーク上で大きな影響力を持った少数のユーザ（インフルエンサ）に対して宣伝等を実施することで，インフルエンサから他ユーザへの口コミを発生させ，効率的に情報を広める手法である．

ソーシャルネットワーク上での影響力に着目した研究は多数行われているが，その中でも，本稿では影響最大化問題 (Influence Maximization) [Domingos 01, Kempe 03] を扱う．影響最大化問題は，特定の情報伝播モデルの下で，ソーシャルネットワーク上で情報を受信する人数を最大化するように，一定人数の情報配信先を決定するという問題である．

本稿では, 影響最大化問題を扱うにあたり，情報配信先同士の嗜好の類似性に着目する．例えばバイラルマーケティングを考えた際に，情報配信先から他のユーザに口コミ (情報伝播) が発生するためには，広告等の情報が情報配信先の嗜好に合ったものである必要がある．このことから，同一の情報を複数の情報配信先に配信するケースを考えると，情報配信先の嗜好の類似性が低い場合には，多数の情報配信先の嗜好にあった情報を配信できず，情報が広まらない可能性がある．

本稿では，影響最大化問題において，情報伝播の発生する確率が情報配信先の嗜好の類似性に依存する状況での情報配信先決定手法を提案する．また，人工データ及び実際の Twitter データを用いて情報伝播効率に関して従来手法との比較評価を行い，提案手法の有効性を示す．

連絡先: 早矢仕 裕，株式会社日立製作所 研究開発グループ，
yu.hayashi.gn@hitachi.com

*1　https://twitter.com (”Twitter”は Twitter, Inc. の商標である．)

*2　https://www.facebook.com (”Facebook”は Facebook, Inc. の商標である．)

2. 関連研究

本稿で対象とする影響最大化問題について，Kempe ら [Kempe 03] は基本的な情報伝播モデルである Linear Threshold(LT) モデル及び Independent Cascade(IC) モデルの下で，影響最大化問題を最適化問題として定式化し,greedy 法による情報配信先の決定手法を提案している．

また，ソーシャルネットワークの影響力分析に関して, 情報の内容やユーザ属性による伝播の違いを扱った研究が行われている．Saito らの研究 [Saito 11] では，情報伝播確率がノード属性に依存する情報伝播モデルを扱い, このような場合での情報伝播確率の推定方法を提案している．Weng らの研究 [Weng 10] では，特定の話題において影響力を有するユーザを発見する手法を提案している．

これまでに，情報の内容やノード属性による情報伝播の違いを考慮した研究が行われているが，選択した情報配信先同士のノード属性の類似性に着目したものはなかった．本稿では，情報伝播確率が情報配信先のノード属性（嗜好）の類似性に依存する状況を対象とする．

3. 提案手法

本章では，情報配信先の嗜好（ノード属性）の類似性を加味した情報伝播モデルを導入し, 本モデルの下で情報配信先の決定方法を提案する．3.1 節で，既存の情報伝播モデルである Independent Cascade(IC) モデルについて説明し，3.2 節で，IC モデルを拡張する形で情報配信先の嗜好の類似性を加味した情報伝播モデルを導入する．3.3 節で，3.2 節で導入した情報伝播モデルの下での情報配信先決定手法を提案する．

3.1 IC モデル

はじめに，情報伝播モデルの一つである IC モデル [Kempe 03] について説明する．ソーシャルネットワークをノード (頂点) 集合 V，有向辺集合 E からなる有向ネットワーク $G = (V, E)$ で表現し，G 上での情報伝播を考える．ここで，ノード $v \in V$ はユーザ，辺 $e = (v, w) \in E$ はユーザ v からユーザ w のつながりに対応する．また，各ノードは「アクティブ」,「非アクティブ」のいずれかの状態を持つ．ノード v に

対応するユーザが伝播した情報を受信した場合に，ノード v はアクティブとなる．

IC モデルでは，ネットワーク G の各辺 $e = (v, w) \in E$ に対して，情報伝播確率 $p_{v,w}(0 < p_{v,w} < 1)$ が定義される．情報伝播確率 $p_{v,w}$ は，辺 $e = (v, w)$ を通じて情報が伝播する確率である．ネットワーク G，情報伝播確率 $p_{v,w}$ が与えられたとき，以下の流れで情報伝播が発生する．

手順 1 G の複数のノードを情報配信先として選択し，選択したノードをアクティブにする．

手順 2 アクティブなノード v を一つ選び，G の有向辺 $e = (v, w)$ でつながっている各ノード w について，確率 $p_{v,w}$ でユーザ w をアクティブにする．

手順 3 全てのアクティブなノードについて，手順 2 を実行する．なお，手順 2 は各ノードについて一回のみ行う．

3.2 嗜好の類似性を考慮した情報伝播モデル

次に,IC モデルを拡張する形で，嗜好の類似性を加味した情報伝播モデルを導入する．本情報伝播モデルでは，最初に複数のノードを情報配信先として選択した際に，配信した情報が各情報配信先に適合するかを考慮する．本稿では，情報配信先全員に同一の情報を配信する状況を仮定し，情報配信先同士の嗜好のばらつきが大きい場合に，情報配信先に情報が適合する確率が小さくなるようなモデルを提案する．

提案モデルでは，ネットワーク G，情報伝播確率 $p_{v,w}$ に加え，各ノード v が嗜好ベクトル $x_v = (x_{v1}, x_{v2}, ...x_{vn}) \in \mathbb{R}^n$ を持つ．ネットワーク G，情報伝播確率 $p_{v,w}$, 嗜好ベクトル x_v が与えられた下で，以下の流れで情報伝播が発生する．

手順 1 G の複数のノードを情報配信先として選択する．

手順 2 各情報配信先 s について，嗜好ベクトル x_s を用いて情報適合確率 q_s を算出し，確率 q_s で s をアクティブにする．

手順 3 アクティブなノード v を一つ選び，G の有向辺 $e = (v, w)$ でつながっている各ノード w について，確率 $p_{v,w}$ でユーザ w をアクティブにする．

手順 4 全てのアクティブなノードについて，手順 3 を実行する．なお，手順 3 は各ノードについて一回のみ行う．

IC モデルとは異なる点は手順 2 である．情報配信先を選択した際に，嗜好の類似性に基いた情報適合確率 q_s により, 各情報配信先に情報が適合するかが決まる. ここで各情報配信先 s に対して情報が適合する確率 (情報適合確率)q_s は以下の通り定義される．

$$\begin{aligned} q_s &= e^{-ad(x_s,c)} \\ c &= \frac{1}{|S|}\sum_{i\in S} x_i \end{aligned}$$

ここで，a は情報伝播に関するパラメータ，c は情報配信先の集合における嗜好ベクトルの平均，$d(x_s, c)$ は x_s と c の間の距離関数，S は情報配信先の集合である．

提案モデルにおいて，情報適合確率 q_s は情報配信先 s の嗜好ベクトル x_s と c の間の距離が大きくなるほど小さい値をとる．すなわち，情報配信先の集合における嗜好ベクトルのばらつきが大きい場合に，情報適合確率が小さくなるモデルとなっている．また，a は嗜好ベクトルのばらつきが，どの程度情報適合確率に影響するかを表すパラメータである．

3.3 情報配信先の決定手法

前節で提案した情報伝播モデルの下での，効率的な情報配信先の決定方法について述べる．影響最大化問題は，情報配信先とするノード数 (情報配信数)n を一定とした下で，アクティブとなるノード数を最大化するように情報配信先を決定することが目的である．

提案手法は，以下の手順からなる．なお, 情報配信先の集合を S, 情報配信数を n とする．

手順 1 各ノード v に対して，嗜好ベクトル x_v を算出する．

手順 2 嗜好ベクトル x_v の類似性に基づいて，ノードをクラスタリングする．

手順 3 各クラスタ C_i に対してクラスタ影響力 $I(C_i)$ を算出し,$I(C_i)$ が最大となるクラスタ C_i を $C_{\max}$ とする．

手順 4 $C_{\max}$ に含まれるノード v の中から, 影響力 $\sigma(v \cap S)$ を最大とするノード $v_{\max}$ を S に順次追加する.$|S| = n$ となったら終了する．

手順 5 $|S| < n$ ならば $C_{\max}$ を除外して, 手順 2 に戻る．

まず，各ノード v に対して，嗜好ベクトル x_v を算出する (手順 1)．例えば，ソーシャルメディアデータを対象とした場合は，各ユーザについて，ユーザの発言における単語頻度等から特徴量を抽出し，嗜好ベクトルを構成する．

次に，嗜好ベクトル x_v に基づいて, ノードを嗜好の類似したクラスタ $C_1, C_2, ...C_n$ に分ける (手順 2)．本稿では k-means 法により嗜好ベクトルのクラスタリングを行う. クラスタの個数 n は事前に手で与えるものとする．

次に，生成された各クラスタ $C_1, C_2, ...C_n$ に対して，平均影響力 $I(C_i)$ を算出する (手順 3). クラスタ C_i に対する平均影響力 $I(C_i)$ は

$$I(C_i) = \frac{1}{|C_i|}\sum_{v\in C_i} \sigma(\{v\})$$

として定義される．ここで情報配信先の集合 S の影響力 $\sigma(S)$ は，S に含まれるノードが全てアクティブとなった際に，アクティブになるノード数の期待値である．すなわち平均影響力 $I(C_i)$ は，C_i に含まれる情報配信先 1 つがアクティブとするノード数の平均値である．また，$\sigma(S)$ は予測する必要があるが，本稿では [Kempe 03] で提案されているモンテカルロシミュレーションを用いた方法で算出する．

次に，$C_{\max}$ に含まれるノード v について，影響力 $\sigma(v \cap S)$ を算出し，影響力の大きい順に S に加えていく (手順 4)．

以上を行った段階で，情報配信先の数が n に満たない場合は，$C_{\max}$ 以外のクラスタについて，手順 2 以降を適用する (手順 5)．

提案手法では, 嗜好の類似したクラスタを作成し，影響力の大きいクラスタ内で情報配信先を決定する．これにより，嗜好が類似しており，かつ影響力の大きいノード集合を情報配信先とでき，情報伝播効率を上げることができる．

4. 評価実験

4.1 人工データを用いた評価

はじめに，人工的なネットワークを生成して評価を行った．一般的なネットワーク生成モデルの一つである Barabasi-Albert モデル [Barabási 99] に従い, ノード数 200, 辺数 390 のネット

ワークを生成した. ネットワークの各辺に対して, 情報伝播確率 $p_{v,w}$ を $[0.1, 0.05, 0.01]$ のいずれかからランダムに割り当てた. ネットワークの各ノードに対して 2 次元の嗜好ベクトル x_v を, 混合数 4 の混合正規分布から生成して割り当てた. これはノードが 4 種類の嗜好を持った状況に対応する. また, 情報伝播モデルにおける距離関数 d はユークリッド距離とし, パラメータ a は定数とした.

生成した人工データを用いて, 嗜好を加味しない既存手法である Kempe らの情報配信先決定手法 [Kempe 03] との比較を行った. 情報配信数 $n = 5, 10, 20, 30$ としたとき, 提案手法及び既存手法により情報配信先 S を決定し, 3.2 節で提案する情報伝播モデルにおいて, アクティブノード数の期待値 (以下, 情報受信者数とよぶ)N を比較した.

図 1 に, 提案手法 (クラスタ数 $k = 4$) 及び既存手法 (Kempe らの手法) における, 情報配信数 n と情報受信者数 N の関係を示す.

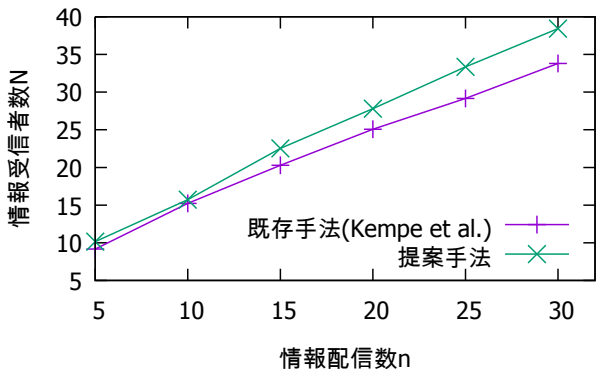

図 1: 既存手法との比較 (人工データ)

図 1 から, 情報配信数 n がいずれの場合でも, 情報受信者数 N が上回ることがわかる. 例えば,$n = 30$ の場合, 既存手法に比べ 13% 情報受信者数 N が増加している. 提案手法では, 嗜好の類似したクラスタ内で影響力の高いノードを選択するため情報適合確率 q_s を高く保ち, 情報受信者数 N を増やすことができる.

また, 提案手法ではクラスタ数 k を事前に決定する必要があることから, クラスタ数 k が情報伝播効率に与える影響を評価した. クラスタ数 k を変更した場合の情報配信数 n と情報受信者数 N の関係を調べた. 提案手法でのクラスタリングにおけるクラスタ数 k を $3, 4, 5$ とした際の比較結果を図 2 に示す.

図 2 より, 図 1 で示したクラスタ数 $k = 4$ の場合と比較して, クラスタ数 $k = 3, 5$ の場合でも情報受信者数 N に大きな差はなく, クラスタ数 k が $3, 4, 5$ のいずれの場合でも, 既存手法と比較して情報受信者数 N が増加していることがわかる.

4.2 Twitter データを用いた評価

次に, マーケティングにおける広告配信を想定し,Twitter データを用いた評価を行った. 評価における問題設定を図 3 に示す.

図 3 において, 広告配信で対象とする製品やサービスを既に利用しているユーザを既存顧客, そうでないユーザを潜在顧客とする. このとき, 既存顧客を対象として広告配信を行い, 潜在顧客を獲得することを目的とする.

評価にあたり, まず以下の手順で Twitter データを収集し, ソーシャルネットワークを構成した.

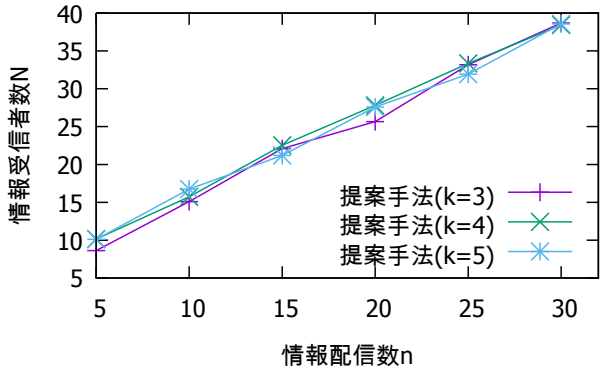

図 2: クラスタ数と情報受信数の比較

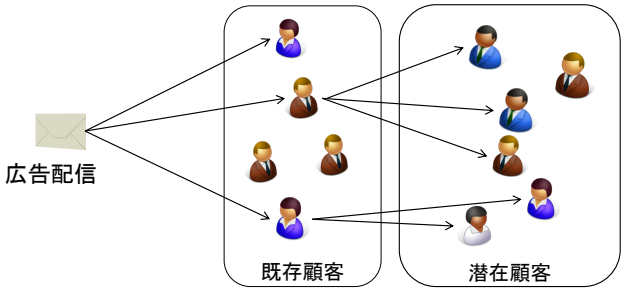

図 3: 評価における問題設定

手順 1 一定期間において, 特定の話題に関するキーワードを含む発言を収集し, 発言を行ったユーザを既存顧客として抽出する.Twitter データの取得には,Twitter API[*3] を用いた.

手順 2 既存顧客からのリプライの中で, 特定の話題に関するキーワードを含むものを収集し, リプライ先であるユーザを抽出する.

手順 3 手順 2 で抽出したユーザのうち, 特定の話題に関するキーワードを含んだ発言を行っていないユーザを潜在顧客として抽出する.

手順 4 各既存顧客 v から潜在顧客 w への発話頻度に基いて, 伝播確率 $p_{v,w}$ を設定する.

手順 5 既存顧客, 潜在顧客それぞれについて, 発言における名詞を対象にした bag-of-words ベクトルを構成し, Latent Dirichlet Allocation[Blei 03] により次元削減を行うことで嗜好ベクトルを作成する.

本稿では, 特定の話題に関するキーワードを「野球」とし, 既存顧客 161 人, 潜在顧客 217 人からなるソーシャルネットワークを構築した.

構築したソーシャルネットワーク上において, 既存手法及び提案手法により既存顧客の中から情報配信先を決定し, その際の情報受信者数を比較した. 情報伝播モデルにおける距離関数 d はユークリッド距離とし, パラメータ a は定数とした.

*3 https://dev.twitter.com/

図 4 において，提案手法及び既存手法 (Kempe らの手法) における, 情報配信数 n と情報受信者数 N の関係を示す.

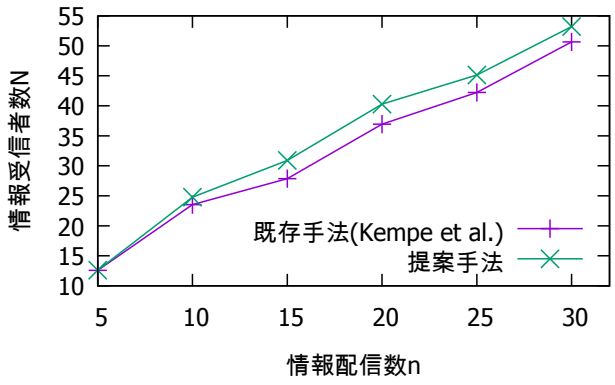

図 4: 既存手法との比較 (Twitter データ)

図 4 から, 情報配信数 n がいずれの場合でも, 提案手法における情報受信者数 N が上回ることがわかる. 既存手法に比べ, 提案手法では最大 10% 程度情報受信者数 N が増加している. 人工データと同様に，提案手法により嗜好の類似した情報配信先集合を選択することで，情報適合確率 q_s を高く保ち，情報受信者数 N を増やすことができる.

5. 結論

本稿では，影響最大化問題において，情報伝播の発生する確率が情報配信先の嗜好の類似性に依存する状況での情報配信先決定手法を提案した．また，人工データ及び Twitter データを用いて情報伝播効率に関して従来手法との比較評価を行った．評価の結果，既存手法に比べ提案手法では最大 10% 程度情報受信者数が増加し，提案手法の有効性を確認した.

今後の課題には，情報伝播モデルにおける嗜好の類似性の影響をデータから推定する方法の検討や，大規模ネットワークを対象とした提案手法の効率化等が挙げられる.

参考文献

[Barabási 99] Barabási, A.-L. and Albert, R.: Emergence of Scaling in Random Networks, *Science*, Vol. 286, No. 5439, pp. 509–512 (1999)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Domingos 01] Domingos, P. and Richardson, M.: Mining the Network Value of Customers, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 57–66 (2001)

[Kempe 03] Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the Spread of Influence Through a Social Network, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 137–146 (2003)

[Saito 11] Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., and Motoda, H.: Learning Diffusion Probability Based on Node Attributes in Social Networks, in *Proceedings of the 19th International Conference on Foundations of Intelligent Systems*, ISMIS'11, pp. 153–162 (2011)

[Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 261–270 (2010)