1H5-3

位置情報を考慮した統計モデルに基づく観光スポットのランキング手法

A Ranking Method for Attractions Based on Statistical Model Reflecting Spatial Trust Factor

山岸 祐己 *1　　斉藤 和巳 *1
Yuki YAMAGISHI　　Kazumi SAITO

静岡県立大学大学院経営情報イノベーション研究科
Graduate School of Management and Information of Innovation, University of Shizuoka

We propose a new item ranking method that is reliable and can efficiently identify high-quality items from among a set of items in review sites using their review scores which were rated and posted by users. Typical ranking methods rely only on either the number of reviews or the average review score. Some of them discount outdated ratings by using a temporal-decay function to make a fair comparison between old and new items. The proposed method reflects trust levels by incorporating a trust discount factor into a spatial-decay function. We bring in the notion of z-score to accommodate the trust variance that comes from the number of reviews available, and propose a z-score version of our statistical model. Finally we demonstrate the effectiveness of the proposed method using the TripAdvisor dataset.

1. はじめに

レビューサイトにおけるレビュー対象オブジェクトのランキングは，殆どの場合，公表されていないサイト独自の手法か，レビュー投稿数やレビュー平均評点といったナイーブなソーシャル情報によって生成されている．確かに，ランキングの秩序を守るためには，独自の方法で最適化を図り，その手法を公表しないというのも重要であるが，その不透明性故に，ユーザからランキングの信頼性を懸念される可能性も大いにある．更に，ナイーブなソーシャル情報によるランキングは，Salganikらの大規模な実験 [10] において，個々の意思決定に多大な影響を与え，市場の不平等性を大いに増加させる要因として明確に示されている．よって，仕組みが明確且つ，ナイーブなソーシャル情報のみに依存しないような，統計モデルに基づくオブジェクトランキングの考案は重要であると言える．

本来ランキングというものは，オブジェクト集合から効率的に高品質なものを見分けるために必要とされている．しかし，レビューサイトでのランキングは，ユーザから提供される情報のみに基づいているため，オブジェクトが登録された時期や，オブジェクトの実際の位置によって，有利不利が生じる可能性が高い．新しいオブジェクトと古いオブジェクトを平等に評価する問題に対しては，時間減衰関数 [1] [8] というものが頻繁に用いられる．実際，時間減衰の考え方は，ソーシャルメディアマイニングの様々な状況において，既にパフォーマンス向上の功績を収めている．例えば，Koren [6] は，時間減衰関数を用いた time-drifting user-preference モデルを提案している．加えて，情報拡散過程の時間減衰影響度は，情報拡散モデル上の情報伝播確率の導入において扱われている [3] [4] [9]．また，投票者モデル [11] [2] の意見形成モデルにおいても，時間減衰関数を組み込んだ手法が提案されている [5]．今回扱う観光スポットのレビューデータは，情報の信頼性が登録時期よりも実際の位置に依存してる可能性が高いため，我々は時間減衰と同様の考え方で，情報の信頼性を考慮することを目的とした空間減衰関数を導入する．

連絡先: 山岸 祐己，静岡県立大学大学院経営情報イノベーション研究科，静岡県静岡市駿河区谷田 52-1，054-264-5436，yamagissy@gmail.com

2. ランキング手法

時刻区間 $\mathcal{T}$ において，整数の評点 $\mathcal{K} = \{1, \cdots, K\}$ によってユーザに評価されたレビュー対象オブジェクトを $\mathcal{V}$ とすると，レビュー集合は $\mathcal{D} = \{(v, k, t) \mid v \in \mathcal{V}, k \in \mathcal{K}, t \in \mathcal{T}\}$ のように書き表せる．任意の $v \in \mathcal{V}$ と $t \in \mathcal{T}$ に対し，時刻 t 以前の時刻 τ からなる v のレビュー集合を $M(v,t) = \{\tau \mid (v,k,t) \in \mathcal{D}, \tau < t\}$ とする．そして，時刻 t におけるオブジェクト v の評点を $g(v,t) \in \mathcal{K}$ とし，$k \in \mathcal{K}$ に対する $M(v,t)$ の部分集合を $M_k(v,t) = \{\tau \in M(v,t) \mid g(v,\tau) = k\}$ とする．いま我々は，過去に投稿された全ての評点を考慮した多項分布モデルを定義する．すなわち，観測されたデータから時刻 t におけるオブジェクト v のレビュー評点分布を予測する以下のモデルを考える．

$$P(g(v,t) = k) = \frac{1 + |M_k(v,t)|}{K + |M(v,t)|}, \ (k = 1, \cdots, K). \qquad (1)$$

ここで，事前分布にはラプラススムージングを施している．このラプラススムージングは，ベイズ統計学において事前分布として頻繁に用いられるディリクレ分布の特殊ケースに相当する．このモデルを基本多項分布モデルとする．

ここから，観測されたデータを用いた，上記のモデルに基づくオブジェクトランキング手法を提案する．時刻区間 $\mathcal{T}$ における平均評点と標準偏差は，それぞれ $\mu = \sum_{k \in \mathcal{K}} kp(k)$, $\sigma = \sqrt{\sum_{k \in \mathcal{K}} (k-\mu)^2 p(k)}$ のように算出される．ここで，$p(k) = \sum_{v \in \mathcal{V}} |M_k(v,T)| / \sum_{v \in \mathcal{V}} |M(v,T)|$ であり，T は $T = \max\{t \in \mathcal{T}\}$ で定義される最終観測時刻である．各レビュー評点が，評点分布 $p(k)$ に従って独立に与えられたと仮定すると，$\mathcal{Q}$ 個のレビュー $S = \{k_1, \cdots, k_{\mathcal{Q}}\}$ が投稿されたときの，期待される平均評点の偏差は以下となる．

$$\begin{aligned} RMSE &= \sqrt{\sum_{k_1 \in \mathcal{K}} \cdots \sum_{k_{\mathcal{Q}} \in \mathcal{K}} \left(\mu - \frac{1}{\mathcal{Q}} \sum_{q=1}^{\mathcal{Q}} k_q \right)^2 \prod_{q=1}^{\mathcal{Q}} p(k_q)} \\ &= \sqrt{\left\langle \left(\mu - \frac{1}{\mathcal{Q}} \sum_{q=1}^{\mathcal{Q}} k_q \right)^2 \right\rangle} \end{aligned}$$

$$= \sqrt{\frac{1}{\mathcal{Q}^2}\left\langle\left(\sum_{q=1}^{\mathcal{Q}}(k_q-\mu)\right)^2\right\rangle}$$
$$= \sqrt{\frac{1}{\mathcal{Q}^2}\left\langle\sum_{q=1}^{\mathcal{Q}}(k_q-\mu)^2+\sum_{x\in\mathcal{Q}}\sum_{q\in\mathcal{Q},q\neq x}(k_x-\mu)(k_q-\mu)\right\rangle}, \quad (2)$$

ここで，$\langle(k_q-\mu)^2\rangle$ は定義によるところの分散 σ^2 であり，$\langle k_q\rangle=\mu$ なので，

$$RMSE = \sqrt{\frac{1}{\mathcal{Q}^2}\sum_{q=1}^{\mathcal{Q}}\sigma^2} = \sqrt{\frac{\sigma^2}{\mathcal{Q}}} = \frac{\sigma}{\sqrt{\mathcal{Q}}}. \quad (3)$$

よって，オブジェクト v の平均評点の z-score は，以下のように考えることができる．

$$z(v)=\frac{\mu(v)-\mu}{\sigma/\sqrt{|M(v,T)|}},\ \mu(v)=\sum_{k\in\mathcal{K}}k\frac{|M_k(v,T)|}{|M(v,T)|}. \quad (4)$$

位置情報を有するレビュー対象オブジェクトを評価する場合，単純に集合全体の情報を考慮した基準を使用するより，位置が近いオブジェクトの情報を強く，位置が遠いオブジェクトの情報を弱く考慮した基準を使用した方が，地理的な有利不利が起こりにくいことが自然と想定できる．この考え方をモデルに反映するために，我々は空間的信頼減衰関数を導入する．単純な一手法としては，$\exp(-\lambda\Delta d)$ のような指数減衰関数が挙げられる．ここで，$\lambda\geq 0$ はパラメータであり，Δd は空間的差異を意味する．一般に， Web 上で得られる位置情報は緯度と経度であるため，オブジェクト v の緯度を $a_v\in\mathcal{A}=\{a_1,\cdots,a_V\}$，経度を $b_v\in\mathcal{B}=\{b_1,\cdots,b_V\}$ とし，それぞれの次元に対応したパラメータを $\boldsymbol{\lambda}=\{\lambda_a,\lambda_b\}^T$ と設定すれば，オブジェクト v，w 間の情報信頼度の重みは以下のように算出できる．

$$\rho(v,w;\boldsymbol{\lambda})=\frac{2\exp(-\lambda_a|a_v-a_w|)\exp(-\lambda_b|b_v-b_w|)}{\exp(-\lambda_a|a_v-a_w|)+\exp(-\lambda_b|b_v-b_w|)}. \quad (5)$$

ここで，信頼度の重みが片方の次元に強く影響を受けないよう，2 値の調和平均をとっている．この $\rho(v,w;\boldsymbol{\lambda})$ を用いれば，各オブジェクト v に対する新たな基準となる評点分布は

$$p_\rho(v,k)=\frac{\sum_{w\in\mathcal{V}}|M_k(w,T)|\rho(v,w;\boldsymbol{\lambda})}{\sum_{w\in\mathcal{V}}|M(w,T)|\rho(v,w;\boldsymbol{\lambda})}, \quad (6)$$

となり，それに伴い $\mu_\rho(v)=\sum_{k\in\mathcal{K}}kp_\rho(v,k)$，$\sigma_\rho(v)=\sqrt{\sum_{k\in\mathcal{K}}(k-\mu_\rho(v))^2p_\rho(v,k)}$ となる．よって，空間的信頼減衰関数を導入したときのオブジェクト v の平均評点の z-score は，以下のように拡張される．

$$z_\rho(v)=\frac{\mu(v)-\mu_\rho(v)}{\sigma_\rho(v)/\sqrt{|M(v,T)|}}. \quad (7)$$

3. データセット

今回使用するデータセットは，TripAdvisor [*1] における，日本の観光スポットのレビューデータである．このデータセットは，緯度と経度を有する観光スポットのみを扱っており，スポット数 N は 11353，総レビュー数は 323868，レビュー評点は 1 から 5 の整数値 ($k\in\mathcal{K}=\{1,\cdots,5\}$) となっている．このデータセットにおける，緯度と経度の差異 Δd の相対度数分布を図 1 に示す．この相対度数分布に基づき，最小二乗法で求めた $\exp(-\lambda\Delta d)$ のパラメータが図 2 である．このパラメータを用いると，今回の実験における $\rho(v,w;\boldsymbol{\lambda})$ は図 3 のようになる．

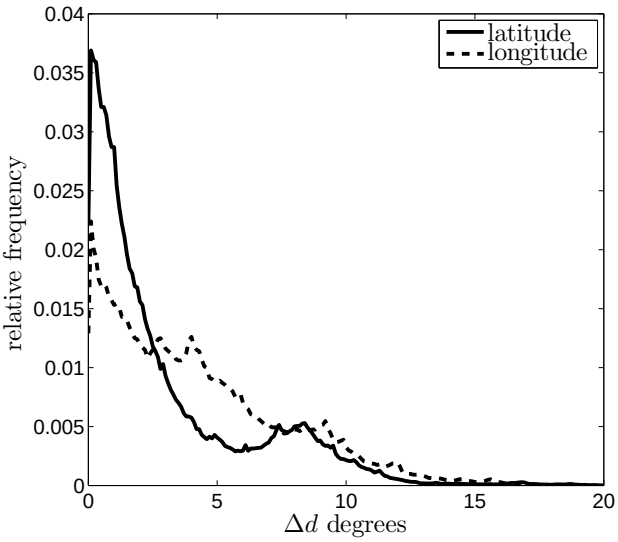

図 1: 緯度と経度の差異 Δd の相対度数分布

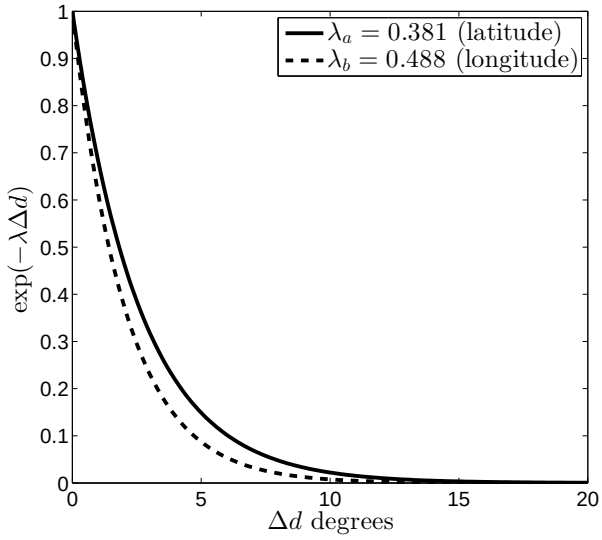

図 2: 指数減衰関数 $\exp(-\lambda\Delta d)$

4. 実験結果

今回のデータセットにおける評価の基準 $\mu_\rho(v)$ の分布を図 4 に示す．これを用いて算出した各スポット v の評価値

*1 http://www.tripadvisor.com/

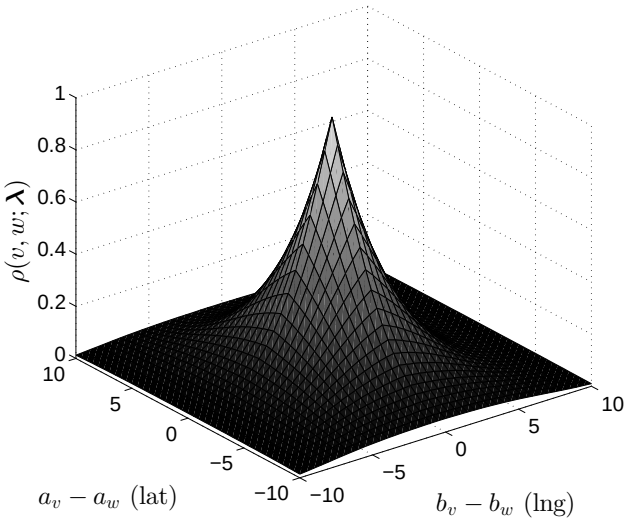

図 3: 空間的信頼減衰関数 $\rho(v, w; \boldsymbol{\lambda})$

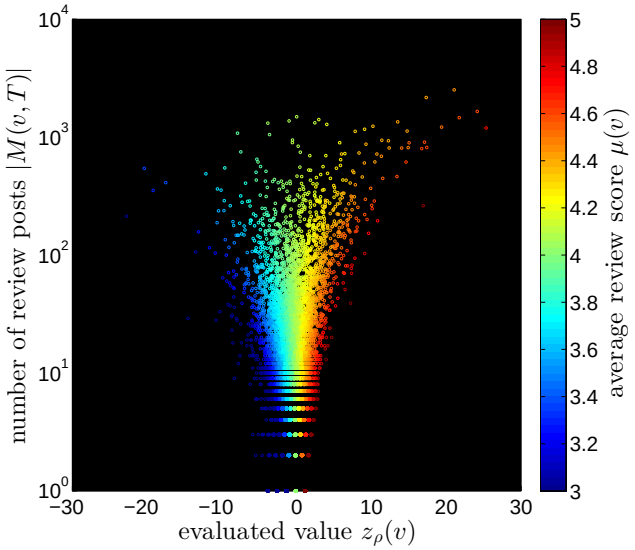

図 5: 投稿されたレビュー数 $|M(v,T)|$ とレビュー平均評点 $\mu(v)$ と提案評価値 $z_\rho(v)$ の関係

$z_\rho(v)$ が図 5 である．図より，$z_\rho(v)$ は，投稿されたレビュー数 $|M(v,T)|$ が多くなるほど評価値の幅が広がるようになっており，単にレビュー平均評点 $\mu(v)$ が高い（又は低い）だけで評価値が極端に高く（又は低く）なっていないことがわかる．また，スポットの位置によって基準となる $\mu_\rho(v)$ が変化するため，投稿数が同程度でも，評価値の大小がレビュー平均評点に完全に準じていないことに注意されたい． 我々は，この提案

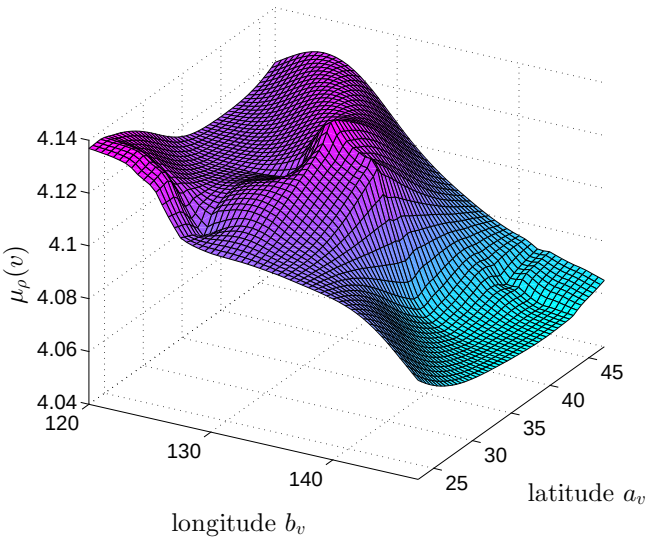

図 4: 緯度 a_v と経度 b_v と評価基準 $\mu_\rho(v)$ の関係

評価値 $z_\rho(v)$ を *proposed*，空間的信頼減衰を考慮しない，すなわち $\rho(v, w; \boldsymbol{\lambda}) = 1$ とした提案評価値を *simple*，投稿されたレビュー評点の合計値 $\sum_{k \in \mathcal{K}} k|M_k(v,T)|$ を *naive* として，それぞれのランキングの地理的な平等性を定量的に評価する．この評価には，以下に述べるカテゴリー評価法の評価値の分散を用いる．

5. カテゴリー評価法

5.1 問題設定

与えられたオブジェクト集合とカテゴリー集合をそれぞれ $\mathcal{I}$ と $\mathcal{J}$ とする．ここで，それぞれの要素数は $I = |\mathcal{I}|$ と $J = |\mathcal{J}|$ とし，各要素は整数と同一視されるとする．つまり，$\mathcal{I} = \{1, \cdots, i, \cdots, I\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ とする．また，オブジェクト i が属すカテゴリーを $j = f(i)$ で表し，各カテゴリに属すオブジェクト数を $I_j = |\mathcal{I}_j| = |\{i\ ;\ j = f(i)\}|$ とする．各オブジェクト i に対し，そのランキングは $1 \leq r_i \leq I$ で与えられるとする．ただし，同順位が起こるケースでは，r_i は平均順位で補正されるとする．

ここでの目的は，カテゴリーとランキング付きのオブジェクトの集合が与えられたとき，ランキングの高い，または逆に低いオブジェクトが有意に多く含まれるカテゴリーを定量的に評価する指標の構築である．以下には，Mann-Whitney の統計量 [7] に基づく自然な拡張法を示す．

5.2 多群順位統計量

Mann-Whitney の二群順位統計量を多群に拡張して適用する方法について述べる．いま，カテゴリー j に着目すれば，このカテゴリーに属すオブジェクト集合 $\mathcal{I}_j$ と，それ以外のオブジェクト集合 $\mathcal{I} \setminus \mathcal{I}_j$ の二群に分割することができる．ここで，$\cdot \setminus \cdot$ は集合差を意味する．よって，Mann-Whitney の二群順位統計量に従い，次式により，カテゴリー j に対し z-score $\dot{z}_j$ を求めることができる．

$$\dot{z}_j \quad = \quad \frac{\dot{u}_j - \dot{\mu}_j}{\dot{\sigma}_j}. \tag{8}$$

ここで，統計量 u_j，順位の平均 $\dot{\mu}_j$，および，その分散 $\dot{\sigma}_j^2$ は次のように計算される．

$$\dot{u}_j \quad = \quad I_j(I - I_j) + \frac{I_j(I_j + 1)}{2} - \sum_{i \in \mathcal{I}_j} r_i, \tag{9}$$

$$\dot{\mu}_j = \frac{I_j(I-I_j)}{2}, \quad (10)$$

$$\dot{\sigma}_j^2 = \frac{I_j(I-I_j)(I+1)}{12}. \quad (11)$$

ただし，同順位が起こるケースでは，標準偏差 σ_j は標準的な方法で補正されるとする．よって，式 (8) で求まる z-score z_j により，各カテゴリー j がランキングの高い，または逆に低いオブジェクトを有意に多く含むか定量的に評価することができる．

既に述べているように，この多群順位統計量は，基本的には 2 クラス分類器の SVM (Support Vector Machine) [12] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる．

5.3 ランキング比較

各スポットを i，TripAdvisor において定められている地域をカテゴリー j としたときの，naive, simple, proposed のそれぞれの評価値のランキングにおけるカテゴリー評価値 z_j の分散を図 6 に示す．この分散が大きい（又は小さい）ということは，ランキングの上位と下位で地域差が大きい（又は小さい）と考えることができる．図より，基本多項分布モデルに基づいている simple は，スポットの単純な人気度に基づいている naive よりも地域差が小さく，また，空間的信頼減衰を考慮した proposed は，更にその simple よりも地域差が小さいことがわかる．

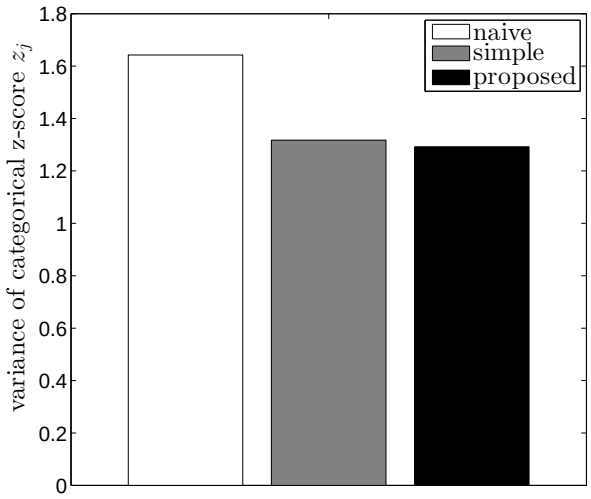

図 6: カテゴリー評価値 z_j の分散

6. まとめ

レビューサイトにおけるユーザの基本評点行動として多項分布モデルを仮定し，投稿されたレビュー数とその平均評点を，統計モデルに基づく評価値に変換した．更に，情報の地理的な信頼性を考慮することを目的とした，空間減衰関数の導入を試みた．提案評価値によるランキングは，ナイーブな評価値によるランキングと比較して，地域による不平等性が低いことを示し，空間減衰関数の導入は，その不平等性を更に低くすることを示した．今後は，空間的信頼と時間的信頼の両方を考慮した評価値を考案する予定である．

謝辞

本研究は，総務省 SCOPE (No.142306004)，及び，科学研究費補助基金基盤研究 (C)(No.25330635) の支援を受けて行ったものである．

参考文献

[1] G. Cormode, V. Shkapenyuk, D. Srivastava, and B. Xu, “Forward decay: A practical time decay model for streaming systems,” in *Proc. of ICDE09*, pp. 138–149, 2009.

[2] E. Even-Dar, and A. Shapira, “A note on maximizing the spread of influence in social networks.,” in *Proc. of WINE’07*, pp. 281–286, LNCS 4858, 2007.

[3] J. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,” *Marketing Letters 12*, pp.211–223, 2001.

[4] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. of KDD’03*, pp. 137–146, 2003.

[5] M. Kimura, K. Saito, K. Ohara, and H. Motoda, “Opinion formation by voter model with temporal decay dynamics,” in *Proc. ECML-PKDD’12*, pp. 565–580, LNCS 7524, 2012.

[6] Y. Koren, “Collaborative filtering with temporal dynamics,” in *Proc. of KDD’09*, pp. 447–456, 2009.

[7] H. B. Mann, and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other”, *Ann. Math. Statist., vol. 18, no. 1*, pp. 572–578, 1947.

[8] G. Papadakis, C. Niederée, and W. Nejdl, “Decay-based ranking for social application content,” in *Proc. of WEBIST’10*, pp. 276–281, 2010.

[9] K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Learning asynchronous-time information diffusion models and its application to behavioral data analysis over social networks,” *Journal of Computer Engineering and Informatics 1*, pp. 30–57, 2013.

[10] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market”, *Science, vol. 311*, pp. 854–856, 2006.

[11] V. Sood, and S. Redner, “Voter model on heterogeneous graphs,” *Physical Review Letters 94*, 17801, 2005.

[12] V. Vapnik, “The nature of statistical learning theory”, *Springer-Verlag New York, Inc.*, 1995.