

時空間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 Spatio-Temporal Trust Factor

山岸祐己*
Yuki Yamagishi

斉藤和巳*
Kazumi Saito

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、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や、公表されていないサイト独自の手法、とりわけユーザ依存の信頼を用いたもの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、独自の手法で信頼性の低いユーザのレビュー情報を淘汰し、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、ユーザがランキングの信頼性を懸念する可能性も大いにある。更に、ユーザに提供するナイーブなソーシャル情報の項目数を増やすと、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ことが Salganik らの大規模実験 [13] によって既に分かっている。よって、既存のランキングに対する代替案の一つとして、ユーザ依存の信頼ではなくオブジェクト依存の信頼を利用し、ナイーブなソーシャル情報のみならず依存しないような、統計モデルに基づくランキングの構築が考えられる。

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、それらの市場における有利不利が生じている可能性が高いということが推察できる。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] [11] 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、推薦システムにおいて、時間減衰関数を用いたモデルを提案している。加えて、情報拡散過程の時間減衰による影響は、情報拡散モデル上の情報伝播確率の導入において頻繁に扱われている [3] [4] [12]。また、投票者モデル [14] [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]。当然、今回扱うような観光スポットのレビューデータも、情報の信頼性が登録時期に依存している可能性が高いため、時間減衰関数の有効性が期待できる。更に、今回のデータは正確な位置情報も有しているため、位置による不平等性問題を考えることもできる。よって我々は、情報の信頼性を考慮することを目的として、時間減衰関数と、それと同様の考え方に基づく空間減衰関数の両方を導入したモデルを構築する。ここで、我々が指している信頼とはオブジェクト依存のものである

ため、推薦システムの研究 [7] [10] で頻繁に用いられているような、ユーザ依存の信頼とは異なることに注意されたい。

2. ランキング手法

時刻区間 \mathcal{T} において、整数の評点 $\mathcal{K} = \{1, \dots, K\}$ 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 $M(v, t) = \{\tau \mid (v, k, \tau) \in \mathcal{D}, \tau < t\}$ 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。

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

ここで、我々は Laplace スムージングとして知られるベイズ事前分布を用いた。式 1 の Laplace スムージングは、各オブジェクトが最初に等確率で $1, \dots, K$ の各評点で評価されたことを仮定している。また、この Laplace スムージングは、ベイズ統計における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しており、実際、ディリクレ分布は多項分布の共役事前分布である。このモデルを基本多項分布モデルとする。

ここから、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、 Q 個のレビュー $S = \{k_1, \dots, k_Q\}$ 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。なお、ここで $\langle \cdot \rangle$ は期待値を表す。

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

*静岡県立大学, University of Shizuoka

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

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、

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

よって、時刻 t におけるオブジェクト v の平均評点の z-score $z(v, t)$ は、以下のように考えることができる。

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

明らかに、この $z(v, t)$ が大きいオブジェクトは、統計的に有意に高い評価を得ているとみなすことができる。

ここまで、過去に投稿された全てのレビューは同じ重みであると仮定してきたが、過去と現在で評価の揺らぎがある場合は、古いレビューの信頼度は低くなると考えることができる。更に、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、位置に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も自然と想定できる。これらの考え方をモデルに反映するために、時空間的信頼減衰関数を導入する。単純な手法としては、 $\exp(-\lambda \Delta \cdot)$ 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、 $\Delta \cdot$ は時空間的差異を意味する。

時間的信頼減衰は、オブジェクト v 単体における問題であるとし、その減衰を $\rho_\alpha(\Delta t; \lambda_v) = \exp(-\lambda_v \Delta t)$ で定めるとすると、基本多項分布モデルの式 (1) は、

$$P(g(v, t) = k) = \frac{1 + \sum_{\tau \in M_k(v, t)} \rho_\alpha(t - \tau; \lambda_v)}{K + \sum_{\tau \in M(v, t)} \rho_\alpha(t - \tau; \lambda_v)}, \quad (5)$$

のように拡張することができる。次に、空間的信頼減衰は、オブジェクト全体 \mathcal{V} における問題であると考えられる。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、GRS80 [9] 準拠楕円体に基づいた、オブジェクト v, w 間の標高を無視した地表面距離を $\Delta d_{v, w}$ とすると、その減衰は $\rho_\beta(\Delta d_{v, w}; \lambda_d) = \exp(-\lambda_d \Delta d_{v, w})$ のように定めることができる。これらの時空間的信頼減衰を考慮し、推定パラメータを $\hat{\lambda}_v, \hat{\lambda}_d$ とすれば、各オブジェクト v に対する新たな基準となる評点確率分布は

$$p(v, k) = \frac{\sum_{w \in \mathcal{V}} \sum_{\tau \in M_k(w, T)} \rho_\alpha(t - \tau; \hat{\lambda}_v) \rho_\beta(\Delta d_{v, w}; \hat{\lambda}_d)}{\sum_{w \in \mathcal{V}} \sum_{\tau \in M(w, T)} \rho_\alpha(t - \tau; \hat{\lambda}_v) \rho_\beta(\Delta d_{v, w}; \hat{\lambda}_d)}, \quad (6)$$

となり、それに伴い各オブジェクト v に対して期待される平均評点と標準偏差は、それぞれ $\mu(v) = \sum_{k \in \mathcal{K}} k p(v, k)$, $\sigma(v) = \sqrt{\sum_{k \in \mathcal{K}} (k - \mu(v))^2 p(v, k)}$ となる。よって、このときの時刻 t におけるオブジェクト v の平均評点の z-score $z_\rho(v, t)$ は、以下のように置き換えられる。

$$\begin{aligned}
z_\rho(v, t) &= \frac{\mu_\rho(v, t) - \mu(v)}{\sigma(v) / \sqrt{\sum_{\tau \in M(v, t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}}, \\
\mu_\rho(v, t) &= \sum_{k \in \mathcal{K}} k \frac{\sum_{\tau \in M_k(v, t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}{\sum_{\tau \in M(v, t)} \rho_\alpha(t - \tau; \hat{\lambda}_v)}. \quad (7)
\end{aligned}$$

この拡張モデルにおける時間的信頼減衰の推定パラメータ $\hat{\lambda}_v$ は、観測データ \mathcal{D} に対するオブジェクト v の対数尤度関数、

$$\mathcal{L}(M(v, T); \lambda_v) = \log \left(\prod_{(v, k, t) \in M(v, T)} P(g(v, t) = k) \right), \quad (8)$$

を最大化することによって得ることが可能である。この対数尤度関数は、式 (5) と $\rho_\alpha(\Delta t; \lambda_v) = \exp(-\lambda_v \Delta t)$ より、

$$\begin{aligned}
\mathcal{L}(M(v, T); \lambda_v) &= \sum_{(v, k, t) \in M(v, T)} \log \left(1 + \sum_{\tau \in M_k(v, t)} \exp(-\lambda_v(t - \tau)) \right) \\
&\quad - \sum_{(v, k, t) \in M(v, T)} \log \left(K + \sum_{\tau \in M(v, t)} \exp(-\lambda_v(t - \tau)) \right), \quad (9)
\end{aligned}$$

と書き直せる。この尤度関数を最大化するパラメータ $\hat{\lambda}_v$ を EM アルゴリズムで推定する。このとき、 Q 関数のヘス行列は半負定値となるため、ニュートン法を用いて Q 関数の大域最適解を得ることができる。また、空間的信頼減衰の推定パラメータ $\hat{\lambda}_d$ は、 $\exp(-\lambda_d \Delta d_{v, w})$ の最尤推定量 $1/E(\Delta d_{v, w})$ を用いる。

3. 実験

3.1. データセット

今回使用するデータセットは、TripAdvisor² に登録されている、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、日本人ユーザのレビュー情報と、英語圏ユーザのレビュー情報で構成されている。取得時期は2014年10月、スポット数 $|\mathcal{V}|$ は11353、総レビュー数 $|\mathcal{D}|$ は296221、レビュー評点は1から5の整数値 ($k \in \mathcal{K} = \{1, \dots, 5\}$) となっている。参考までに、レビュー評点の相対度数分布を図1に示す。

3.2. 実験結果

まず、今回のデータセットにおける、各スポット v の時間的信頼減衰関数の推定パラメータ $\hat{\lambda}_v$ と被レビュー数 $|M(v, T)|$ の関係を図2に示す。図から、被レビュー数が多いスポットほど、推定パラメータが低くなる傾向が見て取れる。つまり、被レビュー数が多いにも関わらず推定パラメータが高いスポットは、直近と過去

²<http://www.tripadvisor.com/>

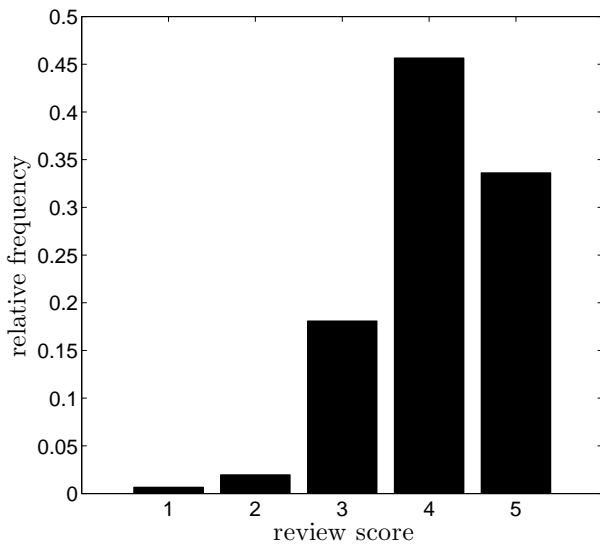
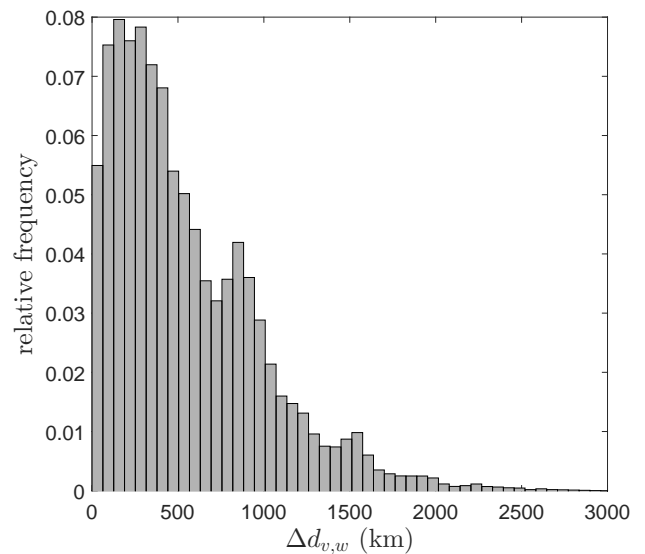
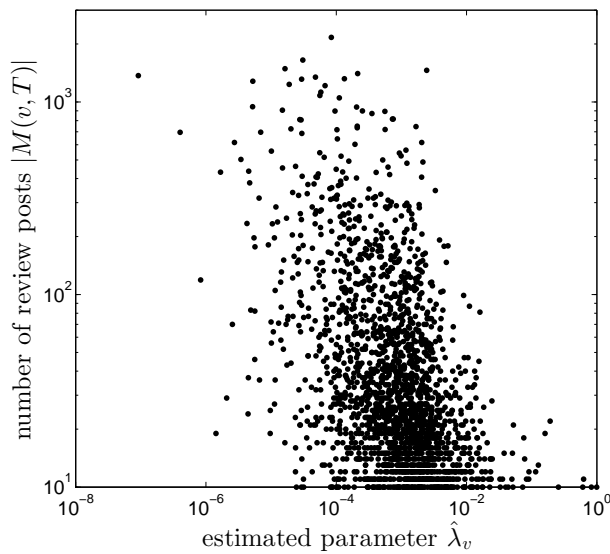
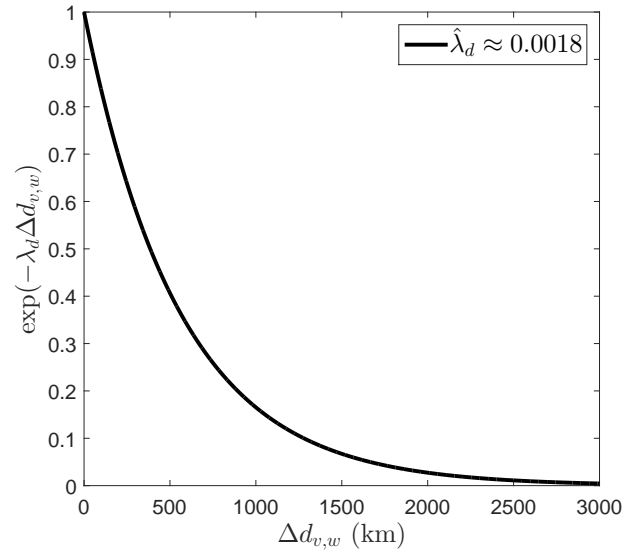


図 1: レビュー評点の相対度数分布

図 3: オブジェクト間距離 $\Delta d_{v,w}$ の相対度数分布図 2: 推定パラメータ $\hat{\lambda}_v$ と被レビュー数 $|M(v, T)|$ の関係図 4: 最尤推定量 $\hat{\lambda}_d$ と空間的信頼減衰関数 $\exp(-\lambda_d \Delta d_{v,w})$

で評価の揺らぎが激しいことが予想できる。次に、今回のデータセットにおけるオブジェクト間距離 $\Delta d_{v,w}$ の相対度数分布を図 3 に示す。そして、データセットから求めた最尤推定量 $\hat{\lambda}_d$ を用いた空間的信頼減衰関数 $\exp(-\lambda_d \Delta d_{v,w})$ を示したのが図 4 である。今回の拡張モデルに基づき、これらの推定パラメータと減衰関数を用いて算出した期待平均評点 $\mu(v)$ の分布は図 5 のようになった。

続いて、基本多項分布モデルに基づいて算出した各スポットの評価値 $z(v, T)$ の分布を図 6 に、拡張モデルと推定パラメータに基づいて算出した各スポット v の

評価値 $z_\rho(v, T)$ の分布を図 7 にそれぞれ示す。図より、両評価値共に、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 $\mu(v, T)$ 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。特に、 $z_\rho(v, T)$ は、過去に投稿されたレビューの影響度と、評価の基準となる期待平均評点 $\mu_\rho(v)$ が v 毎に変化するため、レビュー投稿数が同程度でも、評価値の大小がレビュー平均評点 $\mu(v, T)$ に完全に準じていないことに注意されたい。我々は、基本多項分布モデルに基づく $z(v, T)$ によるランキングを単純法 (*simple*),

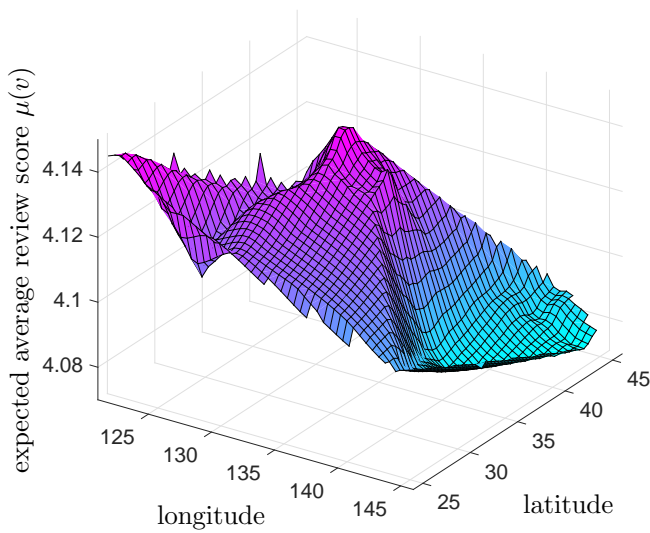


図 5: 推定パラメータと拡張モデルに基づく期待平均評点 $\mu(v)$ の分布

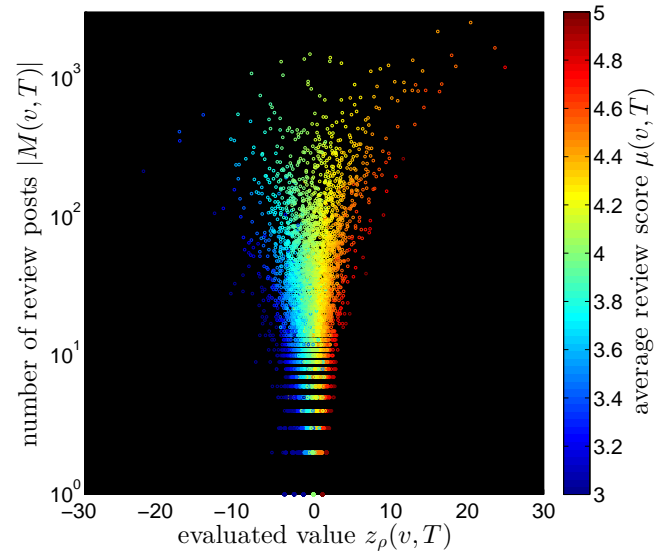


図 7: 投稿されたレビュー数 $|M(v, T)|$ とレビュー平均評点 $\mu(v, T)$ と提案評価値 $z_\rho(v, T)$ の関係

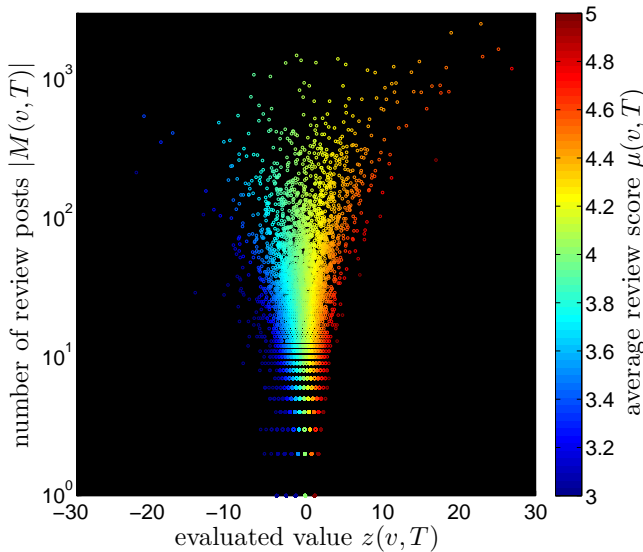


図 6: 投稿されたレビュー数 $|M(v, T)|$ とレビュー平均評点 $\mu(v, T)$ と基本評価値 $z(v, T)$ の関係

拡張モデルと推定パラメータに基づく $z_\rho(v, T)$ によるランキングを提案法 (proposed) とし、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。

4. カテゴリー評価法

4.1. 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、 $\mathcal{I} = \{1, \dots, i, \dots, I\}$ および $\mathcal{J} = \{1, \dots, j, \dots, J\}$ 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、各カテゴリー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 [8] に基づく自然な拡張法を示す。

4.2. 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 \hat{z}_j を求めることができる。

$$\hat{z}_j = \frac{\hat{u}_j - \hat{\mu}_j}{\hat{\sigma}_j}. \quad (10)$$

ここで、統計量 u_j , 順位の平均 μ_j , および、その分散 σ_j^2 は次のように計算される.

$$\hat{u}_j = I_j(I - I_j) + \frac{I_j(I_j + 1)}{2} - \sum_{i \in \mathcal{I}_j} r_i, \quad (11)$$

$$\hat{\mu}_j = \frac{I_j(I - I_j)}{2}, \quad (12)$$

$$\hat{\sigma}_j^2 = \frac{I_j(I - I_j)(I + 1)}{12}. \quad (13)$$

ただし、同順位が起こるケースでは、標準偏差 $\hat{\sigma}_j$ は標準的な方法で補正されるとする。よって、式(10)で求まる z-score \hat{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) [15] を多クラス分類器に拡張するとき利用される one-against-all と類似した考え方となる。

4.3. 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区分をカテゴリー j としたときの、単純法、提案法のそれぞれのランキングにおけるカテゴリー評価値 \hat{z}_j の分布を図 8, 9 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。両図より、僅かではあるが、提案法の方がカテゴリー評価値の散らばりを抑えられていることがわかる。つまり、提案評価値は、基本評価値と比べて地理的な不平等性が低いと言える。この差についての比較を行うため、提案法において時間的信頼減衰を考慮しない、即ち $\hat{\lambda}_v = 0$ で固定した *spatial* と、提案法において空間的信頼減衰を考慮しない、即ち $\hat{\lambda}_d = 0$ で固定した *temporal* で同様の実験を行った。これら4手法におけるカテゴリー評価値 \hat{z}_j の分散比較を示したのが図 10 である。図より、*spatial* と *temporal* も、単純法と比較して地域差が小さくなっているが、これら4手法の比較においても提案法が最も優れていることがわかる。

5. 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の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、情報の時間的信頼性を考慮することを目的として時間減衰関数を、情報の地理的信頼性を考慮することを目的として空間減衰関数を導入した。時空間的信頼減衰を考慮したモデルに基づく提案評価値のランキングは、単純な多項分布モデルに基づく評価値の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た。今後は、空間減衰関数のパラメータも、時間減衰関数のパラメータ同様、オブジェクト毎で推定できるようなモデルを考え、実験と結果の検証を行う予定である。

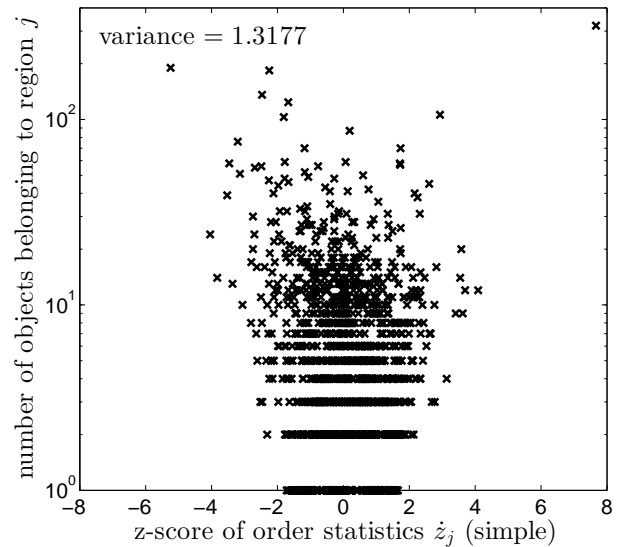


図 8: 単純法におけるカテゴリー評価値 \hat{z}_j の分布

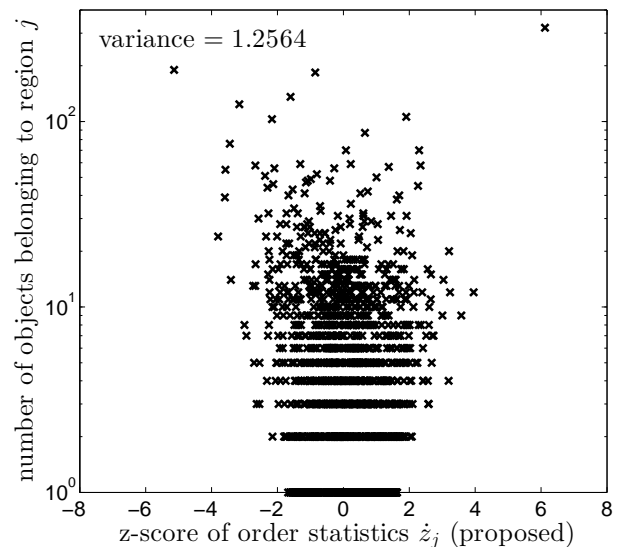


図 9: 提案法におけるカテゴリー評価値 \hat{z}_j の分布

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.

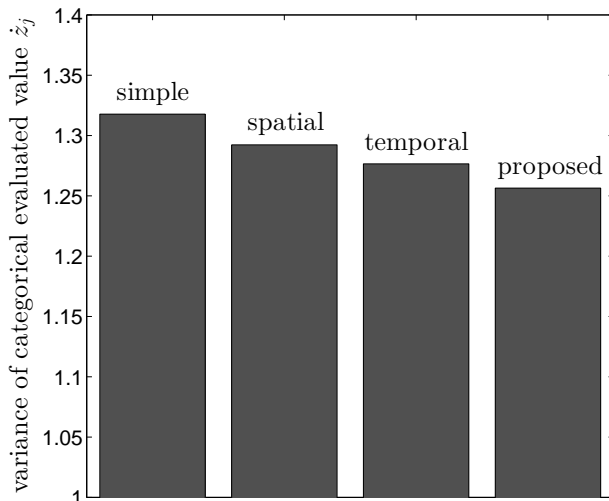


図 10: カテゴリー評価値 z_j の分散比較

- [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. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, “Recommender systems with social regularization,” in *Proc. of WSDM’11*, pp. 287–296, ACM, New York, NY, 2011.
- [8] 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.
- [9] H. Moritz, “Geodetic reference system 1980”, *Journal of Geodesy*, vol. 54:3, pp. 395–405, 1980.
- [10] J. O’Donovan, and B. Smyth, “Trust in recommender systems,” in *Proc. of IUI’05*, pp. 167–174, ACM, New York, NY, 2005.
- [11] G. Papadakis, C. Niederée, and W. Nejdl, “Decay-based ranking for social application content,” in *Proc. of WEBIST’10*, pp. 276–281, 2010.
- [12] 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.
- [13] 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.
- [14] V. Sood, and S. Redner, “Voter model on heterogeneous graphs,” *Physical Review Letters* 94, 17801, 2005.
- [15] V. Vapnik, “The nature of statistical learning theory”, *Springer-Verlag New York, Inc.*, 1995.