D-010

レビューサイトにおける異常検出を用いた重要ユーザ選定法
A Selection Method of Important Users using Anomaly Detection in Online Review Sites

山岸祐己 † Yuki Yamagishi　斉藤和巳 † Kazumi Saito　武藤伸明 † Nobuaki Muto

1. はじめに

レビューサイトとは，商品やサービスについてのレビューを投稿することができるウェブサイトの総称である．レビューは評点・文章・画像から成ることが多く，レビュー評点の平均点が，対象の一般的な評価指標として扱われている．レビューサイトについては，既に多様な分析や研究が展開されている [1].

近年，これらレビューサイトにおけるユーザーのレビュー行動が非常に活発であり，サイトそのものが商品やサービスのプロモーションを左右する重要なメディアになりつつある．しかし，数あるレビューの中から自身にとって有益な情報を探し出すのは簡単ではなく，手作業でレビューを吟味しようとすると膨大な時間を要する．さらに,「やらせ」や「サクラ」に代表されるユーザの異常なレビュー行動も問題視されているため，重要な情報を発信しているレビューやユーザを発見することは非常に難儀であると言える．

そこで我々は，Swan と Allan [2] や Kleinberg [3] と同様に，回顧的 (Retrospective) な立場でレビューの時系列的な異常検出（変化点検出）を行い，それら検出結果を用いて重要ユーザの選定を試みる．提案手法は，ユーザーの基本行動として，レビューの評点を多項分布モデル，レビュー投稿間隔を指数分布モデル，レビューのファイルサイズをガウス分布モデルと仮定し，レビュー時系列データからの変化点検出問題を定式化する．さらに，それら変化点の直前にレビューを投稿したユーザにスコアを付与し，重要ユーザランキングを生成する．本研究の実験では，現実の大規模レビューデータを用いる．

2. 提案手法

2.1. 問題設定

レビュー評点を s_n，レビューファイルサイズを v_n，レビューが投稿された時刻を t_n とし，レビュー時系列データを以下のように表す．

$$\mathcal{D} = \{(s_1, v_1, t_1), \cdots, (s_N, v_N, t_N)\}. \tag{1}$$

ここで各評点は，1 から J の整数値で与えられるとする．即ち，$s_n \in \{1, \cdots, J\}$ となる．モデル記述の都合上，各評点 s_n を以下のように J-次元ベクトルとしてダミー変数を導入する．

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

いま，多項分布モデルを仮定し，評点 j が与えられる確率を p_j とすれば，評点の時系列データの対数尤度関数は次式となる．

$$\mathcal{L}^s(\mathcal{D}; \boldsymbol{p}) = \sum_{n=1}^{N} \sum_{j=1}^{J} s_{n,j} \log p_j. \tag{3}$$

一方，ガウス分布モデルを仮定し，ファイルサイズの平均を μ，標準偏差を σ とすれば，ファイルサイズの時系列データの対数尤度関数は次式となる．

$$\mathcal{L}^v(\mathcal{D}; \mu, \sigma) = \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v_n - \mu)^2}{2\sigma^2}\right). \tag{4}$$

また，指数分布モデルを仮定し，単位時間あたりの平均投稿数を λ とすれば，投稿間隔の対数尤度関数は次式となる．

$$\mathcal{L}^t(\mathcal{D}; \lambda) \quad = \quad \sum_{n=2}^{N} \log \lambda \exp\left(-\lambda(t_n - t_{n-1})\right). \tag{5}$$

いま，K 個の時刻から構成される変化点集合を $\mathcal{C}_K = \{T_1, \cdots, T_K\}$ とし，便宜上 $T_0 = t_1$ かつ $T_{K+1} = t_N$ と設定しておく．また，$T_{k-1} < T_k$ であるとし，$\mathcal{C}_K$ による $\mathcal{D}$ の分割を $\mathcal{D}_k = \{(s_n, v_n, t_n); T_{k-1} < t_n \leq T_k\}$ で定義する．すなわち，$\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{K+1}$ となり，$|\mathcal{D}_k|$ は区間 $(T_{k-1}, T_k]$ に含まれる観測時刻数を表す．ここで，任意の $k \in \{1, \cdots, K+1\}$ に対して，$|\mathcal{D}_k| \neq 0$ とする．一方，各対数尤度関数で用いるパラメータを，区間 $\mathcal{D}_k$ 毎に対応させて $\boldsymbol{\theta}_{K+1} = \{\theta_1, \cdots, \theta_{K+1}\}$ で定義すれば，変化点集合 $\mathcal{C}_K$ が与えられたときの観測データ $\mathcal{D}$ に対する対数尤度は，次式のように一般化して表される．

$$\mathcal{L}(\mathcal{D}; \boldsymbol{\theta}_{K+1}, \mathcal{C}_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}; \theta_k). \tag{6}$$

よって，式 (6) の尤度を最大にするパラメータの最尤推定値を $\hat{\boldsymbol{\theta}}_{K+1}$ とすれば，我々の変化点検出問題は，$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{\theta}}_{K+1}, \mathcal{C}_K)$ を最大化する変化点集合 $\mathcal{C}_K$ を求める問題となる．ただし，変化点集合 $\mathcal{C}_K$ の導入効果を直接的に評価するため，この問題の別表現として，尤度比検定の目的関数として我々の変化点検出問題を定式化する．つまり，変化点が K 個存在するとしたときと，存在しないとしたとき，即ち $\mathcal{C}_0 = \emptyset$ のときの尤度比の対数を次式で定義する．

$$\mathcal{LR}(\mathcal{C}_K) = \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{\theta}}_{K+1}, \mathcal{C}_K) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{\theta}}_1, \mathcal{C}_0). \tag{7}$$

本論文では，式 (7) で定義した $\mathcal{LR}(\mathcal{C}_K)$ を最大化する変化点集合 $\mathcal{C}_K$ を求める問題を考える．この問題に対

†静岡県立大学, University of Shizuoka

し，網羅的な解法を適用すると，計算量は $O(N^K)$ となる．このため，観測点数 N が十分に大きくなると，実用的な時間で結果が得られるのは $K = 2$ 程度までである．よって，実用的な時間で結果を得るための解法として，既に我々が提案した局所改善法 [4] を用いる．この解法は貪欲法が基となっており，必要とする計算量は $O(NK)$ の数倍程度に留まる．さらに，単純な貪欲法の結果から十分解品質が向上することも分かっている．

3. データセット

今回使用したデータセットは，@cosme[1] のレビューデータである．

@cosme は，株式会社アイスタイル[2] が運営する日本最大級の化粧品レビューサイトであり，1999 年 12 月にサービスが開始された．

このデータセットは，2013 年 3 月から 2013 年 4 月にかけて@cosme をクロールして取得したものであり，446839 ユーザー，21211 ブランド，168126 アイテム，8810651 レビューを有する．レビューの評点は，0〜7 の整数値をとりうるが，0 点は「評価としての 0 点」と「無評価」が混在しているため，今回の実験では使用していない．

4. 実験設定

今回の実験対象は，被レビュー回数が 200 以上の 7941 アイテムとし，レビュー評点 ($s_n \in \{1, \cdots, 7\}$) の変化点のみを扱う．一般に，データ数 N が十分に大きいとき，$\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の 2 倍は漸近的に χ^2 分布となることが知られているため，予め設定した有意水準における自由度 $J-1$ の χ^2 の棄却点を変化点の採用基準とする．今回の場合，$2(\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1}))$ が，$p = 0.005$，自由度 6 の χ^2 の棄却点 18.5476 を超える場合には T_k を変化点として採用し，そうでなければそこで変化点の探索を打ち切る．

また，ユーザ u 毎のスコアを A_u，スコアを付与する範囲を整数 $R(>0)$ とし，変化点 T_k のレビュー時刻 t_n から過去に遡る形で $t_{n-b}(b \in \{0, \cdots, (R-1)\})$ のレビューを投稿したユーザの A_u に $R-b$ を加算する．

5. 実験結果とまとめ

図 1 に $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の度数分布を，図 2 に A_u と u のレビュー投稿数の相関係数の推移を示す．

図 1 より，$\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ にはスケールフリー性が見られる．これについては，アイテムの被レビュー数そのものがスケールフリーになっていることが大きく関わっているはずである．よって，分布を考慮した正規化，または，ノンパラメトリックな尺度水準を用いて，T_k 毎に付与するスコアを変化させ，ユーザ u の重要性を有意に計る必要がある．

また，図 2 より，A_u と u のレビュー投稿数の相関係数は，$R = 20$ 辺りで 0.8 を超えているため，R を大きくしすぎると，レビュー投稿数のみでユーザをランキングした場合と差別化ができないことが危惧される．

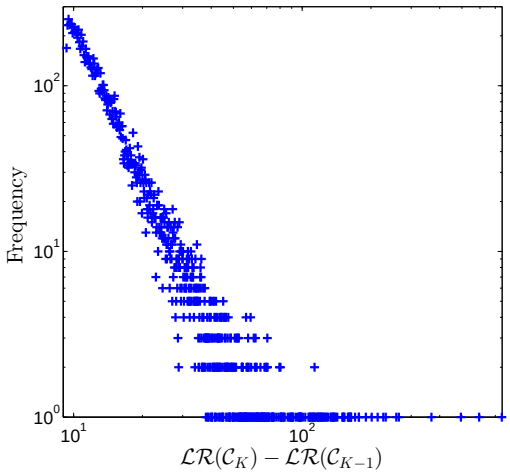

図 1: $\mathcal{LR}(\mathcal{C}_k) - \mathcal{LR}(\mathcal{C}_{k-1})$ の度数分布

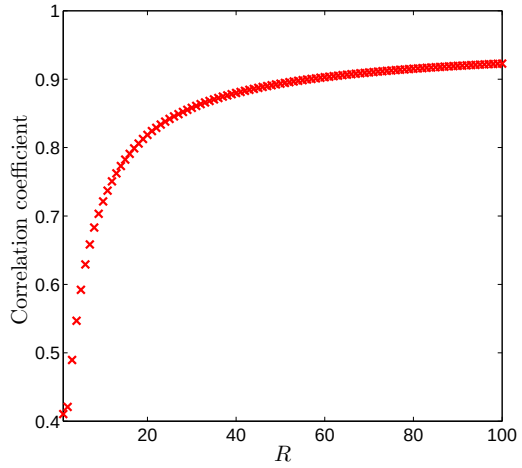

図 2: A_u と u のレビュー投稿数の相関係数の推移

謝辞

本研究は，豊田中央研究所との共同研究，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] M.J.Salganik, P.S.Dodds, and D.J.Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market", Science 10, pp.854-856, February 2006.

[2] R.Swan and J.Allan, "Automatic Generation of Overview Timelines", SIGIR 2000, pp.49-56, 2000.

[3] J.Kleinberg, "Bursty and Hierarchical Structure in Streams", KDD 2002, pp.91-101, 2002.

[4] 山岸 祐己，斉藤 和巳，"オンラインレビューサイトにおけるレビュー変化点検出法"，第 5 回 Web とデータベースに関するフォーラム 2012，2012.

[1] http://www.cosme.net
[2] http://www.istyle.co.jp/