DEIM Forum 2012 F6-6

オンラインレビューサイトの評点時系列データからの異常検出

山岸　祐己† 　斉藤　和巳† 　大久保誠也†

† 静岡県立大学経営情報学部　〒 422–8526 静岡県静岡市駿河区谷田 52-1
E-mail: †{b08107,k-saito,s-okubo}@u-shizuoka-ken.ac.jp

あらまし　インターネット上で展開されている商品レビューサイトにおける，評点時系列データからの異常検出法を提案する．我々は，人間の評点行動のモデル化を第一に考えている．しかし，レビューサイトの評点とそれに付随する文章は，純粋に評価をしているものから意図的なものまで様々である．よって，より現実的で実用性のあるモデルの構築のためにも，レビューサイトにおけるユーザーの異常な評点行動の検出は重要である．提案法は，ユーザーの基本評点行動として多項分布モデルを仮定し，尤度比検定により異常期間を検出することを特徴とする．実験結果より，短期間に一定の評点を評価対象系列に与える特徴的なユーザーの存在や，評価対象に対する評価が著しく変動した期間等，いくつかの異常を検出できたことを示す．

キーワード　レビューサイト, 異常検出, 尤度比検定, 時系列

An Anomaly Detection in Time Series Data of Online Review Sites

Yuki YAMAGISHI†, Kazumi SAITO†, and Seiya OKUBO†

† School of Management and Information, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka, 422–8526 Japan
E-mail: †{b08107,k-saito,s-okubo}@u-shizuoka-ken.ac.jp

Abstract　In recent years, those who write a review at online review sites are increasing. And we are considering a modeling of people's such evaluation action. However, there are a variety of scores and accompanying documents on those sites, going from the pure rating to the intentional or indifferent. Thus, in this paper, we propose a method for anomaly detection in time series data of online review sites. The proposing method assumes a multinomial distribution model as a user's basic evaluation action, and is characterized by detecting an unusual period by a likelihood ratio test. It is shown from an experimental result that anomalies, such as characteristic users who gave same scores to a brand in short period and periods when the evaluation to goods fluctuated remarkably, have been detected.

Key words　Review Site, Anomaly Detection, Likelihood Ratio Test, Time Series

1. はじめに

オンライン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．オンラインレビューサイトについては，多様な分析や研究が展開されている [1]．初期のレビューサイトとして Amazon.com(http://www.amazon.com) がよく挙げられるが，レビューサイトの歴史は各国によって様々であり，日本国内では古くよりゲームソフトのレビューサイトが数多く存在した．そして，インターネットの爆発的な普及により，ネットショッピングの一般化と共にレビューサイトのユーザーが急増し，日々大量のレビューがあらゆるサイトに投稿される現状に至った．結果，1 つの商品に対して多種多様なレビューが付くこととなり，有益なレビューを判別することが難しくなったため，レビューに付随する評点の平均点が一般的な評価指標として扱われるようになった．しかし，殆どのレビューサイトが投稿回数制限やレビュー内容の吟味を行なっていないため，主観的思考が強いユーザーによる極端な評価が書かれたレビューも飛び交い，この評点平均ですら信頼性を失いつつある．あまりにも肯定的なレビューは，商品の製造会社や関係会社が意図的に書いたのではないかと疑われ，あまりにも否定的なレビューは，競合他社や個人の嫌がらせとして見做される場合もある．そういった信頼性の無いレビューに対し，近年は「専門家レビュー」なるものも見かけるようになってきた．専門家レビューとは，レビューと共に書いた個人の氏名，職業，経歴，評価等が明記されているものを指し，専門誌が行なっている商品レビューのウェブ版

に当たる．専門家レビューはある程度の信頼性が保証されていて，レビュー 1 つでかなりの力を持っていると考えられるが，絶対数が少ないうえに投稿されている商品も限られているため，ウェブサイトという場に於いては柔軟性に欠ける．さらには，金銭を受け取って好意的なレビューを書いたり書かせたりする「やらせ業者」の特定も相次いでいるため，オンラインレビューサイトに対する不信感は益々強まるばかりである．従って，オンラインレビューサイトにおける異常検出は重要な研究課題と言える．本論文では，Swan と Allan [2] や Kleinberg [3] と同様に，回顧的 (Retrospective) な立場で異常を検出する新たな手法を提案する．提案法は，ユーザーの基本評点行動として多項分布モデルを仮定し，尤度比検定により異常期間を検出することを特徴とする．

本論文の構成は以下となる．まず，評点時系列データから異常期間を検出する提案法について説明する．次に，実験で用いたデータセットの詳細を述べると共に，実験結果を報告する．最後に，本研究のまとめについて述べる．

2. 提 案 法

評点の時系列データを以下とする．

$$\mathcal{D} = \{(a_1, t_1), \cdots, (a_N, t_N)\}. \tag{1}$$

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，$a_n \in \{1, \cdots, J\}$ となる．モデル記述の都合上，各評点 a_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．

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

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．

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

この尤度関数に対して最尤推定量を求めれば，以下となる．

$$\hat{p_j} = \frac{\sum_{n=1}^{N} a_{n,j}}{N} \tag{4}$$

一方，区間 $S = [t_u, t_v]$ が異常期間であるとして，通常とは違う多項分布に従うと考える．そこでは評点 j が与えられる確率を q_j とし，区間 S 以外では評点 j が与えられる確率を r_j とする．ここで，区間 S に入る評点データを記述するため，以下の集合を導入する．

$$X(S) = \{u, u+1, \cdots, v\}. \tag{5}$$

このとき，評点時系列データの対数尤度関数は次式となる．

$$\mathcal{L}(\mathcal{D}; \boldsymbol{p}, \boldsymbol{q}, S) = \sum_{n \in X(S)} \sum_{j=1}^{J} a_{n,j} \log q_j + \sum_{n \notin X(S)} \sum_{j=1}^{J} a_{n,j} \log r_j. \tag{6}$$

この尤度関数に対して最尤推定量を求めれば，以下となる．

$$\hat{q_j} = \frac{\sum_{n \in X(S)} a_{n,j}}{|X(S)|}, \; \hat{r_j} = \frac{\sum_{n \notin X(S)} a_{n,j}}{N - |X(S)|}. \tag{7}$$

本提案法では，最も顕著な異常区間として，尤度比の対数を最大にする区間 S を求める．即ち，次式を最大にする $\hat{S}$ である．

$$\hat{S} = \arg\max_{S} \{\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, S) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{p}})\}. \tag{8}$$

今回の実験では，区間 S の候補として，評点時系列データの任意の観測時間のペア $[t_u, t_v]$ を考える．ここで，

$$A = \max\{\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, S) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{p}})\}, \tag{9}$$

とし，この A を区間 $\hat{S}$ における異常の度合いとして扱う．一般に，データ数 N が十分に大きいとき，A の 2 倍は漸近的に χ^2 分布となることが知られている．

3. データセット

1 つ目のデータセットは，@cosme(http://www.cosme.net) の商品レビューデータである．@cosme は，日本最大級の化粧品のレビューサイトであり，ユーザーのレビューを中心に、化粧品の情報提供、オリジナル商品の企画などを行っていて，利用者は 20 代の女性が主である．このデータは，2008 年 12 月から 2009 年 12 月にかけてサイトから取得したものであり，48548 個の商品，331084 個のレビュー，45024 人のユーザー，7139 種類のブランドを有する．各レビューに含まれる属性情報は，ユーザー，商品，ブランド，得点，投稿時間である．ここで得点とは，@cosme における評価指標を意味するものであり，1 から 7 の整数値をとりうる．

2 つ目のデータセットは，価格.com(http://kakaku.com) の商品レビューデータである．価格.com は，あらゆるジャンルの商品やサービスを網羅したネットショッピング支援サイトとして知られている．我々は 2011 年 10 月 20 日から同年 10 月 26 日にかけて価格.com のサイト内をクロールし，登録されている商品に付けられたレビューを取得した．このサイトは，全ての商品にレビューを付けられるわけではないため，ユーザーによる評価行動が活発に行われているジャンルには偏りがある．取得データの分析から，価格.com 内のジャンル区分に倣うと，PC 関連 (pc)，家電 (kaden)，カメラ (camera)，車関連 (kuruma)，ゲーム類 (game) といったジャンルでのレビューが大半を占めていることが分かった．このデータセットは，24406 個の商品，164660 個のレビュー，79269 人のユーザーを有する．各レビューに含まれる属性情報は，ユーザー，商品，満足度，投稿時間，参考人数である．ここで満足度とは，価格.com における評価指標を意味する得点であり，1 から 5 までの整数値をとりうる．レビューには，満足度とそれに対応する文章及び画像が添えられているもの（コメント有り）と，満足度のみ（コメント無し）の 2 種類が存在する．今回，レビューの取得時にコメントの有無判別を行なってはいるが，実験時には区別なく全てのレビューを扱うこととした．

4. 実験結果

2 つのデータセットと提案法を用いた実験結果を示す．今回，@cosme のレビューデータはブランドによる区分のもと実験を行い，価格.com のレビューデータは商品による区分のもと実験を行った．実験結果を式 9 で示した A で降順ソートし，上位のブランド及び商品を対象に検証を行った．各実験結果における A の上位 10 ブランド及び商品を表 1, 2 に示す．

表 1　A の上位 10 ブランド（@cosme）

Rank	Brand	A	Reviews
1	ラッシュ	56.61014	8300
2	ザ・ボディショップ	44.76309	3902
3	マキアージュ	39.90004	2925
4	クリニーク	34.14218	2953
5	ボビイ ブラウン	32.33880	1132
6	エリクシール シュペリエル	29.67255	584
7	ルナソル	29.57952	3683
8	オーブ クチュール	27.47581	772
9	ヘレナ ルビンスタイン	26.89848	1385
10	無印良品	26.52386	3694

表 2　A の上位 10 商品（価格.com）

Rank	Item	A	Reviews
1	Wii [ウィー] クロ	48.60031	991
2	MEDIAS N-04C [MEDIAS Black]	42.44330	209
3	docomo PRIME series F-06B	30.23166	267
4	AQUOS SHOT SH006	27.21675	202
5	docomo PRO series SH-04A	23.23475	279
6	biblio	23.03822	186
7	iida G9	21.29133	247
8	WILLCOM 03 WS020SH	21.04830	106
9	LYNX 3D SH-03C [Pure White]	20.05467	168
10	ファイナルファンタジー XIII	19.11675	289

@cosme での実験結果における A の上位ブランドの異常期間 $\hat{S}$ を検証すると，短期間において 1 ユーザーが一定の評価を投稿し続けたことが原因であるものが多かった．例えば，表 1 で 1 位に位置する「ラッシュ」の $\hat{S}$ では，あるユーザーが 2009 年 10 月 25 日 13 時 57 分 56 秒から同日 19 時 30 分 28 秒までに，評点 3 のレビューを 38 回投稿していた．同様に，4 位に位置する「クリニーク」の $\hat{S}$ では，また別のユーザーが 2009 年 11 月 11 日 5 時 17 分 24 秒から同日 11 時 48 分 10 秒までに，最低評価である評点 1 のレビューを 10 回投稿していた．

以上のことより，@cosme の実験では，異常な評点行動をとったユーザーを検出することに成功したと言えるだろう．参考までに，「ラッシュ」，「マキアージュ」，「クリニーク」の $\hat{S}$ における $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{q}}$ の比較を図 1, 2, 3 に示す．

価格.com での実験結果における A の上位商品の $\hat{S}$ を検証すると，極端な評価の変動が起こっていることは分かるが，数値だけでは原因が分からないので，各 $\hat{S}$ で投稿されたレビューの文章を実際に見ていくことにした．因みに，価格.com も含め，一般的なオンラインレビューサイトでは，1 ユーザーが 1 商品に対してレビューを多重投稿することはできないので，こ

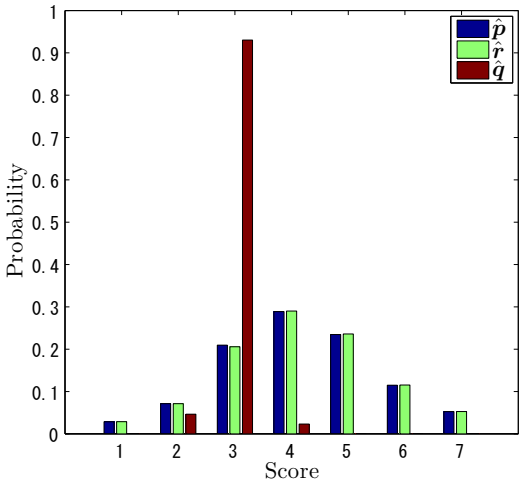

図 1 「ラッシュ」の確率分布比較

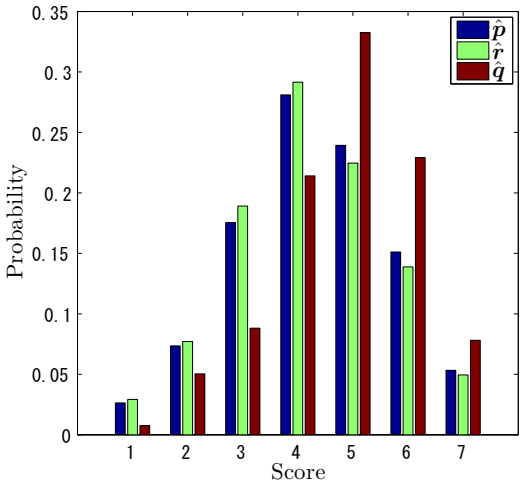

図 2 「マキアージュ」の確率分布比較

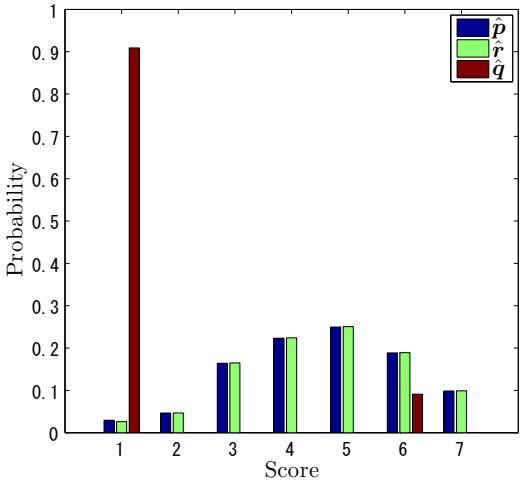

図 3 「クリニーク」の確率分布比較

の実験の区分ではユーザーの異常行動は検出されない．例えば，表 2 で 2 位に位置する「MEDIAS N-04C [MEDIAS Black]」は，2011 年 7 月 12 日 22 時 35 分から 2011 年 10 月 13 日 22 時 2 分にかけて，最低評価である評点 1 のレビューが半分を占めていた．これは，2011 年 3 月 15 日に発売してから不具合による修理や交換が相次いで発生したらしく，購入したユーザーがその時のことを含めてレビューの執筆を行ったためと考えられる．発売当初から $\hat{S}$ まで約 4ヶ月のラグが発生しているあたりから，修理や交換等のほとぼりが冷めたユーザーから一気にレビューが寄せられたことが伺える．一方，$\hat{S}$ での評価が異常に高かった 6 位の「biblio」や 7 位の「iida G9」等は，$\hat{S}$ が発売直後から約半年間であった．半年間に及んで付けられた評価が，それ以降の評価と食い違っていることが原因であることは明確であるが，その原因の詳細については解明できなかった．この 2 つの実験結果は，全区間における明らかな評価の変化点（または変化点を含む短い区間）が 1 箇所存在する場合，提案手法は，変化点以前の区間と以降の区間の 2 区間に分割することにより，分割した時刻でそれらを検出できることを示している．

以上のことより，価格.com の実験結果は，商品の特定時期におけるイベントや，評価の変化点を検出していたことが判明した．参考までに，「MEDIAS N-04C [MEDIAS Black]」，「biblio」，「iida G9」の $\hat{S}$ における $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{q}}$ の比較を図 4, 5, 6 に示す．

これらの実験の本来の目的は，不自然なレビューやユーザーを検出することである．しかし，購入予定の商品や，既に所持している商品の評価に時期的なブレが無いかを確認するために，対象とする商品から異常値を逆引き的に調べられるという有用性も期待できる．言い換えれば，「異常値の低さ」を「平均評点の信頼度」として考えることができるということである．

因みに，両実験において，有意水準 0.05 における自由度 2 の χ^2 の棄却点 10.60 を考えると，A が上位のものは，「$\hat{S}$ は異常区間ではない」という帰無仮説が棄却されることが示唆される．

5. おわりに

日本の代表的な 2 つのオンラインレビューサイトの時系列データを用い，提案法によって，レビューの期間的な異常の検出を試みた．実験結果の検証より，明らかに異常な評点行動をとっているユーザーや，時事的な原因による商品の評価変動，及び評価の変化点を検出することに成功したことを示した．今後は，今回の結果を評点行動モデルの構築に活かすと共に，提案法を応用した平均評点の調整法を検討するつもりである．

謝　辞

本研究は，科学研究費補助基金基盤研究 (C)（No. 23500312）の支援を受けて行ったものである．

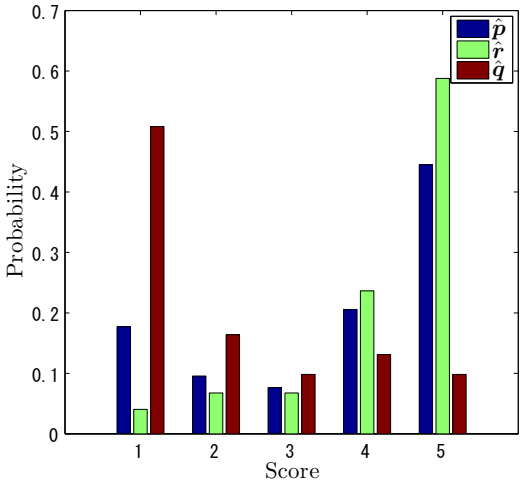

図 4 「MEDIAS N-04C [MEDIAS Black]」の確率分布比較

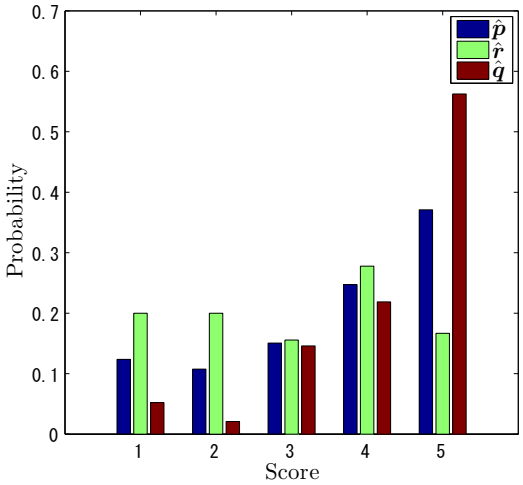

図 5 「biblio」の確率分布比較

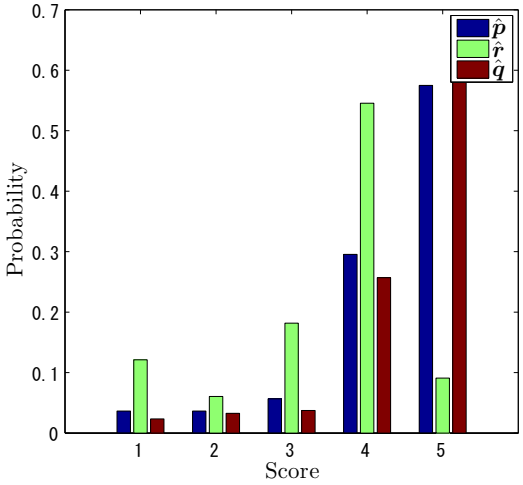

図 6 「iida G9」の確率分布比較

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.