1C4-5

レビューサイトにおける異種情報に基づく変化点検出法

The Change-Point Detection by Heterogeneous Information in Online Review Sites

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

The word of mouth information of online review sites on the internet are affecting various activities from person to person. However, it is not easy to look for useful information from too many reviews. Furthermore, in recent years, some user's unusual evaluation actions that represented by "Fakery" is also regarded as problems. Thus, we propose a methods for change-point detection in time series data of online review sites. We assume evaluation scores, file sizes and review-intervals to be a multinomial distribution model, a exponential distribution model and a gaussian distribution model respectively, and formulate change-point detection problems from the review time-series data. In the experiments, we use a real big data and compare the results of each technique.

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変化の検出を試みる．本研究では，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．実験では現実の大規模データを用い，各手法の結果を比較する．

本稿の構成は以下となる．まず，レビュー時系列データからの変化点検出問題を定式化し，解法を提案する．次に，実験で用いたデータセットの詳細を説明する．最後に，実験結果と本研究のまとめについて述べる．

2. 提案手法

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.

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

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, j12115@u-shizuoka-ken.ac.jp

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．

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

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．

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

$$= (N-1)\log\lambda - \lambda(t_n - t_1). \tag{6}$$

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．

$$\mathcal{L}(\mathcal{D}; \boldsymbol{\theta}_{K+1}, \mathcal{C}_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}; \theta_k). \tag{7}$$

よって，式 (7) 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．

$$\mathcal{LR}(\mathcal{C}_K) = \mathcal{L}(\mathcal{D};\hat{\boldsymbol{\theta}}_{K+1},\mathcal{C}_K) - \mathcal{L}(\mathcal{D};\hat{\boldsymbol{\theta}}_1,\mathcal{C}_0). \qquad (8)$$

本論文では，式 (8) 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．

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．ユーザーのレビューを中心として，化粧品の情報提供，オリジナル商品の企画などが行われており，サイトの利用者は 20 代の女性が主であることが分かっている．

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．各レビューに含まれる属性情報は，ユーザー，ブランド，アイテム，評点，投稿時刻，文章のファイルサイズ，参考票数である．また，1 人のユーザーが 1 つのアイテムに対して口コミを多重投稿することができ，必要ならばその際に評点を変更できる仕様になっている．レビューの評点は，0〜7 の整数値をとりうる．

図 1，2，3 に，レビュー評点の分布，レビューファイルサイズの分布，レビュー投稿間隔の分布をそれぞれ示す．

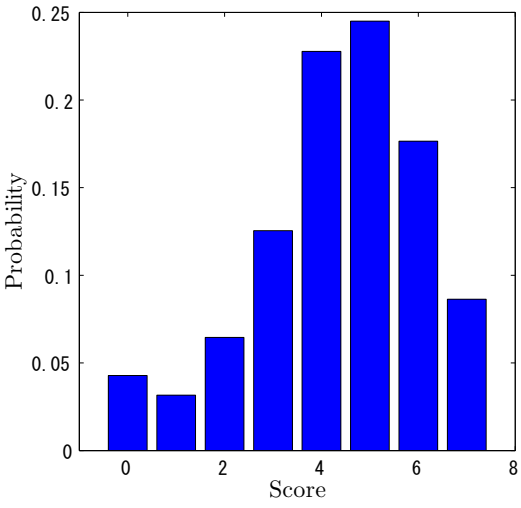

図 1: レビュー評点の分布

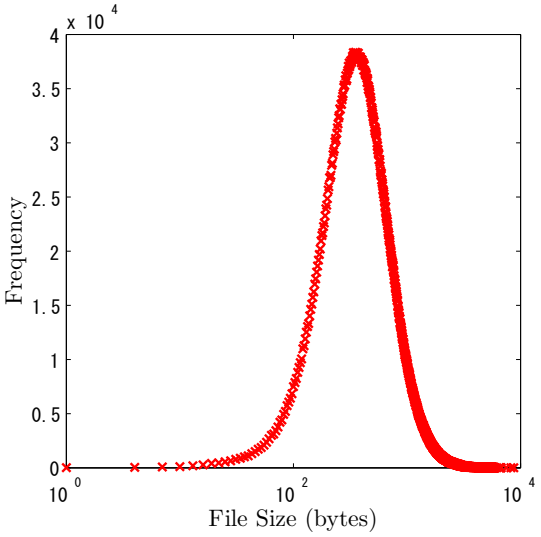

図 2: レビューファイルサイズの分布

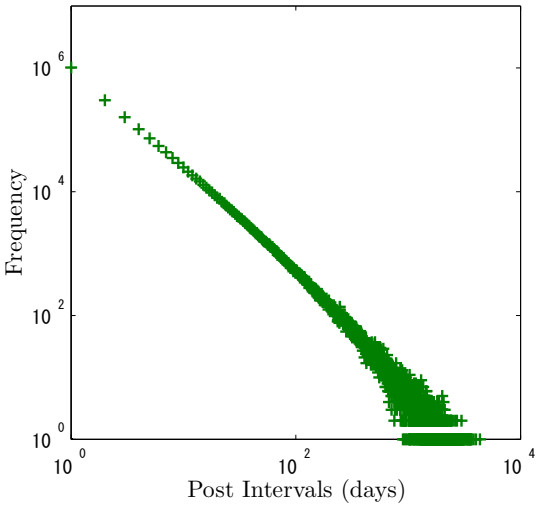

図 3: レビュー投稿間隔の分布

4. 実験結果と考察

提案法を適用するため，データ上の 0〜7 の評点を 1 ずつ加算して 1〜8 とし，ファイルサイズは対数をとって使用した．また，レビューデータはブランド毎に分け，実験対象は総レビュー数が 100 以上の 3558 ブランドとした．さらに，実験結果の比較のしやすさを考慮して，今回変化点数は 1，即ち $\mathcal{C}_1$ で固定した．ここで，$\mathcal{LR}^s$，$\mathcal{LR}^v$，$\mathcal{LR}^t$ は，それぞれ式 (8) で定義した，評点，ファイルサイズ，投稿間隔の対数尤度比を示す．

表 1，2，3 に $\mathcal{LR}^{\cdot}$ の上位 15 ブランドとその詳細を，表 4 に $\mathcal{LR}^{\cdot}$ の相関係数を，図 4，5，6 に $\mathcal{LR}^{\cdot}$ のプロットを示す．表と図より，レビューのファイルサイズと投稿間隔は，提案法において関連性が高く，評点は他の 2 項目と関連性が低いことが分かる．

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

表 1: $\mathcal{LR}^s$ の上位 15 ブランド

Brand	$\mathcal{LR}^s$	UNIX time	Reviews
DHC	1.560863e+3	1139902725	161279
キャンメイク	9.848753e+2	1242892943	88256
ランコム	9.505432e+2	1280970646	99022
クリニーク	8.920518e+2	1116829423	79507
ピエヌ	8.715662e+2	1131618570	40945
ビオレ	8.342914e+2	1263469711	66920
ディオール	8.223021e+2	1179492575	81370
ベビーピンク	7.723887e+2	1030683170	15390
シャネル	7.477878e+2	1215361612	103042
ヘレナ ルビンスタイン	7.163254e+2	1306386769	46331
大島椿	5.560947e+2	1090163526	26523
メイベリン ニューヨーク	5.456240e+2	1210485520	96100
メリット	5.423438e+2	1274147220	5520
コスメデコルテ	4.914082e+2	1077426528	60597
ドクターシーラボ	4.754689e+2	1288505453	36054

表 2: $\mathcal{LR}^v$ の上位 15 ブランド

Brand	$\mathcal{LR}^v$	UNIX time	Reviews
ラッシュ	2.950836e+3	1167792643	163709
シャネル	1.455860e+3	1264276214	103042
マジョリカ マジョルカ	1.417639e+3	1154439136	110710
DHC	1.128838e+3	1285090130	161279
資生堂	1.076777e+3	1281008003	45476
ランコム	1.011683e+3	1281333988	99022
キャンメイク	1.003719e+3	1288191847	88256
オルビス	9.658480e+2	1286820328	154658
専科	9.410325e+2	1285988869	19944
ザ・ダイソー	9.064579e+2	1157017950	107288
クリニーク	6.898432e+2	1054298156	79507
ちふれ	6.887544e+2	1287990484	74071
ロクシタン	6.682499e+2	1168849339	56386
ビオレ	6.485380e+2	1277213495	66920
メイベリン ニューヨーク	6.051582e+2	1284295587	96100

表 3: $\mathcal{LR}^t$ の上位 15 ブランド

Brand	$\mathcal{LR}^t$	UNIX time	Reviews
ラッシュ	2.704959e+4	1043588154	163709
ルナソル	1.733585e+4	1107154352	58206
ロクシタン	1.655206e+4	1098849241	56386
オルビス	1.613024e+4	1007435656	154658
無印良品	1.380696e+4	1123462026	73111
DHC	1.281484e+4	1007353399	161279
ケイト	1.255704e+4	1027993504	107438
キャンメイク	1.242230e+4	1250465141	88256
ザ・ダイソー	1.107623e+4	1033607521	107288
クリニーク	1.046435e+4	1124067407	79507
ピエヌ	1.024806e+4	1209044689	40945
ディオール	9.964376e+3	1028701367	81370
テスティモ	9.573530e+3	1225720420	38377
ザ・ボディショップ	9.416698e+3	1067003658	81293
メイベリン ニューヨーク	8.732692e+3	1006989824	96100

表 4: $\mathcal{LR}$ の相関係数

	$\mathcal{LR}^s$	$\mathcal{LR}^v$	$\mathcal{LR}^t$
$\mathcal{LR}^s$	1.0000	0.6769	0.6521
$\mathcal{LR}^v$	0.6769	1.0000	0.7873
$\mathcal{LR}^t$	0.6521	0.7873	1.0000

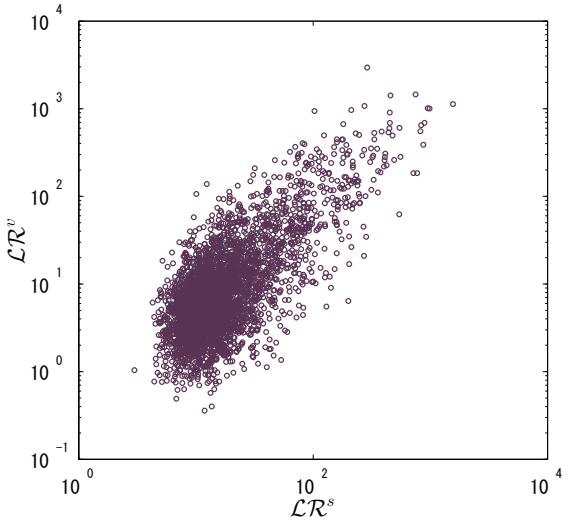

図 4: $\mathcal{LR}^s$ と $\mathcal{LR}^v$ のプロット

5. おわりに

ユーザーの基本行動として，評点を多項分布モデル，投稿間隔を指数分布モデル，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を，尤度比検定の枠組みで試みた．現実の大規模データを用いた実験では，提案法における各項目の関連性を見ることができた．今後は，今回の実験結果の詳細な分析を行うと共に，変化点数を可変とし，それを用いて重要ユーザの検出を行うつもりである．

謝辞

本研究は，科学研究費補助基金基盤研究 (C) (No.2533065) 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 311, 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.

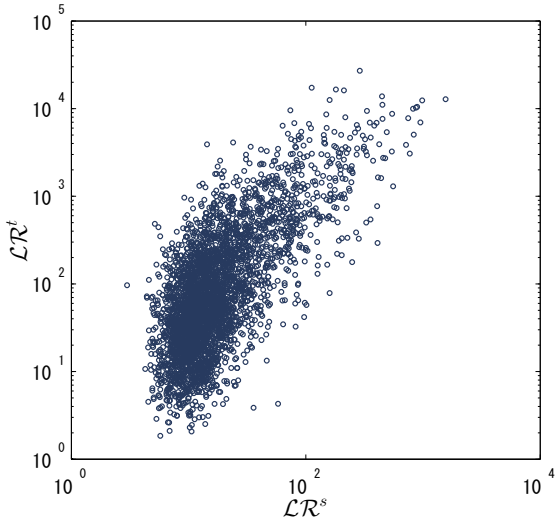

図 5: $\mathcal{LR}^s$ と $\mathcal{LR}^t$ のプロット

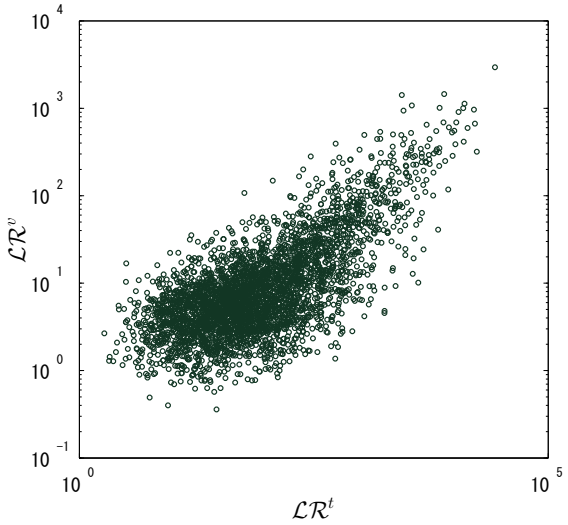

図 6: $\mathcal{LR}^t$ と $\mathcal{LR}^v$ のプロット