5M-5

レビューサイトにおけるユーザ間の動的類似度分析

山岸　祐己†　　斉藤　和巳†　　池田　哲夫†
†静岡県立大学経営情報イノベーション研究科

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．レビューは点数・文章・画像から成ることが多く，レビュー点数の平均点が，アイテムに対する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].

レビューサイトでのユーザによるレビュー行動は，オンライン環境での購買行動を始めとし，ユーザ間で多様な活動に影響を及ぼしている．具体的には，多くのユーザから信頼されているユーザにより，ある商品に対して高評価のレビューが投稿されれば，周囲のユーザ，特に類似度が高いユーザに対し購買意欲を引き起こさせ，結果その商品の販売が大幅に促進されるようなことも起こり得る．よって，レビューサイトにおけるユーザ間の類似度の定式化や，類似度の時系列的変化の観測は，ウェブ情報学において極めて重要といえる．本研究では，ユーザ間の類似度が時間とともに変化する動的な側面を考慮し，現実の大規模レビュー時系列データを用いて統計的分析を行う．

2　分析手法

分析手順と，その要素技術について説明する．ここで，対象データのユーザ集合を V，アイテム集合を $I(|I| = N)$，レビュー点数が取りうる整数値を $M = \{1, \cdots, m\}$，最も多くのレビューを投稿したトップユーザを $v_0 \in V$ とし，以下の手順により分析を行う．

1. v_0 と $v \in V \setminus \{v_0\}$ の動的類似度ベクトル $\boldsymbol{y_v}$ を計算；
2. $\boldsymbol{y_v}$ を K-median 法によりクラスタリング；

2.1　動的レビュー類似度

データ上の最古のレビュー投稿時刻を 0，最新のレビュー投稿時刻を T とすれば，時刻区間 $[s, s+\Delta](0 \le s \le T-\Delta)$ を定めることできる．この時刻区間において，ユーザ $v \in V \setminus \{v_0\}$ が，n 番目のアイテム $i_n \in I$ に対して付与したレビュー得点を $E(v, i_n) \in M$ とし，この区間の v のレビュー得点を要素とする N 次元ベクトルで

$$\boldsymbol{x}_v(s) = \{E(v, i_1), E(v, i_2), \cdots, E(v, i_N)\}^T, \tag{1}$$

と書き表せば，トップユーザ v_0 と v の類似度は

$$y_v(s) = \frac{\boldsymbol{x}_{v_0}(s)^T \boldsymbol{x}_v(s)}{\|\boldsymbol{x}_{v_0}(s)\| \cdot \|\boldsymbol{x}_v(s)\|}, \tag{2}$$

のように算出することができる．さらに，設定した全ての s における $y_v(s)$ を時刻順で要素に持つ動的類似度ベクトルを $\boldsymbol{y_v}$ とすれば，ユーザ u，v 間の動的類似度の類似度は

$$\rho(u, v) = \frac{\boldsymbol{y}_u^T \boldsymbol{y}_v}{\|\boldsymbol{y}_u\| \cdot \|\boldsymbol{y}_v\|}, \tag{3}$$

と求めることができる．

2.2　*K*-median クラスタリング

K-median（K-medoid とも呼ばれる）法は，非階層クラスタリングで有名な K-means 法と同様に，N 個のオブジェクト集合 $\mathcal{V}$ が与えられたとき，オブジェクト集合を K 個のクラスタに分割する手法である．任意のオブジェクトペア $u, v \in \mathcal{V}$ 間に，2.1 節でも用いた類似度 $\rho(u, v)$ を定義し，オブジェクト集合の中から他のオブジェクトとの類似度の和が高い代表オブジェクトを選定し，類似度の高いオブジェクトペアは同じクラスタに，類似度の低いオブジェクトペアは異なるクラスタに属するように分割する．一般的に，平均（mean）より中央値（median）の方が頑健であることが知られている．K-median の解法には反復法や貪欲法があるが，本研究では解の一意性が保証される貪欲法を採用する．さらに，貪欲解法の目的関数のサブモジュラ性より，厳密解ではないものの，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [2]．貪欲法とは，既に選定した代表オブジェクトを固定し，ある評価関数値を最大にするオブジェクトを求め，目的関数が増加するならば代表オブジェクト集合に追加することで，結果の代表オブジェクト集合を求める方法である．各オブジェクトは，最も類似度の高い代表オブジェクトと同じクラスタに割り当てられる．既に選定した代表オブジェクト集合を $\mathcal{P}$ とし，新たに追加を試みるオブ

Dynamic Similarity Analysis Between Users in Online Review Sites
†Yuki YAMAGISHI　†Kazumi SAITO　†Tetsuo IKEDA
†Graduate School of Management and Information of Innovation, University of Shizuoka

ジェクトを w とするとき，本稿では，以下の目的関数を考える．

$$f(\mathcal{P} \cup \{w\}) = \sum_{v \in \mathcal{V}} \max\{\mu(v;\mathcal{P}), \rho(v,w)\}. \quad (4)$$

ここで，$\mu(v;\mathcal{P})$ は既に選定された代表オブジェクトとの類似度の最大値を表し，$\mu(v;\mathcal{P}) = \max_{w \in \mathcal{P}}\{\rho(v,w)\}$ で定義される．以下に貪欲法による K-median 法のアルゴリズムを説明する．

1. $k \leftarrow 1, \mathcal{P}_0 \leftarrow \emptyset$，各オブジェクト $v \in \mathcal{V}$ に対し，$\mu(v;\emptyset) \leftarrow 0$ と初期化する；
2. 式 4 で $\hat{p}_k = \arg\max_{w \in V \setminus \mathcal{P}_{k-1}}\{f(\mathcal{P}_{k-1} \cup \{w\})\}$ を求め，$\mathcal{P}_k \leftarrow \mathcal{P}_{k-1} \cup \{\hat{p}_k\}$ とする；
3. $k = K$ ならば $\hat{\mathcal{P}}_K = \{\hat{p}_1, \ldots, \hat{p}_K\}$ を出力し終了する；
4. 各オブジェクト $v \in \mathcal{V}$ に対し，$\mu(v;\mathcal{P}_k)$ を求め，$k \leftarrow k+1$ としステップ 2. へ戻る．

各オブジェクトを，最も類似度の高い代表オブジェクト $p_k \in \mathcal{P}$ のクラスタ C_k に割り当てる．

3 @cosme データセット

今回使用したデータセットは，@cosme*のレビューデータである．@cosme は，株式会社アイスタイル§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，2012 年 11 月に@cosme をクロールして取得したものであり，131238 アイテム，480182 ユーザー，6606951 レビュー，20307 ブランドを有する．

4 結果とまとめ

今回は，プロット図の見易さを考慮して，レビュー回数が 300 以上の 1190 ユーザを分析対象とし，s を日単位に離散化させ，Δ =365 日，K = 15 とした．図 1 と図 2 に C_3 と C_7 のプロット図を示す．赤い太線は代表オブジェクト $p_k \in \mathcal{P}$ である．

C_3 に属するユーザは，近年になるにつれてレビューの頻度が少なくなった先行活発型，いわば古参である可能性が高いことがわかる．実際，このクラスタのユーザの平均年齢は高い．一方，C_7 に属するユーザは，ある期間だけレビューが活発になったイベント型と言えよう．先ほどと逆で，このクラスタのユーザの平均年齢は低い．

*http://www.cosme.net
§http://www.istyle.co.jp/

以上のことから，今回のユーザクラスタリングは，年齢等のユーザ属性と関連性が見出だせることがわかった．

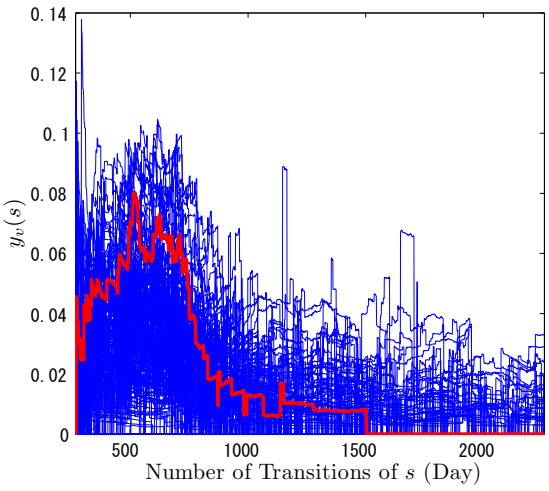

図 1: C_3のプロット

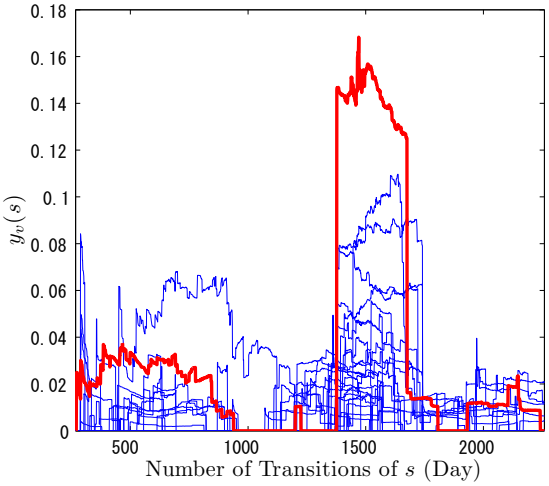

図 2: C_7のプロット

謝辞 本研究は，科学研究費補助金基盤研究 (C)(No.23500312) の補助を受けた．

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*, Vol. 311, No. 5762, pp. 854–856, 2006.

[2] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, Vol. 14, pp. 265–294, 1978.