

ソーシャルメディアにおける情報の伝搬がもたらす商品の売れ行き に対する影響について

奥田 輔[†]

安田 孝美[†]

水野 政司[‡]

[†] 名古屋大学大学院 情報科学研究科 社会システム情報学専攻

[‡] 株式会社クエリーアイ

1 はじめに

近年、ソーシャルメディアの普及に伴いユーザが情報の受け手だけでなく送り手にもなることができるようになった。それに伴い、いわゆる「口コミ」がソーシャル上でも行われるようになった。

本研究では、その効果を実証するために、ソーシャルメディアの例として Twitter を、商品の題材として iTunes App Store で販売されているモバイル向けアプリを対象とし、統計処理を行った。

2 Twitter のつぶやきと iTunes App Store の関連

Twitter において、140 文字という制限上、多量の情報を伝えるためにハイパーリンクを用いることがしばしば見られる。本研究ではまずこれに注目し、S. Brin らの PageRank[1] の考え方に即して、特定のアプリへの被リンクが増加する要因、つまりつぶやきの中にアプリへのハイパーリンクが存在しているものを、そのアプリに関するつぶやきとして収集した。

3 変化点検出エンジン ChangeFinder

非定常データの変動を検出するため、山西らが開発 [2, 3] した変化点検出エンジン ChangeFinder を用いた。

3.1 AR モデル

ChangeFinder では、典型的な時系列モデルの一つである AR モデルを基本としている。まず、初期値の平均が 0 であるような定常時系列変数 $\{z_t : t = 1, 2, \dots\}$ を考える。それぞれのデータ z_t は、 d 次元のベクトルであるとする。この時系列変数 z_t が、 k 次の AR モデルに従って発生すると仮定すると、式 1 のように表される。

$$z_t = \sum_{i=1}^k \omega_i z_{t-i} + \epsilon \quad (1)$$

ここで、 $\omega_i \in \mathbf{R}^{d \times d}$ は d 次パラメータ行列、 ϵ はガウス分布 $N(0, \Sigma)$ に従うノイズ項である。実際に観測される時系列データを $\{\mathbf{x}_t : t = 1, 2, \dots\}$ とすると、その平均 μ を用いて式 2 のように表される。

$$\mathbf{x}_t = z_t + \mu \quad (2)$$

$\mathbf{x}_{t-k}^{t-1} = (\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-k})^T$ とするとき、 \mathbf{x}_t の確率密度関数は式 3 で与えられる。

$$p(\mathbf{x}_t | \mathbf{x}_{t-k}^{t-1} : \theta) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{\xi^T \Sigma^{-1} \xi}{2}\right) \quad (3)$$

ただし、 $\xi = \mathbf{x}_t - \sum_{i=1}^k \omega_i (\mathbf{x}_{t-i} - \mu) + \mu$ であり、パラメータをまとめて $\theta = (\omega_1, \dots, \omega_k, \mu, \Sigma)$ と記した。

3.2 SDAR アルゴリズム

一般に、AR モデルは情報源が定常であることが仮定されており、また新しいデータごとに以前のデータからのバッチ学習を行うため計算量が多くなる。一方、本研究で取り扱う Twitter のつぶやき数や iTunes App Store のランキングなどは、平均や分散が変動する非定常なデータであると考えられる。

そこで、新しいデータと以前のパラメータとの重みつき平均でパラメータ更新する忘却型逐次学習を行うことで、非定常のデータに対応し、なおかつ計算量の削減を行なった。Algorithm 1 に本研究で採用したオンライン忘却型学習 SDAR (Sequentially Discounting AR model learning) のアルゴリズムを示す。

Algorithm 1 SDAR Algorithm

Given:

$\{\mathbf{x}_t | t = 1, 2, \dots\}$ {time series data}
 $0 < r < 1$ {discounting parameter}
 k {order of AR Model}

Initialization:

Set $\hat{\mu}, C_j, \hat{\omega}_j (j = 1, \dots, k), \hat{\Sigma}$

Update Parameter:

for all time series data \mathbf{x}_t and its index t **do**

$\hat{\mu} := (1 - r)\hat{\mu} + r\mathbf{x}_t$

$C_j := (1 - r)C_j + r(\mathbf{x}_t - \hat{\mu})(\mathbf{x}_{t-j} - \hat{\mu})^T$

Solve the Yule-Walker equation for ω_j :

$$\sum_{i=1}^k \omega_i C_{j-i} = C_j (j = 1, \dots, k)$$

$\hat{\omega}_i := \omega_i$

$\mathbf{x}_t := \sum_{i=1}^k \hat{\omega}_i (\mathbf{x}_{t-i} - \hat{\mu}) + \hat{\mu}$

$\hat{\Sigma} := (1 - r)\hat{\Sigma} + r(\mathbf{x}_t - \hat{\mu})(\mathbf{x}_t - \hat{\mu})^T$

end for

3.3 学習

まず, 3.2 で示した SDAR アルゴリズムを用いて学習し, 確率密度関数の列 $\{p_t(\mathbf{x}_t|\mathbf{x}_{t-k}^{t-1} : \theta) : t = 1, 2, \dots\}$ を得る. ここで θ は $(\mathbf{x}_1, \dots, \mathbf{x}_{t-1})$ より学習されたパラメータ群 $(\hat{\omega}_1, \dots, \hat{\omega}_k, \hat{\mu}, \hat{\Sigma})$ である. 次に, \mathbf{x}_t の外れ値スコアを対数損失により, 式 4 で定める.

$$Score(\mathbf{x}_t) = -\log p_{t-1}(\mathbf{x}_t) \quad (4)$$

次に, ウィンドウ幅 T を与え, $Score(\mathbf{x}_t)$ の外れ値スコアを式 5 による移動平均で平滑化し, 新たな時系列データ $\{y_t : t = 1, 2, \dots\}$ を構成する.

$$y_t = \frac{1}{T} \sum_{i=t-T+1}^t Score(\mathbf{x}_i) \quad (5)$$

得られた時系列データ $\{y_t : t = 1, 2, \dots\}$ に対して, 再び SDAR アルゴリズムによって学習する. これにより得られた確率密度関数の列を $\{q_t(y_t|y_{t-k}^{t-1} : \theta') : t = 1, 2, \dots\}$ とする. ウィンドウ幅 T' を与え, 式 6 より移動平均を求め, その値を変化点検出のスコアとする.

$$Score(t) = \frac{1}{T'} \sum_{i=t-T'+1}^t (-\log q_{i-1}(y_i)) \quad (6)$$

ここで得られた $Score(t)$ の値が高いほど, 時刻 t における変化の度合いが高いとみなすことができる.

4 結果と考察

ここでは例として PDF リーダモバイルアプリ, 「GoodReader for iPhone」を取り上げ, 本手法について考察する. Twitter の正規化されたつぶやき数と iTunes App Store 有料総合カテゴリにおけるランキング, また ChangeFinder によって計算されたそれぞれの変化点スコアを図 1 に示す. パラメータはそれぞれ, $k = 4, r = 0.005, T = 10, T' = 10$ とした.

まず, time の 0~500, 700~1000, 1300~1700 の間にそれぞれ Twitter の一連のつぶやき群がある. これらつぶやき群に対して, ChangeFinder は立ち上がり部分に反応し, 高いスコアを算出している. 一方, つぶやき群の後半ではつぶやきが次第になくなっていく様子が上図から見られる. これに対しては, ChangeFinder は反応せず, 緩やかな変動に対しては検知しないことがわかる. 通常 Twitter では, 一部のトレンドに敏感なユーザー群が特定の事象に対して反応し, それに追随する形で周囲のユーザーに拡散していく. その情報拡散は急激に収束をせず, 緩やかに減少していくのが一般的であると考えられる.

このことを考慮すると, 本手法は Twitter の情報拡散のモデルによく当てはまると思われる. 特に図 1 の time の 0~500 周辺つぶやき群に対し, スコアは time index が 100 周辺の初期つぶやき群に対してのみ反応している. これによりつぶやきの立ち上がり部分を得るのに適していると思われる.

また, 図 1 の iTunes の変動を見てみると, 大きな変動はあまりない. このような場合でも, ChangeFinder はそれまでと違った挙動を見せる部分に対して反応する

GoodReader for iPhone (id:306277111)

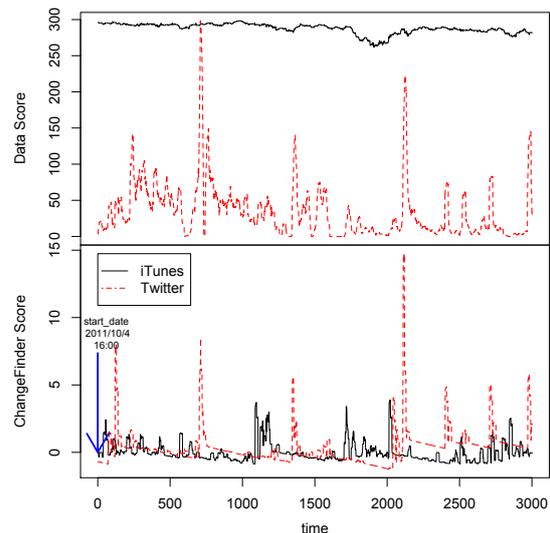


図 1 Twitter つぶやき数と iTunes App Store 有料総合ランキングのデータ, およびその変化点スコア

ことがある. 図 1 中では, 1100~1200 周辺の緩やかな下降に対して, 1700~2000 周辺の上下動に対して反応している.

こういった特徴を持つ ChangeFinder を, Twitter, iTunes App Store 双方に対して組み合わせた結果, Twitter の変化点スコアが高い時間に対して iTunes App Store の変化点スコアが反応する場合, 反応していない場合がみられた.

5 考察と課題

本研究では, オンライン忘却型学習アルゴリズムを用いることで非定常な時系列データである, Twitter のつぶやき数, iTunes App Store のランキングの変化点をよみとり, それぞれがどのように変化するかを調査した.

今後は調査量を増やし, ChangeFinder のパラメータの最適解, また変化点スコアの上昇度によるクラスタリングを行うことで, Twitter のつぶやき数と iTunes App Store のランキングの間の関連性を見つけ出したい.

参考文献

- [1] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine" Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp. 107-117, 1998.
- [2] 山西健司, "データマイニングによる異常検知", 共立出版, 2009
- [3] 竹内純一, 山西健司, "忘却型学習アルゴリズムを用いた外れ値検出と変化点検出の統一的扱い", 2002 情報論的学習理論ワークショップ予稿集 (2002), pp:156-161, 2002