3E3-1

行動履歴からのユーザの活性度の推定

User Enthusiasm Estimation from Behavior History

柳本 豪一 *1
Hidekazu Yanagimoto

*1大阪府立大学
Osaka Prefecture University

In this paper I propose enthusiasm analysis, which estimates both active states and inactive states from observations simultaneously. To realize the enthusiasm analysis I improve burst analysis to detect not only active states but also inactive states. Speaking concretely, I construct a generative model that assumes the observations from Poisson distributions and change state transition costs in the burst analysis. The enthusiasm analysis was applied to customer's service usage histories and predicted customers who would not use the service. From results of the experiments, the enthusiasm analysis can estimate inactive states from the histories and predict the customers who will not use the service.

1. はじめに

観測されたデータからそのデータが生成された構造を予想することができれば、その知見を用いてデータの予測や分類が可能である。例えば、ユーザの行動履歴から、その行動をとった要因である要素を推定することができれば、ユーザの行動予測や分類などに応用することができる。このような考えをもとに、本論文ではユーザの行動履歴である観測データからその行動の取りやすさを表す活性度の推定を行うことを考える。ここでの活性度とは、注目している行動をユーザが頻繁に行うかを数値化してものであり、高いほどその行動を取りやすく、低いほどその行動を行わなくなると想定している。

ユーザの行動ではなく、イベントがどれだけ発生しやすいかという研究としては、Burst 解析 [Kleinberg] がある。バースト解析では、イベント発生の発生はある確率分布に基づいていると仮定し、観測データ列に対してあらかじめ用意した複数の確率分布を割り当てることで、そのイベントの状態を推定するものである。特に、イベントの盛り上がりを検出するための手法である。Burst 解析を用いた研究としては、ニュースにおける話題の盛り上がりの検出 [Takahashi et. al.] やソーシャルメディアの解析への応用 [Yamakawa et. al.] などがある。また、ユーザの行動傾向の予測への応用として、検索クエリーの利用パターンの検出 [Izumi et. al.] にも用いられている。

しかし、Burst 解析では盛り上がりを精度よく検出することはできる反面、盛り下がりに関しては十分な解像度を持っているとは言い難い。このため、ユーザの活性度が下がっているような盛り下がり状態を適切に検出することは難しい。したがって、盛り下がりの状態を検出するために、Burst 解析を修正する必要がある。盛り下がりの検出は重要性の低いタスクではなく、ユーザの行動パターンを知る上では、盛り上がり同様に重要な情報となる。例えば、活性度が下がっているユーザを発見した場合に、そのユーザに対して適切なアプローチすることができ、活性度を引き上げることができれば、ユーザのサービス利用を維持することができ、新規ユーザの獲得と同じ効果があると言える。

連絡先: 柳本 豪一，大阪府立大学，大阪府堺市中区学園町 1 番 1 号，072-254-9279，072-254-9279，
hidekazu@kis.osakafu-u.ac.jp

本論文では、Burst 解析を改良することで、盛り上がりと盛り下がりを同時に検出する Enthusiasm 解析を提案する。そして、Enthusiasm 解析の有効性を確認するため、上記の問題設定に対して応用する。具体的には、サービス利用履歴のデータに対して解析を行い、得られた Enthusiasm レベル (ユーザの活性度) により、以後サービスを利用しなくなる可能性が高いユーザを発見するタスクに取り組む。そして、実験より提案手法が約 78%の正解率でユーザを発見できることを確認した。

2. Burst 解析

2.1 Burst 解析

トピックの盛り上がりを観測値から推定する手法として、Burst 解析がよく用いられる。これは以下のような確率分布モデルを仮定し、観測値に対して適切な確率分布の割り当てを行うことで、バーストレベルの推定を行っている。

まず、イベント発生モデルをイベントが発生する時間間隔 x を用いて、指数分布 $p(x|\alpha_i)$ により定義する。

$$p(x|\alpha_i) = \alpha_i e^{-\alpha_i x} \tag{1}$$

α_i は平均イベント発生時間間隔 α_0 を基準にて計算する。

$$\alpha_0 = \frac{T}{n} \tag{2}$$

$$\alpha_i = \alpha_0 s^i \quad (s > 1) \tag{3}$$

ここで、T は観測期間、n は観測期間中のイベント発生回数を表し、i はバーストレベルを表す。バーストレベルの最大値 $i_{\max}$ は以下より求め、$\delta(x)$ は最小イベント発生時間間隔である。

$$i_{\max} = \lceil 1 + \log_s T + \log_s \delta(x)^{-1} \rceil \tag{4}$$

次に、観測系列に指数分布を割り当てる。この割り当ては対数尤度と遷移コストを利用した評価関数の最小化で実現する。

$$c(i_{1\cdots n}|x_{1\cdots n}) = \sum_{t=1}^{n} -\ln p(x_t|\alpha_{i_t}) + \sum_{t=0}^{n-1} \tau(i_t, i_{t+1}) \tag{5}$$

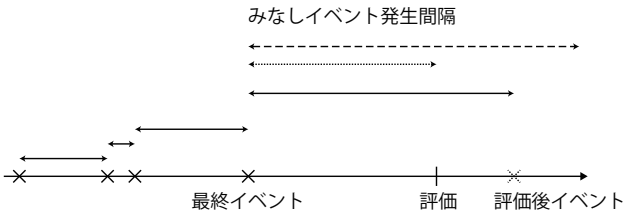

図 1: イベント観測と評価時点との関係

ここで、$\tau(s_i, s_{i+1})$ は隣接するバーストレベルが s_i から s_{i+1} へ推移した時のコストであり、以下のように定義される。

$$\tau(s_i, s_{i+1}) = \begin{cases} 0 & (s_i \geq s_{i+1}) \\ \gamma(s_{i+1} - s_i) \ln n & (s_i < s_{i+1}) \end{cases} \quad (6)$$

式 (5) の最小化は動的計画法により計算される。

2.2 Burst 解析による盛り下がり検出における問題点

Butst 解析では盛り上がりを検出することが目的となっており、そのまま盛り下がりを検出しようとすると、(1) 最終イベント観測から評価時点までの情報が利用できない、(2) 最低のバーストレベルに引き戻されやすく適切なバーストレベルを推定できない、という問題がある。

まず、(1) の最終イベントから評価時点までの情報の利用について検討する。図 1 に評価時点とイベントの関係を示す。評価時点までに観測されているイベントからイベント発生時間間隔を求める場合、最終イベント以降については、(a) 考慮しない、(b) 評価時点で観測があったとみなす、(c) 十分に長い時間が経過後に新しいイベントが発生したとみなす、の 3 通りが考えられる。しかし、(a) を用いると、最終イベント以降にイベントが観測されていない事実を解析に利用することができない。また、(b) や (c) を用いた場合は、真のイベント発生時間間隔を、過小または過大に評価していることとなる。このため、最終イベント発生から評価までの時間が長い場合には、Burst 解析では適切に評価ができない区間が長くなるという問題点がある。

次に、(2) の最低バーストレベルへの引き戻しによるバーストレベルの検出への影響について検討する。式 (6) によるバーストレベルの推移に対してコストが発生する。このため、バーストレベルが小さくなるような遷移が優先されることとなり、最低バーストレベルに引き戻されやすくなる。このため、平均イベント発生時間間隔 α_0 より小さい α_i を用いたとしても、バーストレベルの推移に発生するコストのため、過剰に盛り下がっているように判定される。

3. Enthusiasm 解析

提案手法である Enthusiasm 解析では、イベントの発生をポアソン分布によりモデル化する。つまり、単位時間当たりのイベント発生数 k を用いることとする。

$$p(k|\lambda_i) = \frac{\lambda_i^k}{k!} e^{-\lambda_i} \quad (7)$$

ポアソン分布を用いたモデルのイメージを図 2 に示す。ポアソン分布を用いてモデル化することによって、最終イベントから評価時点の期間も、$k = 0$ として利用できる。λ_i について以下のようにして求める。

次に、状態の遷移に関係したコストについて説明する。

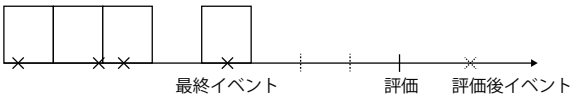

図 2: ポアソン分布を用いたモデル

- λ_0 は平均イベント発生頻度とする。

$$\lambda_0 = \frac{n}{T} \quad (8)$$

- 負の値で表されるレベルはイベントの発生が沈滞していることを表す。最小の Enthusiasm レベルは、イベントの発生が最も盛り下がっている状態を想定して、手動で決定する。
- 他の λ_i については、以下の計算より求める。

$$\lambda_i = \lambda_0 \alpha^i \quad (\alpha > 1) \quad (9)$$

- 最大の Enthusiasm レベルは以下の計算式より求める。ただし、入力データ系列における最大イベント発生回数を k_{max} とする。

$$\lambda_{max} = \left\lceil \frac{\frac{k_{max}}{\lambda_0}}{\alpha} \right\rceil + 1 \quad (10)$$

次に、Enthusiasm レベルが s_i から s_{i+1} へ遷移した時のコストを以下の式で求める。

$$\tau(s_i, s_{i+1}) = \begin{cases} \gamma |s_{i+1} - s_i| \ln N & (s_{i+1} \geq s_i \geq 0, \\ & 0 \geq s_i \geq s_{i+1}) \\ \gamma |s_{i+1}| \ln N & (s_{i+1} \geq 0 \geq s_i, \\ & s_i \geq 0 \geq s_{i+1}) \\ 0 & (\text{otherwise}) \end{cases} \quad (11)$$

この遷移コストは平均イベント発生頻度の状態に回帰するように設定されており、Enthusiasm レベルが負の場合もあるため、上記のように設定されている。

Enthusiasm 解析での評価関数は Burst 解析で用いられている式 (5) と同じものを用い、動的計画法により各観測値に対する最適な Enthusiasm レベルを求める。

3.1 Enthusiasm 解析による離脱可能性ユーザの発見

サービスを受けていたユーザが時間経過とともにサービスを利用しなくなることはよくある。このような顧客はサービス離脱ユーザと呼ばれ、これらのユーザをサービスに引き止めることができれば、新規ユーザを獲得するのと同じ効果がある。そこで、ユーザの過去の行動履歴から、サービスを離脱する可能性が高いユーザを発見することを考える。

まず、サービスから離脱する可能性のあるユーザは、サービスの利用頻度が低くなり、その結果としてサービスから離脱すると仮定する。すべてのサービスから離脱するユーザがこのプロセスを経ず、突然サービスを利用しなくなることも考えられる。ただ、この場合も離脱した時にすぐに検知はできないとしても、サービスを使用しない期間が長くなるにつれて検知できると考えられる。

離脱可能性ユーザを発見する方法としては、上記の仮定に基づき、Enthusiasm レベルが低くなった時に、サービスから離脱する可能性が高い状態であると判断し、離脱可能性ユーザとして検知することとする。

表 1: 実験対象の顧客

実験対象顧客	23,457 人
離脱可能性顧客	5,050 人

表 2: Enthusiasm 解析におけるパラメータの設定

パラメータ	値
α	2
γ	1.0

4. 実験

実験には、顧客のサービス利用実績データを用いた。具体的には、そのデータに対して Enthusiasm 解析を行い、顧客のサービス利用に関する Enthusiasm レベルを推定する。そして、推定された Enthusiasm レベルにより、将来そのその顧客が半年に渡ってサービスを利用しなくなる顧客 (離脱可能性顧客) かどうかを判定し、提案手法の有効性について議論する。

4.1 実験データ

2013 年 1 月 6 日から 2015 年 6 月 30 日にわたって収集された顧客のサービス利用実績のデータを用いた。このうち、2013 年 1 月 6 日から 2014 年 11 月 30 日までのデータを用いて Enthusiasm 解析を行い、2014 年 11 月 30 日における顧客の Enthusiasm レベルを求め、離脱可能性顧客の推定に利用する。2015 年 1 月 1 日以降一度もスーパーを利用していない顧客を離脱可能性顧客とみなす。実験に用いた顧客の総数と離脱可能性顧客数は表 1 である。

4.2 パラメータ設定

顧客の利用履歴データは、顧客ごとに一週間あたりの利用頻度を用いて入力データを作成する。また、最小の Enthusiasm レベルは-2 とし、このときの λ_{-2} の値は約半年間利用していない状態とし、$\frac{1}{30}$ とする。最大の Enthusiasm レベルは式 (10) を用いて、顧客ごとに計算する。他のパラメータについては表 2 に示す。

4.3 実験結果

表 3 に推定された Enthusiasm レベルに該当する顧客数及び離脱可能性顧客数を示す。この結果より、Enthusiasm レベルが負となっている顧客のみに着目すると、約 78.5%の正解率で、全離脱可能性顧客の約 60.9%を検出することができた。しかし、約 40%の離脱可能性顧客は平均的な利用頻度状態であり、本実験では検出できていない。この顧客層に対して離脱可能性を判定するため、顧客の行動パターンのチェック、Enthusiasm 解析のパラメータ調整などを行う必要がある。

5. おわりに

本論文では、盛り上がりだけではなく、盛り下がりも同時に検出するため、Burst 解析を拡張した Enthusiasm 解析を提案した。Enthusiasm 解析では、Burst 解析では不可能であった、(1) 最終イベントと評価までの観測情報の利用、(2) 適切なユーザ状態 (Enthusiasm レベル) の推定、を実現した。また、Enthusiasm 解析の応用例として、顧客のサービス利用履歴を対象とした顧客の活性度状態 (Enthusiasm レベル) の推定を行い、その推定された Enthusiasm レベルから離脱可能性顧客の推定を行った。実験より、平均的なサービス利用実績より利用状況が悪化していると推定された顧客の約 78%が実際にサービスから離脱する可能性が高いことが分かった。また、全離脱可能性顧客の約 60%を検出できることも分かった。

表 3: Enthusiasm レベルを用いた離脱可能性顧客の推定結果

Enthusiasm レベル	該当顧客数	離脱可能性顧客
-2	1,544 人	1,381 人
-1	2,375 人	1,695 人
0	19,256 人	1,971 人
1	264 人	3 人
合計	23,440 人	5,050 人

今後としては、さらに検出精度を向上させるため、Enthusiasm 解析のパラメータ調整を行う必要がある。また、今回検出できなかった顧客を調べることで、Enthusiasm 解析では検出が困難と思われる行動パターンについて検討を行う予定である。

参考文献

[Kleinberg] Kleinberg, J.:Bursty and Hierarchical Structure in Streams Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2002)

[Takahashi et. al.] 高橋 佑介, 横本 大輔, 宇津呂 武仁, 吉岡 真治:ニュースにおけるトピックのバースト特性の分析, 情報処理学会研究報告, Vol.2011-NL-204, No. 6, pp.1-6, (2011)

[Yamakawa et. al.] 山川 義介, 小野 広司:Twitter 解析のための技術と 2013 年ヒット予測, オペレーションズ・リサーチ:経営の科学, Vol. 58, No.8, pp.462-468, (2013)

[Izumi et. al.] Izumi, A. and Yanagimoto. H.: Query analysis for site visits with burst analysis, Proc. of ICKM2015, (2015)