6S-06

アクセスログデータの区間分割に基づくユーザ行動分析

山岸　祐己 †　　斉藤　和巳 †　　湯瀬　裕昭 †　　武藤　伸明 †
† 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

時系列データにおけるバーストや異常区間の検出手法は，観測されたデータの時刻間隔や，データの頻度を利用するものが代表的である [1] [2] [3]．しかし，これらの手法は，単一のトピックについてのバーストを検出するためのものであるため，多様なトピックやその分布の変化については扱っていない．即ち，これらの手法は，時刻間隔や頻度の変化を有しない観測データには適応できず，また，データの内容が著しく変化していたとしてもそれらを検出するには不向きである．これに対し我々は，データ内容の分布変化のみに着目し，多項分布モデルを仮定した尤度比検定の枠組みで，時系列データの区間分割手法を提案した [4]．本論文は，この手法をアクセスログデータに適応し，ユーザの行動分析における有用性を確かめるものである．

2　問題設定

あるユーザの n 番目のアクセスのアクセス先サーバーとタイムスタンプを s_n と t_n とすると，観測されたアクセスログは $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ のように示すことができる．なお，アクセス先サーバーは J 種類 ($s_n \in \{1, \cdots, J\}$) とする．モデル記述の都合上，s_n を J 次元ベクトルに変換する以下のダミー関数を導入する．

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

ここで，サーバー j にアクセスする確率 p_j は多項分布であると仮定すると，パラメータベクトル $\boldsymbol{p} = \{p_1, \cdots, p_J\}$ によって計算される $\mathcal{D}$ の対数尤度は以下となる．

$$\mathcal{L}(\mathcal{D}; \boldsymbol{p}) = \sum_{n=1}^{N} \sum_{j=1}^{J} s_{n,j} \log p_j. \tag{2}$$

式 (2) の最尤推定量は $\hat{p}_j = \sum_{n=1}^{N} s_{n,j}/N$ によって与えられる．

次に，k 番目の分割点を T_k ($t_1 < T_k < t_N$) とし，K 個の分割点集合を $C_K = \{T_1, \cdots, T_K\}$ とする．便宜上，$T_0 = t_1$，$T_{K+1} = t_N$，$T_{k-1} < T_k$ としておく．更に，C_K による $\mathcal{D}$ の分割を $\mathcal{D}_k = \{n; T_{k-1} < t_n \leq T_k\}$，即ち $\mathcal{N} = \{1, 2, \cdots, N\} = \{1\} \cup \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{K+1}$ とし，$|\mathcal{D}_k|$ は $(T_{k-1}, T_k]$ の観測点数とする．なお，$k = 1, \cdots, K+1$ に対して $|\mathcal{D}_k| \neq 0$，即ち少なくとも t_n が存在すること，加えて $t_n \in \mathcal{D}_k$ が満たされるとする．この分割点検出問題は，分割点集合 $C_K \subset \mathcal{T}$ ($\mathcal{T} = \{t_1, t_2, \cdots, t_N\}$) の探索問題と同等である．

分割点集合 C_K が与えられた $\mathcal{D}$ の対数尤度は，区間 k 毎のパラメータベクトル $\boldsymbol{P}_{K+1} = \{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{K+1}\}$ によって以下のように計算される．

$$\mathcal{L}(\mathcal{D}; \boldsymbol{P}_{K+1}, C_K) = \sum_{k=1}^{K+1} \mathcal{L}(\mathcal{D}_k; \boldsymbol{p}_k). \tag{3}$$

式 (3) の最尤推定量は，$k = 1, \cdots, K+1$ と $j = 1, \cdots, J$ に対し，$\hat{p}_{k,j} = \sum_{n \in \mathcal{D}_k} s_{n,j}/|\mathcal{D}_k|$ によって与えられるため，それら推定量に置き換えると以下の式が導かれる．

$$\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) = \sum_{k=1}^{K+1} \sum_{n \in \mathcal{D}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{4}$$

従って，この分割点検出問題は，式 (4) を最大化する分割点集合 C_K の探索問題と言い換えることができる．しかし，式 (4) 単体では C_K の導入による効果を直接的に評価することができないので，この問題を対数尤度比の最大化問題として定式化する．観測されたアクセスログが分割点を有しない，即ち $C_0 = \emptyset$ と仮定すると，式 (4) は $\mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{1,j}$ となり，このときの最尤推定量は $\hat{p}_{1,j} = \sum_{n \in \mathcal{N}} s_{n,j}/|\mathcal{N}|$ となる．故に，K 個の分割点を有する場合と，分割点を有しない場合との対数尤度比は，以下のように与えられる．

$$\mathcal{LR}(C_K) = \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_{K+1}, C_K) - \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{P}}_1, C_0). \tag{5}$$

最終的に，この分割点検出問題は，式 (5) を最大化する C_K の探索問題に帰着する．

3　解法

式 (5) の最大化問題を網羅的に解くと，最適解が得られることが保証されるが，計算量は $O(N^K)$ となるため，実用的な時間で解が得られるのはせいぜい $K \geq 3$ のときである．従って，我々はあらゆる K に対して効率的に解を得られる解法として，貪欲法による解法 (*Simple*)

User Behavior Analysis Based on Stream Dividing of Access Log Data
†Yuki YAMAGISHI　†Kazumi SAITO　†Hiroaki YUZE　†Nobuaki MUTOH
†Graduate School of Management and Information of Innovation, University of Shizuoka

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

と，既に提案した *Sequential* と *Iterative* [4] を実験で用いる．

4 データセット

今回用いたデータセットは，第7回静岡おまちバル*のために設置したおよそ100箇所のFree Wi-Fi Spotのアクセスログデータである．ユーザIDはデバイスのMACアドレスで分類し，アクセス先サーバーIDはアクセス先の問い合わせ回答にオーソリティを持つDNSで分類をしている．データセットの対象時刻は11/08/2014 09:08:35から同日の22:01:09であり，238ユーザ，299アクセス先サーバーを有する．

5 実験結果

今回の実験における各手法の解品質の比較を表1に，計算時間 (Intel(R) Xeon(R) CPU X5690 @ 3.47GHz) の比較を表2にそれぞれ示す．表より，提案法のIterativeは他の2手法よりも計算時間を要しているが，実用的な時間で解くことができており，解品質においては他の2手法よりも優れている．

また，Iterativeの実験結果の代表として，ユーザID 13のアクセスログと分割点を図1に，各区間の主要DNSアクセス確率分布を図2にそれぞれ示す．図より，本手法の分割点は，アクセス分布の変化を基に，アクセスログを有意に分割することができていることが見て取れる．例えば，$\mathcal{D}_1$，$\mathcal{D}_3$，$\mathcal{D}_5$ は，似たようなアクセス分布となっており，特定のドメイン (DNS 14) へのアクセスが多いことから，仕事関連の作業をしていることが推察される．更に，$\mathcal{D}_2$，$\mathcal{D}_6$ は，大手ポータルサイト (DNS 10) へのアクセスが多いことから，ウェブブラウザで検索作業をしていることが分かる．また，$\mathcal{D}_4$，$\mathcal{D}_7$ は，他の区間には出現しないアクセス先サーバーへのアクセスが多いため，デバイスの通信状態が他の区間とは明確に異なっている．

表 1: 解品質の比較

	Sum of $\mathcal{LR}(C_K)$	Rate
Iterative	34805.6837	101.01%
Sequential	34771.0212	100.91%
Simple	34459.0665	100.00%

表 2: 計算時間の比較

	Calculation time	Rate
Iterative	104.99sec	1418.78%
Sequential	24.13sec	326.08%
Simple	7.40sec	100.00%

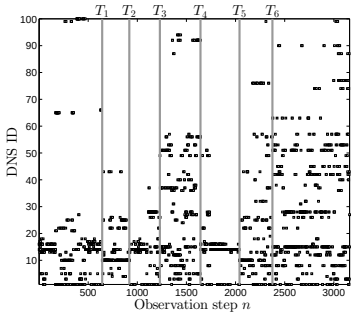

図 1: ユーザ ID 13 のアクセスログと分割点

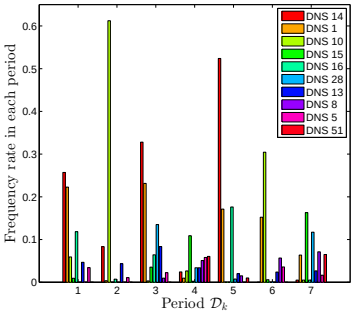

図 2: ユーザ ID 13 の主要 DNS アクセス確率分布

謝辞

本研究は，総務省SCOPE (No.142306004)，ふじのくに地域・大学コンソーシアム学術研究，及び，科学研究費補助金基盤研究 (C)(No.25330635) の補助を受けた．また，アクセスログデータの収集にはタケショウ株式会社の協力を得た．

*http://www.omachibar.com

参考文献

[1] J. Kleinberg, “Bursty and hierarchical structure in streams,” in *Proc. of KDD 2002*, pp. 91–101, 2002.

[2] Y. Zhu, and D. Shasha, “Efficient elastic burst detection in data streams,” in *Proc. of KDD 2003*, pp. 336–345, 2003.

[3] A. Sun, D. Zeng, and H. Chen, “Burst detection from multiple data streams: A network-based approach,” *IEEE Transactions on Systems, Man, & Cybernetics Society, Part C*, pp. 258–267, 2010.

[4] Y. Yamagishi, S. Okubo, K. Saito, K. Ohara, M. Kimura, and H. Motoda, “A Method to Divide Stream Data of Scores over Review Sites,” in *Proc. of PRICAI 2014*, pp. 913–919, 2014.

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.