L-016

# インターネットトラヒックのポアソン分布の密度パラメータが時間変動する時系列モデルを用いた解析に関する一考察

小泉 大城* 松嶋 敏泰† 平澤 茂一†

Daiki KOIZUMI Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

## 1 まえがき

インターネットのネットワーク通信におけるトラヒック解析の研究のうち，確率モデルを仮定するものについて，どのような確率分布がよくトラヒックを表わすかという点についてはその種類によって様々な実験的評価がある [2]．たとえば，Telnet や FTP セッションのようなユーザーによるリクエストは電話網のトラヒック解析に使われていたような定常ポアソン分布で比較的良好に表現されることが指摘されている [5] が，これ以外のプロトコル，あるいはコネクション内のパケット到着間隔などにはポアソン分布以外の分布のほうが適合するとする解析例が報告されている [5][2]．

特に http プロトコルのトラヒックに関する研究については，名部ら [4] が Proxy の TCP パケットのうち http プロトコルのものに着目し，ドキュメントサイズの分布が対数正規分布で良好に表されることを実データにより明らかにしている．[4] ではさらに待ち行列モデルのシミュレーションによる遅延特性の評価を行っているが，到着間隔に (定常) ポアソン分布，サービス時間に対数正規分布を仮定している．一方で篠ら [3] は，http セッションのリクエストが密度パラメータ $\lambda_t$ の時間的に変化するようなポアソン分布に従うことを検定によって示したうえで，$\lambda_t$ を推定して http のトラヒックの解析を行っている．しかし，密度パラメータをフーリエ級数展開し，さらに最ゆう推定の際に遺伝的アルゴリズムを用いており，近似的な推定となっている．

これに対して本研究では，岩田ら [1] によって提案されたポアソン分布の密度パラメータ $\lambda_t$ が時間変化する時系列解析モデルを利用して，http セッションのリクエスト解析を試みる．このモデルでは，現在の $\lambda_t$ に依存して次の $\lambda_{t+1}$ が決まり，その変動の大きさをパラメータ $\rho$ を使って表現している．このとき，岩田らのモデルには密度パラメータ $\lambda_t$ の事後分布が解析的に求まり，さらに $\rho$ も比較的簡単な数値計算で最ゆう推定が可能であるという特徴がある．さらに予測値と観測値とで計算される二乗誤差の期待値が最小になるという理論的保証も得られる．本研究ではこのモデルを使って実際の http 通信リクエストからなるトラヒックを解析し，定常ポアソン分布を仮定したモデルよりも精密な解析が行える解析例を示すものである．

*早稲田大学大学院理工学研究科, 169-8555 東京都新宿区大久保 3-4-1, Email: dkoizumi@matsu.mgmt.waseda.ac.jp

†早稲田大学理工学部経営システム工学科

## 2 インターネットトラヒックのモデル化

### 2.1 非定常ポアソン分布によるトラヒックの表現

本研究で対象とするインターネットトラヒックは，篠ら [3] にならい，ユーザーが発生させる http プロトコルの通信リクエストとする．この種のトラヒックを独立な点過程の重ね合わせとみなすと，その極限分布はポアソン分布となる．いま，離散時刻 $t = 1, 2, \cdots$ のリクエストの到着数 $x_t$ に以下で定義される非定常なポアソン分布を仮定する：

$$p(x_t|\lambda_t) = \frac{e^{-\lambda_t}}{x_t!}(\lambda_t)^{x_t}. \quad (1)$$

ただし，$\lambda_t$ は $\lambda_t \geq 0$ を満たしながら時間変化する密度パラメータである．

### 2.2 篠ら [3] による密度パラメータの定式化

篠ら [3] は，この $\lambda_t$ の変化に周期 $T$ を仮定し，以下のようなフーリエ級数展開をしている：

$$\lambda_t = a_0 + \sum_{n=1}^{\infty}\{a_n \cos(\frac{2\pi}{T}nt) + b_n \sin(\frac{2\pi}{T}nt)\}, \quad (2)$$

$$a_0 = \frac{1}{T}\int_0^T \lambda_t dt, \quad (3)$$

$$a_n = \frac{2}{T}\int_0^T \lambda_t \cos(\frac{2\pi}{T}nt)dt, \quad (4)$$

$$b_n = \frac{2}{T}\int_0^T \lambda_t \sin(\frac{2\pi}{T}nt)dt, \quad (5)$$

(2) の項の個数 $n$ をいくつ取るかは AIC (Akaike Information Criterion) によってモデル選択し，係数 $(a_0, a_n$

および $b_n$) のうち，非負のものを遺伝的アルゴリズムを用いて最ゆう推定しているが，これは近似推定となっている．

### 2.3 岩田ら [1] による密度パラメータの定式化

岩田ら [1] はポアソン分布の密度パラメータ $\lambda_t$ が以下のように変動する (非定常な) モデルを提案している：

$$\lambda_{t+1} = \frac{U_t}{\rho}\lambda_t \ , \ (t = 1, 2, \cdots). \tag{6}$$

ただし，$\rho$ は $0 < \rho \leq 1$ を満たす定数であり，$t = 1$ における $\lambda_1$ の事前分布にはパラメータ $\alpha_1, \beta_1$ のガンマ分布 $Ga(\alpha_1, \beta_1)$ を仮定する．さらに，$U_t$ はパラメータ $\rho\alpha_t$ および $(1-\rho)\alpha_t$ のベータ分布 $Be(\rho\alpha_t, (1-\rho)\alpha_t)$ に従うものとする．

## 3 岩田らのモデルの性質

### 3.1 $\lambda_t$ の変化を決めるパラメータ $\rho$ について

(6) において，$\rho = 1$ のときはベータ乱数 $U_t$ の分散がゼロとなって $\lambda_t$ は時間的に変化せず，定常ポアソン分布を仮定していることに対応する [1]．$0 < \rho < 1$ のときは $\lambda_t$ は時間的に変化する (=非定常である) が，ベータ乱数 $U_t$ の分散は，$\rho$ の値が小さくなるにつれて大きくなり，$\rho = 0.5$ で最大となる．したがって $\rho$ は $\lambda_t$ の変化の激しさを表わすパラメータになっている [1]．

### 3.2 密度パラメータ $\lambda_t$ の事後分布

(6) のように変化する $\lambda_t$ に対し，$(t-1)$ 期までの観測値の系列 $x_1^{t-1} = x_1 x_2 \cdots x_{t-1}$ が得られたもとでの $\lambda_t$ の分布を $Ga(\alpha_{t-1}, \beta_{t-1})$ とする．ガンマ分布とベータ分布の性質を使うと，新たに $x_t$ を観測したもとでの $\lambda_{t+1}$ の事後分布 $f(\lambda_{t+1}|x_1^t)$ は以下のようになる：

$$\begin{aligned}
f(\lambda_t|x_1^{t-1}) &= Ga\left(\alpha_{t-1}, \beta_{t-1}\right), & (7)\\
f(\lambda_t|x_1^t) &= Ga\left(\alpha_{t-1} + x_t, \beta_{t-1} + 1\right), & (8)\\
f(\lambda_{t+1}|x_1^t) &= Ga\left(\rho(\alpha_{t-1} + x_t), \rho(\beta_{t-1} + 1)\right), & (9)\\
&= Ga\left(\alpha_t, \beta_t\right). & (10)
\end{aligned}$$

### 3.3 $\rho$ が未知の場合の最ゆう推定

岩田のモデルでは，$\rho$ が既知という条件の下で $\lambda_t$ の事後分布が解析的に計算できるが，実データを扱う際には事後分布の更新にパラメータ $\rho$ を推定することが必要である．本研究では $\rho$ の最ゆう推定を行うものとすると，ゆう度関数は以下のようになる：

$$\begin{aligned}
&L(\rho)\\
&= \prod_{k=1}^{t-1} p(x_{k+1}|x_1^k, \rho) p(x_1|\lambda_1)\\
&= \prod_{k=1}^{t-1} \int_0^\infty p(x_{k+1}|x_1^k, \lambda_{k+1}) f(\lambda_{k+1}|x_1^k) d\lambda_{k+1}\\
&= \prod_{k=1}^{t-1} \frac{(\beta_k)^{\alpha_k}\Gamma(\alpha_k + x_{k+1})}{(\beta_k + 1)^{\alpha_k + x_{k+1}}\Gamma(\alpha_k) x_{k+1}!}. \qquad (11)
\end{aligned}$$

ここで，(7)-(10) より，

$$\begin{aligned}
\alpha_k &= \rho^{k-1}\alpha_1 + \sum_{i=1}^{k} \rho^{k-i} x_i \ , & (12)\\
\beta_k &= \rho^{k-1}\beta_1 + \sum_{i}^{k-1} \rho^i \ , & (13)
\end{aligned}$$

が成り立つので，

$$\begin{aligned}
&L(\rho)\\
&= \prod_{k=1}^{t-1} \frac{\left(\rho^{k-1}\beta_1 + \sum_{i=1}^{k-1}\rho^i\right)^{\left(\rho^{k-1}\alpha_1 + \sum_{i=1}^{k}\rho^{k-i}x_i\right)}}{\left(\rho^{k-1}\beta_1 + \sum_{i=1}^{k-1}\rho^i + 1\right)^{\left(\rho^{k-1}\alpha_1 + x_{k+1} + \sum_{i=1}^{k}\rho^{k-i}x_i\right)}}\\
&\quad \times \frac{\Gamma(\rho^{k-1}\alpha_1 + x_{k+1} + \sum_{i=1}^{k}\rho^{k-i}x_i)}{\Gamma(\rho^{k-1}\alpha_1 + \sum_{i=1}^{k}\rho^{k-i}x_i)} \times \frac{1}{x_{k+1}!}. \qquad (14)
\end{aligned}$$

$$\begin{aligned}
&\log L(\rho)\\
&= \sum_{k=1}^{t-1}\left(\rho^{k-1}\alpha_1 + \sum_{i=1}^{k}\rho^{k-i}x_i\right)\log\left(\rho^{k-1}\beta_1 + \sum_{i=1}^{k-1}\rho^i\right)\\
&- \sum_{k=1}^{t-1}\left\{\left(\rho^{k-1}\alpha_1 + x_{k+1} + \sum_{i=1}^{k}\rho^{k-i}x_i\right)\right.\\
&\quad \left.\times \log\left(\rho^{k-1}\beta_1 + \sum_{i=1}^{k-1}\rho^i + 1\right)\right\}\\
&+ \sum_{k=1}^{t-1}\log\left\{\Gamma(\rho^{k-1}\alpha_1 + x_{k+1} + \sum_{i=1}^{k}\rho^{k-i}x_i)\right\}\\
&- \sum_{k=1}^{t-1}\log\left\{\Gamma(\rho^{k-1}\alpha_1 + \sum_{i=1}^{k}\rho^{k-i}x_i)\right\}\\
&- \sum_{k=1}^{t-1}\log(x_{k+1}!). \qquad (15)
\end{aligned}$$

この対数ゆう度関数は解析的には解けないものの，次章の予備実験で数値的に示すように，関数の形状が単峰となるため単純な計算アルゴリズムで最ゆうパラメータ $\hat{\rho}$ の推定が可能となる．

### 3.4 観測値を得たもとでのベイズ基準による最適な予測

いま，前章のモデルに対し観測値の系列 $x_1^t = x_1 x_2 \cdots x_t$ を得たもとで $x_{t+1}$ を予測する問題を考える．$\hat{x}_{t+1}$ を $x_{t+1}$ の予測値として，観測値と予測値との間に以下のような二乗損失を仮定する：

$$L(x_{t+1}, \hat{x}_{t+1}) = (x_{t+1} - \hat{x}_{t+1})^2. \qquad (16)$$

このとき，(16) で定義された損失の期待値を最小にするという意味でベイズ最適な予測を特に $\hat{x}^*_{t+1}$ とすると，$\hat{x}^*_{t+1}$ は以下のように簡便な算術計算にて解析的に求めることが可能である：

$$\begin{aligned}
&\hat{x}^*_{t+1} \\
&= \sum_{x_{t+1}} x_{t+1} \int_0^\infty p(x_{t+1}|\lambda_{t+1}, x_1^t) f(\lambda_{t+1}|x_1^t) d\lambda_{t+1} \\
&= E[x_{t+1}|x_1^t] \\
&= \frac{\sum_{k=1}^t \rho^{t-k} x_k + \rho^t \alpha_1}{\sum_{k=1}^t \rho^{t-k} + \rho^t \beta_1}. \qquad (17)
\end{aligned}$$

## 4 実際のトラヒックデータに対する応用

### 4.1 http リクエストのデータ諸元

早稲田大学内の 2 種類の Web サーバ (A および B) の http リクエストのデータから，同一の接続元のうち，初めの接続から 2 秒以内の接続は同一とみなしたうえ [3]，実接続数をカウントし，24 時間を 5 分毎の計 288 離散時間に分割して集計してから解析を行った．

表 1: 解析に用いた http リクエストのデータ諸元

| | サーバ A | サーバ B |
|---|---|---|
| 年月日 | 2004 年 8 月 6 日 | 2004 年 8 月 6 日 |
| 時間帯 | 0:00 - 24:00 | 0:00 - 24:00 |
| 総接続数 | 16395 | 64056 |
| 実接続数 | 442 | 37822 |

### 4.2 予備実験

サーバ A, B の http リクエストのデータを用いて，$\lambda_t$ の事前分布であるガンマ分布のパラメータを $\alpha_1 = \beta_1 = 3.0$ に定め，(15) の対数ゆう度関数 $\log L(\rho)$ を数値計算にてプロットしたところ，どちらも単峰となった．縦軸にサーバ B のデータによる対数ゆう度関数 $\log L(\rho_B)$ の値，横軸に $\rho_B, (0 \le \rho_B \le 1)$ を取ったときの対数ゆう度関数のプロットを図 1 に示す．このときの $\rho$ の最ゆう推定値 ($\hat{\rho}$) はそれぞれ表 2 に示すようになった．

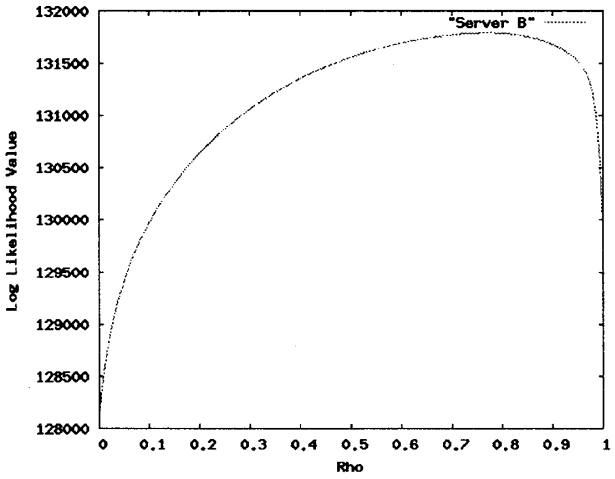


図 1: サーバ B の対数尤度関数 $\log L(\rho_B)$

表 2: 数値計算による $\rho$ の最ゆう推定結果

| | サーバ A | サーバ B |
|---|---|---|
| $\hat{\rho}$ の値 | 0.904 | 0.780 |

### 4.3 本実験

それぞれのサーバの http リクエストデータに対し，予備実験で推定した $\hat{\rho}$ を使って接続数の予測を行い，非定常な $\lambda_t$ をパラメータに持つ岩田のモデルと定常な $\lambda$ のモデルとで予測二乗誤差の平均を算出し，観測値と予測値をプロットして比較評価を行った．

表 3: 予測二乗誤差平均の比較

| 予測二乗誤差平均 | 非定常ポアソン | 定常 |
|---|---|---|
| サーバ A | 6.20 | 7.92 |
| サーバ B | $1.49 \times 10^3$ | $3.74 \times 10^3$ |

## 5 考察

### 5.1 予備実験についての考察

図 1 からわかるように，(15) で表わされる対数ゆう度関数は一見複雑に見えるものの，右辺第一項と第二項に含まれる変数の違いはわずか (第三項と第四項も同様) であり，$\log L(\rho)$ の形状は単峰形となった．したがって $\rho$ の最ゆう推定は解析的には不可能であるが，山登り型のアルゴリズムで数値計算を行うことで推定が可能であることがわかる．今回は非定常なモデルを仮定しているため $\rho \neq 1$ となるが，この予備実験で得られた $\hat{\rho}$ の値とデータの特性との関係については次節で論ずる．

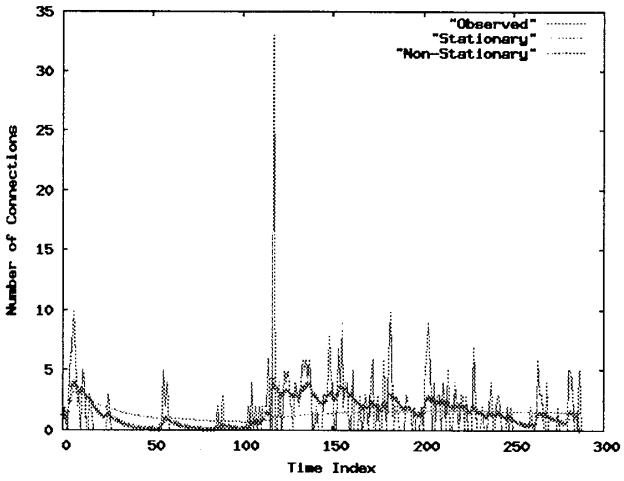


図 2: サーバ A の接続数の予測値および観測値

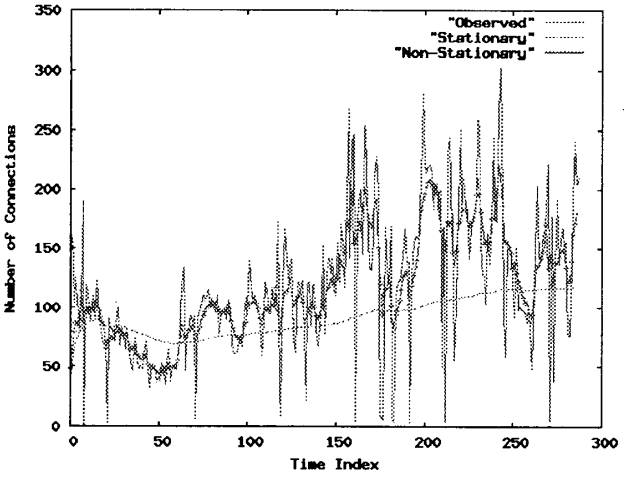


図 3: サーバ B の接続数の予測値および観測値

### 5.2 本実験についての考察

表 3 の通り，サーバ A, B のデータとも非定常なポアソンモデルのほうが定常のものよりも予測二乗誤差平均が小さい値となった．これはポアソン分布のパラメータ $\lambda_t$ に非定常性を考慮した効果が予測値にも表われたということになるだろう．さらに両モデルの予測二乗誤差平均の比はサーバ A で 1.00:1.27 であったのに対し，サーバ B では 1.00:2.51 と約 2 倍になっている．そこで，表 2 の $\hat{\rho}$ の値を比べてみると，サーバ A では $\hat{\rho_A} = 0.904$, サーバ B では $\hat{\rho_B} = 0.780$ となっている．(6) を見てもわかるとおり，今回適用したモデルでは $\rho$ の値が小さいほど $\lambda_t$ の変化は激しくなるので，サーバ B のように接続要求数の変動の激しいサーバに対しては今回適用した岩田の非定常モデルが定常モデルに比べて有効に働いていることがわかる．これは観測値と予測値をプロットした図 2 および図 3 を見ても，接続数のスケールの違いはあるものの視覚的に見てとることができる．以上より，Web サーバ毎のトラヒックデータから $\rho$ を推定して接続要求数の変動の指標とすると，Web サーバの仕様決定や運用のチューニング等に有用となる可能性がある．

## 6 まとめおよび今後の課題

本研究では，http アクセスの実データに対して，非定常なポアソン分布のパラメータを解析的に更新可能な時系列モデルを採用し，いくつかの解析例を示した．このモデルは非定常性を表現するパラメータ $\rho$, $(0 < \rho \leq 1)$ を利用しているため，$\rho = 1$ の定常ポアソン分布を仮定したモデルに比べて様々なタイプのトラヒックデータに適応する可能性がある．また $\rho$ の値をサーバ毎に推定することでサーバの設計や運用の際の参考となる可能性がある．

今後の課題としては，まず対象となるトラヒックデータに関して，他のプロトコルのデータについての検討や，パケット単位のデータを用いた詳細なトラヒック解析などが挙げられる．さらに今回取り上げた解析モデルがネットワークトラヒックの特徴といわれる自己相似性や長期依存性を表現しうるかどうかの検討も挙げられる．また，このモデル自体に関して，$\rho$ の最ゆう推定法以外の推定法の検討や，$\lambda_t$ の事前分布のパラメータの初期値 $(\alpha_1, \beta_1)$ の設定のしかたの検討などが挙げられる．

## 参考文献

[1] 岩田錦弥, 吉田隆弘, 松嶋敏泰, “ポアソン分布に従う非定常な時系列のモデル化に関する一考察,” 電子情報通信学会研究技術報告, IT2003-39(2003-07), pp.93-98, 2003 年 7 月.

[2] 佐藤昌平, 吉田万貴子, “次世代インターネットとトラヒック工学”, 電子情報通信学会論文誌, Vol.J-85-B, No.6, pp.875-889, 2002 年 6 月.

[3] 篠秀明, 北澤慶一, 八名和夫, “インターネットアクセスネットワークにおける HTTP 通信リクエスト発生の非定常解析法,” 電子情報通信学会論文誌, J84-B, No.8, pp.1494-1504, 2001 年 8 月.

[4] 名部正彦, 馬場健一, 村田正幸, 宮原秀夫, “インターネット・アクセスネットワーク設計のための WWW トラヒックの分析とモデル化,” 電子情報通信学会論文誌, J80-B-I, No.6, pp.428-437, 1997 年 6 月.

[5] Vern Paxon and Sally Floyd, “Wide Area Traffic: The Failure of Poisson Modeling, ” *IEEE/ACM Trans. on Networking*, Vol.3, No.3, June 1995.