F-009

粒子フィルタを用いたWebサイトのトレンド推定

Trend Estimations of Web Sites using Particle Filter

佐藤 哲 †
Tetsu R. Satoh†

1. はじめに

テレビの視聴率と同様，インターネット上でもWebサイトのアクセス数はビジネスにおいて重要な情報であり，様々な測定方法が提案されている．有名な例ではAlexa社[1]，日経BP社[2]，サイボウズ・ラボ社[3]等による各発表があり，検索サイトを運営している場合は検索語単位の，Google社[4]，Yahoo!Japan社[5]等による各発表がある．その他にもブログサイトに特化した調査や検索タグを対象にした調査等様々な調査が発表されており，より正確な実態を反映させるための改良が行われている．

そこで本発表では，対象を単純なWebアクセスとした上で粒子フィルタにより正確なアクセス数を隠れ変数とみなした推定を行い，時系列データのトレンドを推定する手法を提案する．

2. 粒子フィルタ

粒子フィルタ [1] は逐次モンテカルロ法と呼ばれる非線形確率分布の推定手法の一種で，本研究では粒子フィルタの実装としてはモンテカルロ・フィルタ [2] を用いる．

粒子フィルタは確率密度分布を粒子の集合により近似し，(a) システムを観測した場合のノイズを含む観測結果を表す観測モデル，(b) システムの変化を表すシステムモデル，の二つのモデルにより現される．本研究では，観測モデルは時刻 t_k におけるWebサイトへのアクセス数そのものであり，システムモデルは時刻 t_{k-1} でのシステムの状態にガウスノイズまたはコーシーノイズを加えたものとする．すなわち，観測モデルおよびシステムモデル

$$\begin{cases} x_k &= F(x_{k-1}, v_k) \\ y_k &= H(x_k, \omega_k) \end{cases} \qquad (1)$$

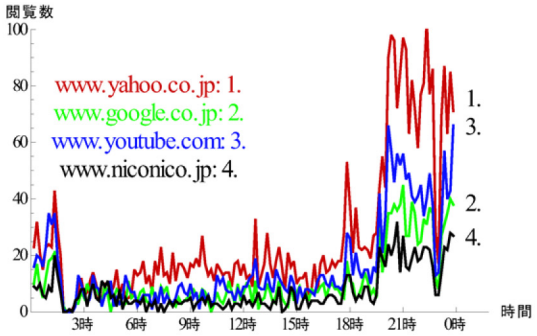

図 1: 入力データ

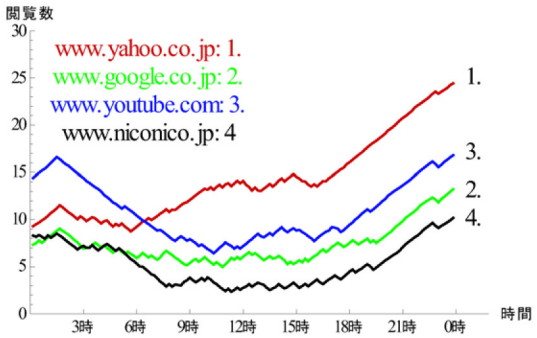

図 2: ガウスノイズ（分散値小）

において，

$$\begin{cases} x_k &= x_{k-1} + v_k \\ y_k &= g(t_k) \end{cases} \qquad (2)$$

である．ただし，

$$v_k \sim \begin{cases} \dfrac{1}{\sqrt{2\pi\rho^2}} \exp\left(-\dfrac{x_{k-1}^2}{2\rho^2}\right) & \cdots \text{ ガウスノイズ} \\ \dfrac{\tau}{\pi(x_{k-1}^2 + \tau^2)} & \cdots \text{ コーシーノイズ} \end{cases}$$

であり，$g(t_k)$ は時刻 t_k でのアクセス数測定値である．また，t_k には隠れ変数の真の値にガウスノイズが加わっていると仮定し，従って再サンプリングの際の粒子の尤度計算にはガウス分布を用いる．

3. 隠れ変数値推定実験

2009年5月17日の24時間の，株式会社オプトリンクスによるWebサイトアクセス数測定結果を

† 株式会社オプトリンクス，Optlynx Co.,Ltd.

[1] http://www.alexa.com/topsites
[2] http://www.nikkeibpm.co.jp/bz/chosa/w_brand/
[3] http://pathtraq.com/ranking
[4] http://www.google.com/trends
[5] http://searchranking.yahoo.co.jp/

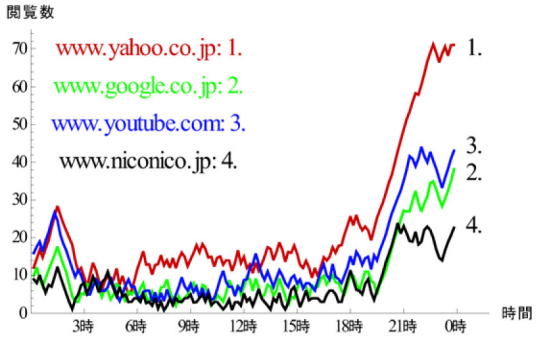

図 3: ガウスノイズ（分散値大）

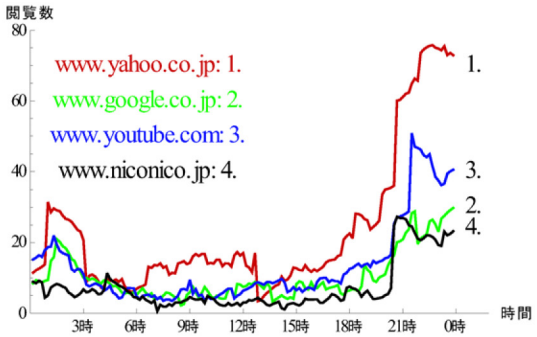

図 4: コーシーノイズ

オリジナルデータとし，粒子フィルタを適用した結果を報告する．オリジナルデータの中で，本稿では 24 時間のアクセス数の多さの上位 4 位までのサイトである，www.yahoo.co.jp，www.google.co.jp，www.youtube.com，www.niconico.jp（以下，yahoo，google，youtube，niconico と記す）を対象とする．

図 1 に，時系列オリジナルデータを示す．変動が激しく，yahoo のアクセスが最も多いことは容易に確認できるが，一日の傾向や他の 3 サイトとの相対差が理解し難いことが分かる．

次に，システムノイズの分散を $\rho = 0.1$ として，オリジナルデータに粒子フィルタを適用した結果を図 2 に示す．深夜 0 時頃から明け方 6 時頃までの深夜時間帯は youtube へのアクセスが多く，それ以外の時間帯は yahoo へのアクセスが多いことが容易に理解できるが，オリジナルデータとのアクセス数の差が大きく，アクセス数の絶対値の評価が難しい．

また，図 3 にシステムノイズの分散を $\rho = 1.0$ と大きく取った場合のシミュレーション結果を示す．図 2 に比べると，オリジナルデータのアクセス数が比較的よく反映されており，オリジナルデータである図 1 に比べるとノイズ等による変動が平滑化されており，データ分析が容易になっていることが分かる．

ところで，Web サイトへのアクセス数測定のような極めてノイズの混入や日時による変動が激しい時系列データに対しては，コーシーノイズのような非線形の確率分布を仮定することが有効であることが知られれている．そこで，システムノイズとしてコーシーノイズを導入し，パラメータ $\tau = 0.02$ としてオリジナルデータに対し粒子フィルタを適用した結果を図 4 に示す．この場合，オリジナルデータの数値も良好に反映されており，さらに午後以降の時間帯に一時的に yahoo や google へのアクセス数が落ち込み youtube へのアクセスが増加する様子や，夜間 21 時過ぎに niconico へのアクセスが急増する様子が確認できる．

以上の分析結果はいずれもパソコンユーザへのアンケート結果や長期間のログ分析結果と合致しており，粒子フィルタの適用結果が妥当であることを示唆していると思われる．

4. サイトアクセス数測定の考察と課題

一般に，Web サイトのアクセス数の測定や推定は困難である．それには多くの理由があるが，特に以下のような問題が重要である：

- 多様性
  他のメディアに比べ，数や種類が非常に多い（いわゆるロングテール）
- ノイズ
  クライアントまたはサーバソフトウェアの種類や設定，セキュリティソフト等の設定，回線状況，時間帯，プライバシーを考慮した記録方法等により，測定結果に大きなバラつきが生じる
- 単体 PC の複数ユーザ利用
  1 台の PC を家族で共有するような状況自体は他のメディアと同様であるが，より時間帯や曜日による複雑な変化が現れる

これらの解決には，他の統計的手法や情報量基準を併用する階層的なアプローチが有効であると思われる．

5. おわりに

いくつかの Web サイトへのアクセス数を測定し，その真の値を隠れ変数として粒子フィルタを用いて時系列のトレンドを推定する手法を報告した．今後はさらに個人別にデータを分類し [3]，Web サイトの個人へのレコメンデーションシステムへ拡張する予定である．

参考文献

[1] 樋口知之：粒子フィルタ, 電子情報通信学会誌, Vol. 88, No. 12, pp. 989–994 (2005).

[2] 北川源四朗：モンテカルロ・フィルタおよび平滑化について, 統計数理, Vol. 44, No. 1, pp. 31–48 (1996).

[3] 佐藤哲：文字列圧縮を用いた Web サイトのクラスタリング, 日本応用数理学会 2007 年度年会講演予稿集, pp. 170–171 (2007).