5B-4

粒子フィルタを用いたユーザのネット上での検索要求背景推定

佐藤 哲 †

楽天株式会社 †

1. はじめに

インターネット上でのキーワード検索は，ユーザの能動的な行動でありユーザの要求やユーザそのものの属性など，豊富な情報が含まれていると考えられる．そこで本発表では，ユーザの検索動向を追跡することによってユーザが何を求めているか，そのユーザはどのような属性を持っているかを推定する試みについて述べる．また，追跡には粒子フィルタが適していることを説明する．

2. キーワード検索についての仮説

インターネット上には大量のデータが存在するため，希望するデータを得るために検索技術が利用される．多くの情報検索手法が提案されているが，1つ以上の単語を利用するキーワード検索が未だ広く利用されている．そこでユーザがインターネット上でキーワード検索を行う場合の状況を想定すると，次のような仮説は妥当であるように思われる：

(1) 何か非常に興味がある事項がある場合，同一または類似の単語の検索を繰り返す
(2) 検索キーワードが変化した場合，興味も変化した可能性がある
(3) 興味が強い事項は繰り返し検索の間隔が短い
(4) 興味が強い事項は繰り返し検索の期間が長い
(5) 多くのユーザが同一のキーワードを検索していることが検出された場合，まだその単語を検索していないユーザもそのキーワードに興味を持つ可能性がある
(6) 繰り返し検索の間隔・期間にはユーザの特性が現れる

これらの妥当性を検証するために，ユーザの検索ログの解析を試みる．しかしよく知られているように，Web のログ解析にはノイズ混入・各ユーザによる差異の大きさ等の問題があり，統計処理が難しい．

そこで本研究では，ユーザが個人を特定できない ID を持っていて識別可能であるという仮定のもとで，データの異常値や欠損値に強いと言われている粒子フィルタを用いてユーザの検索ログの追跡を行う．

Estimating Purposes of User's Search Behaviour on the Internet by Particle Filters
†Tetsu R. Satoh, Rakuten Inc.

3. 粒子フィルタ

粒子フィルタは，時系列の信号入力に対しデータの再サンプリングを繰り返しながら逐次的に保持データを更新していく逐次的モンテカルロ法と呼ばれる手法である [1]．粒子フィルタでは，次のような時系列状態空間モデルを用いる：

$$x_n = F(x_{n-1}, v_n) \tag{1}$$

$$y_n = H(x_n, w_n) \tag{2}$$

ここで，x は状態ベクトル，y は観測値である．v 及び w は白色雑音で，それぞれシステムノイズ，観測ノイズと呼ばれる．関数 F 及び H は任意の関数で，式（1）をシステムモデル，式（2）を観測モデルと呼ぶ．粒子フィルタはこのモデルを用いて観測値 y から状態 x を推定することが目的である．

推定する状態 x は確定値ではなく確率分布によって表され，m 個の粒子の実現値によって近似される．これを次のように表す：

$$\{f_n^{(1)}, f_n^{(2)}, \cdots, f_n^{(m)}\} \sim p(x_n|y_n) \tag{3}$$

式（3）は，時刻 t_n での観測値 y_n から粒子フィルタが推定した状態 x_n が，粒子によって $\{f_n^{(1)}, f_n^{(2)}, \cdots, f_n^{(m)}\}$ という形で保持されることを表す．$f_n^{(i)}$ は，次のような手法で求められる．

時刻 t_{n-1} の状態 x_{n-1} を表す粒子分布 $f_{n-1}^{(i)}$ が与えられているとする．また，あらかじめシステムノイズを表す確率分布 v を近似する粒子 $v_n^{(i)}$ が生成されているとする．すると，時刻 t_{n-1} の観測値 y_{n-1} から現在の状態を推定する分布 $p(x_n|y_{n-1})$ を近似する粒子 $p_n^{(i)}$ は予測分布と呼ばれ，次式で得られる：

$$p_n^{(i)} = F(f_{n-1}^{(i)}, v_n^{(i)}) \tag{4}$$

次に，予測分布に対し現在の観測値を考慮して修正を加えることを考える．そのためにまず，実際の観測値 y_n とシステムモデルからの予測である粒子 $p_n^{(i)}$ の間の違いを $\alpha_n^{(i)}$ として次のように計算する：

$$\alpha_n^{(i)} = r(G(y_n, p_n^{(i)})) \left| \frac{\partial G}{\partial y_n} \right| \tag{5}$$

ここで G は H の逆関数，r は観測ノイズの確率密度関数である．そして粒子の集合 $\{p_n^{(1)}, p_n^{(2)}, \cdots, p_n^{(m)}\}$ から，$\alpha_n^{(i)}$ に比例した確率で再サンプリングして $f_n^{(i)} = p_n^{(j)}$ とする．ただし j は次式を満たす：

$$\sum_{k=1}^{j-1}\alpha_n^{(k)}/\sum_{k=1}^{m}\alpha_n^{(i)} < \frac{j-1/2}{m} \leq \sum_{k=1}^{j}\alpha_n^{(k)}/\sum_{k=1}^{m}\alpha_n^{(k)} \tag{6}$$

このように計算された $f_n^{(i)}$ は，$p(x_n|y_n)$ すなわち観測値 y_n が得られた時に推定される状態 x_n の値を近似している．

4. 推定実験

楽天株式会社は，2010 年の年間商品売れ筋ランキングを発表した ††．上位 10 位の中で，特にヘアメディカル製品（以下，キーワード A）と新米（以下，キーワード B）が複数回ランクインしていることが目立つ．そこでこの二つのキーワードに注目し，以下のような方法でデータを抽出した．

まず，2010 年 12 月 15 日（水）から 2010 年 12 月 19 日（日）の，あるログの一部よりキーワード A と B の全ユーザ対象検索回数ランキングを作成した．次に，作成した二つのランキングの上位 3 名のユーザをグループ化した．すなわち，キーワード A に着目した仮想的なユーザ X と，キーワード B に着目した仮想的なユーザ Y の，2 名の仮想的なユーザの識別 ID を作成した．そしてこの仮想ユーザの全ての検索キーワードについて 1 時間毎の検索回数を集計し，仮想ユーザ毎に検索回数の変化を追跡した．図 1 は，実線がヘアメディカル製品に着目した仮想ユーザ X の検索回数，破線が新米の商品名に着目した仮想ユーザ Y の検索回数を表している．縦軸が検索回数で，横軸が時間を表す．横軸は 24 時間毎に区切り線を入れている．実観測データは変動が激しく，2 仮想ユーザの差異や特徴が理解しにくいことが分かる．

この観測データに対し，粒子フィルタを適用して粒子の位置の平均値の推移を図 2 に示す．実線の仮想ユーザ X は，朝，昼，夜に検索行動を起こしていることから，例えば昼休みのある社会人である可能性がある．ファッションブランド名や旅行関係の検索キーワードもあることから，購買意欲は高いと推定される．一方仮想ユーザ Y は，夕方から深夜前の時間帯に規則正しくキーワード検索を行っていることから，新米の購入タイミングを考えている可能性があり，該当の時間に商品情報を提示すれば推薦効果は高くなると思われる．

このように，検索行動を追跡することによりある程度のユーザ特性が推定でき，また検索行動活動時間のパターンを抽出できることから，効果的な情報配信に利用することも可能であることが分かる．さらに，ユーザは検索行動を止めたあとは時間がたつにつれ検索対象に対する興味が徐々に薄れていくと考えられるが，入力データが無い場合に粒子フィルタが確率分布に基づいて即時に入力無しとは判断せずに徐々に検索回数がゼロに落ちていく様子は，まさに人間の興味が薄れる様子をシミュレートしていると言える．粒子フィルタのパラメータは，システムノイズは平均 0.0，分散 4.0 の正規分布を用い，観測ノイズは台が $[-2.5, 2.5]$ の一様分布を用いた．粒子数は 3000 個で，システムモデル・観測モデル共にノイズを加算するランダムウォークモデルを用いた．

5. おわりに

粒子フィルタを用いて検索回数を追跡することで，ユーザを特定せずに検索行動パターンやユーザのデモグラフィック・サイコグラフィック特性の推定が可能であること示唆する実験結果及び粒子フィルタの特性が Web のログ解析に適することを示した．今後，システムモデルとして混合正規分布等を用い，ユーザの複数の関心事項を追跡する実験を行う予定である．

参考文献

[1] 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).

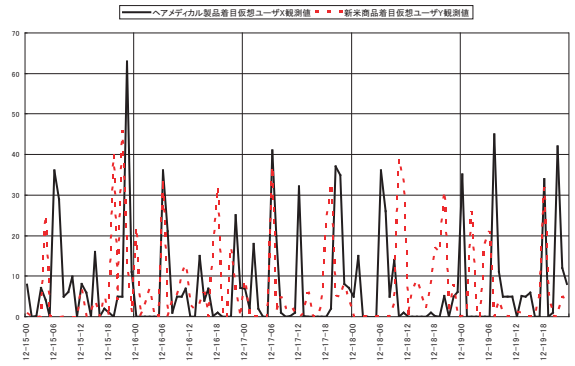

図 1: 時系列検索回数

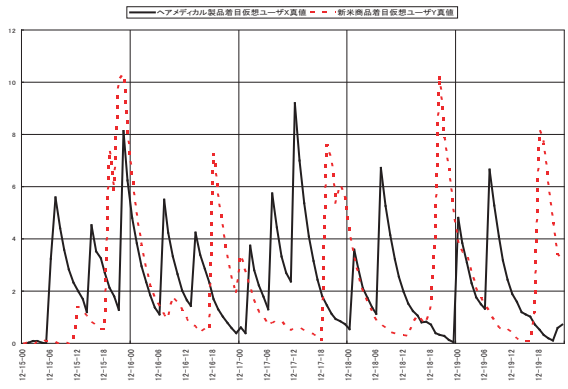

図 2: 推定検索回数真値

††http://ranking.rakuten.co.jp/yearly/