社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－144 (4)
2008－GN－ 66 (4)
2008／1／24

利用者の情報アクセス行動を推定する逐次適応モデル

陳 健　シュティフ ロマン　金 群
早稲田大学大学院人間科学研究科
{wecan_chen@fuji., roman@akane., jin@}waseda.jp

本研究では、利用者の情報アクセス行動のデータを、ショート、ミディアム、ロングといった期間や Remarkable や Exceptional などの特殊カテゴリに分けて収集・解析し、完全ベイジアン推定を基にした逐次適応モデルを提案する。

Gradual Adaptation Model for Estimation of Human Information Access Behavior

Jian Chen, Roman Y. Shtykh, and Qun Jin
Graduate School of Human Sciences, Waseda University, Japan

In this study, we present a gradual adaptation model to estimate human information access behavior. A variety of users' information access data are collected in terms of short, medium, long periods, or by categories such as remarkable and exceptional. The proposed model is then established by analyzing the preprocessed data based on the Full Bayesian Estimation.

1 はじめに

膨大な情報に囲まれている今日、効率的に情報アクセスすることは大変重要な課題になっている。インターネットは世界中の彼方此方の情報をつなげており、ウェブ上の情報検索は人間の仕事、生活の一部になった。しかし、ウェブ検索によって返ってくる結果は今では数千件から数百万件に及ぶ。利用者がその全ての検索結果を確認することはとても無理である。

情報検索行動に関して、どのように利用者の欲しい情報を優先的に提示するか、どのように利用者の興味に合わせて情報を推薦するか、が本研究の目的である。本稿では、利用者の情報アクセス行動を推定する逐次適応モデルを提案する。このモデルでは、まず利用者のアクセス行動の日常データを収集し、コンセプト（各ウェブページを分類する上位概念）ごとに記録する。次に、完全ベイジアン推定（Full Bayesian Estimation）を利用し、アクセスされる情報をショート、ミディアム、ロングといったタームごとに、利用確率を算出する。さらに、利用者の情報アクセス行動に応じて、逐次的に利用者が最も欲しい（確率の高い）情報を洗い出し、利用者に提供する。

2 関連研究

近年、情報推薦の研究は注目されている。従来の情報推薦モデルには、利用者プロファイリングが利用されている。この技術では、明示的（直接的）手法（explicit method）と暗黙的（間接的）手法（implicit method）の二つの手法がある。

明示的手法では、アンケートのような手法を利用し、利用者のフィードバックに基づいて、直接に評価する手法である。この手法で取得した情報は、利用者が直接答えたものであるために、信頼性が高いという評価がある。AntWord [Kantor 2000]では、この手法を利用し、利用者の良い評価がつけられたものを記録する。ある利用者は検索エンジンに検索をかけ始めると、入力されたキーワードを記録と較べて、同じキーワードの記録がある場合、このキーワードでの検索により最終的にたどり着いたページ、即ち良い評価のあるページを推薦する。

しかし、アンケートのようなものを評価させるという手間を、利用者側に負担させるという問題がある。また閲覧したページに評価を付けさせる方法では、ページにつけられた興味の度合いからキーワード単位の重み付けや検索式に変換する必要がある。多くの場合、ページ全体のテキストからキーワードを洗い出す方法が利用されている。この方法では、洗い出したキーワードが利用者の興味と無関係のものが入っている可能性もあるという指摘がある [Hijikata 2004]。

暗黙的手法では、①利用者プロファイルに基づく情報フィルタリング、②協調フィルタリングなどがある。

①は利用者の情報アクセス行動記録などから利用者プロファイルを作成し、それに基づいて情報推薦を行うものである。Google の "Personalized Search" というサービスでは、単一利用者の検索履歴に基づいて、より利用者の興味のあるコンテンツを提示する一つの応用例である。

②は複数の利用者が協調して利用することを前提とした情報推薦手法である。嗜好の類似度によって、利用者グループを編成し、このグループに属する利用者のアクセス情報をお互いに協調することである。Amazon.com や Movielens.umn.edu は、協調フィルタリングを用いて書籍や映画などを利用者に推薦する。

暗黙的手法は利用者のページ閲覧時間や、マウスの動きなどの情報による評価手法である。この手法は、明示的手法よりも客観的であるため、この手法に関する研究がますます進んでいる。

暗黙的手法のうち、TextExtractor [Hijikata 2004] という単一利用者のプロファイルに基づくフィルタリング手法がある。TextExtractor 手法は利用者の興味と関連して発生する操作を分析することによる暗黙的手法の一つである。具体的には、利用者のなぞり読み、リンクポインティング、リンククリック、テキスト選択などの操作を事前に観察することによって、情報推薦を行う。

ウェブアクセスのほか、利用者プロファイルによるテレビ視聴情報の推薦も提案されている [Takama 2007]。この提案はヒューマン・ロボットを使って、利用者の視聴番組ログと、視聴中の発話を収集し、プロファイルを作成する。このプロファイルに基づいて、情報を推薦する。

協調フィルタリングの手法のほうは、嗜好を利用した協調フィルタリングによる Web 情報推薦の提案がある [Kohara 2005]。この提案は Blogger によるアクセス情報から利用者の嗜好傾向を示した利用者プロファイルを生成し、その利用者の嗜好に類似した情報を推薦する方式である。

単なる嗜好の類似度による情報推薦の上、Support Vector Machine（SVM）を基にした C-SVM-CF [Oku 2006]という提案がある。この提案は利用者のコンテクスト類似度と嗜好類似度に基づいて、情報を推薦するものである。

しかし、いずれの手法も利用者の個性化、利用者情報アクセス行動の不確実性を考慮していない。テレビ視聴情報推薦と C-SVM-CF などの提案はいずれも利用者興味の変化に適応していない。TextExtractor 手法は利用者の興味の変化に適応するが、この手法で推薦されるものは最近よく触れた情報に限られている。即ち、利用者の一時的な操作を誤算し、本来の興味を無視する可能性がある。

それに対して、我々が提案する Gradual Adaptation Model（以下、GAM モデルと略する）はショート、ミディアム、ロングといったタームや Remarkable と Exceptional のような特殊カテゴリを分けて、利用者の興味変化などの不確実性によって、各タームやカテゴリにおけるコンセプトの確率を考慮することにより、より有効な情報推薦を可能にしようとするものである。

特に、今回の提案で利用されるベイジアン・ネットワークは学習機能を持っているため [Zhang 2006]、協調フィルタリングによる情報推薦のような、データが少ないと類似利用者が発見できないという問題 [Schein 2002]を避けられる。

3 ベイジアン・ネットワーク

ベイジアンネットワーク（Bayesian Networks）は 1980 年代中期に、人工知能の不確実性問題に対する研究の過程で生まれた手法である [Zhang 2006]。近年、ベイジアンネットワークの応用がさまざまな分野に広がった。確率、統計応用などの複雑な不確実性推測、データマイニング、知能情報システム構築の有力な手段になっている。

技術の側面から見ると、ベイジアン・ネットワークはランダム変数群の間の関係を記述する言語で、主な目的は確率の推定である。即ち、その原因となるランダム変数群から結果となるランダム変数群の間にある確率を推定することである。一つの例としては、原因 A ならば、結果 B になる確率は P(A | B)である。A という事象を観測することで、B に関する確信度を示す B の確率分布を推定できる。

今は、ベイジアン・ネットワークは、統計学、システム工学、情報理論などの分野で数多くの多次元確率モデルの共通フレームワークになっている。例えば、膨大な売買履歴から顧客の個人属性と商品属性との背後にある規則性を抽出する統計学領域の応用 [Shikemasu 2006]、利用者の情報アクセス履歴から嗜好などをモデル化する認知心理学領域の応用 [Kimura 2006]などの応用がある。ベイジアン・ネットワークを通じて、一つの領域の研究結果を他の領域にも適用できるため [Zhang 2006]、我々は既存の学習機能持つ完全ベイジアン推定を利用し、これを基にした逐次適応モデルを提案する。

3.1 最尤推定

この提案の前提は事前にコンセプトごとに情報リンク（以降はリンクと呼ぶ）を分類することである。この分類によって、一回のキーワード A での情報アクセス操作に関して、アクセスされたリンクのデータ・サンプル（Data Sample）が式（1）のようになる。

$$\mathrm{D} = \{D_1, D_2, ..., D_n\} \qquad (1)$$

式（1）の中にある D_1 は D_1 コンセプトに属するリンクのアクセスサンプル数である。D_2, ..., D_m は D_1 と同じことである。

例えば、一回の検索結果の各リンクに関して、二つの値を取る可能性がある。即ちクリックされる：t (true)、或いはクリックされない：f (false) である。ここで、コンセプト D_m のリンクがクリックされる確率 θ は式（2）の通りである。この式の中の X が、コンセプト D_m のリンクのクリック結果である。

$$\theta = P(X = t) \qquad (2)$$

もし 6 回のクリックのうち、コンセプト D_m のリンクが2回クリックされた場合、コンセプト D_m のリンクがクリックされた確率は $\frac{2}{6}$ で、即ち $\theta = \frac{2}{6}$ である。理論的に言うと、ここに求めた確率値は 100 パーセントの正解とは言えないが、正解に近ければ近いほど良い。この値を最大尤度と呼ぶ。

複数回、情報アクセスをすると、複数の D の集合がデータ・コレクション（Data Collection）D になる。D の条件付け確率 $P(\text{Đ}|\theta)$ が θ の尤度（Likelihood）と呼ばれる。

最尤推定（Maximum Likelihood Estimation）に関する理論に従って、コンセプト D_m のリンクがクリックされた最大尤度は式（3）の通りである [Zhang 2006]。

$$\theta^* = \frac{d_t}{d_t + d_f} = \frac{d_t}{d} \qquad (3)$$

式（3）の中で、θ^* はコンセプト D_m のリンクの最大尤度、d_t はコンセプト D_m のリンクがクリックされた回数、d_f はコンセプト D_m のリンクがクリックされなかった回数、d はクリックされた回数の合計である。

最大尤度推定では、経験値は無視されるので、一つの偶然でも推定結果に大きな影響を与える。そこで、本研究では完全ベイジアン推定を利用する。

3.2 完全ベイジアン推定

完全ベイジアン推定とは、アクセスの経験値、即ち θ の事前分布（Prior Distribution）と尤度関数（Likelihood Function）を結合し、θ の事後分布（Posterior Distribution）を求めることである。

$$P(D_{m+1} = t|\text{Đ}) = \int P(D_{m+1} = t, \theta|\text{Đ})d\theta$$
$$= \int P(D_{m+1} = t|\theta, \text{Đ})p(\theta|\text{Đ})d\theta$$
$$= \int \theta p(\theta|\text{Đ})d\theta \qquad (4)$$

完全ベイジアン推定式（4）の積分計算は大変複雑である。通常は、(a) D 内にある各サンプルが互いに独立して、同一の分布に従うという i.i.d.（independent and identically distributed）の仮定を満たす。(b) 今回クリックされたコンセプト D_m のリンクのサンプル数は d_t で、クリックされなかったサンプル数は d_f である。これと関連する事前分布はベータ分布 $B[\alpha_t, \alpha_f]$ であるという二つの仮定を前提にし、次回にコンセプト D_m のリンクがクリックされる確率の完全ベイジアン推定の結果は式（5）の通りである [Zhang 2006]。

$$P(D_{m+1} = t|\text{Đ}) = \int \theta p(\theta|\text{Đ})d\theta$$
$$= \frac{\Gamma(d_t + \alpha_t + d_f + \alpha_f)}{\Gamma(d_t + \alpha_t)\Gamma(d_f + \alpha_f)} \int \theta\theta^{d_t+\alpha_t-1}(1-\theta)^{d_f+\alpha_f-1}d\theta$$
$$= \frac{d_t + \alpha_t}{d_t + d_f + \alpha_t + \alpha_f} \qquad (5)$$

式（5）の中に、α_t は D_m のリンクがクリックされたサンプル数の経験値で、α_f は D_m 以外のリンクがクリックされたサンプル数の経験値である。式（5）において、今回クリックされた回数 d のサンプル数が少ない場合には、事前確率が結果に大きな影響を与える。d が大きくなると、事前確率の影響は弱くなる。

4 完全ベイジアン推定を用いる逐次適応モデルの提案

前述の完全ベイジアン推定を用いて、我々は利用者の情報アクセス行動のデータをショート、ミディアム、ロングの各種のタームや Remarkable と Exceptional のような特殊カテゴリに分けて収集・解析し、逐次適応モデル提案をする。

4.1 逐次適応推定モデルの提案

我々が提案するモデルの概念図は図 1 に示す。このモデルでは、先ずリンクのコンテンツによって、コンセプトごとに分類し、コンセプト(Concept)の Knowledge Base を作成する。次に、利用者を特定できる利用者情報を記録し、利用者が情報アクセスをする時に利用したキーワードとクリックしたリンクを Keyword-Concept ごとに、アクセス日付とアクセス回数をデータベースに記録する。さらに、Probability Estimator で、この記録情報に基づいて、Keyword-Concept-Term ごとにこの次にリンクがクリックされる確率を算出して、結果を Estimation Results に保存する。

我々は利用者の情報アクセス行動に応じて、情報を推薦する部分を Gradual Adaptation Recommender と呼ぶ。利用者がキーワードを入力し、検索をかけ始める場合、Gradual Adaptation Recommender は先ず事前に算出された確率結果によって、均等・ランダムで、タームやカテゴリごとにクリックされる確率の高いものを一回目の推定結果として提出する。利用者の操作に応じて、即ち、クリックされたリンクが属されているタームをより高い重み付けで、二回目の推定に優先的に提示する。

利用者の操作が続けられる場合、利用者の次の操作によって、クリックされたリンクが属しているタームまたはカテゴリの情報を優先的に提示し、次の推定を行う。

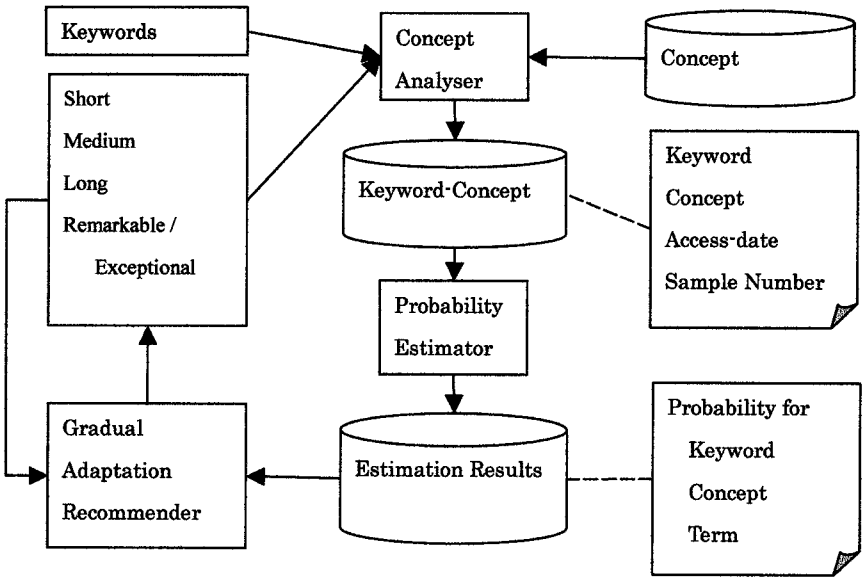

図１：逐次適応モデルの概念図

収集されたデータを処理し分析するため、利用者が使用したキーワードをコンセプトという上位概念で集計し、ある一定な期間にわたる使用の頻度により、ショート・ターム（例えば、一週間）、ミディアム・ターム（一ヶ月）、ロング・ターム（三ヶ月）といったように分ける。

例えば、利用者は最近一週間に"Apple"というキーワードを利用して、人の名前や会社名称などの情報を収集したとする。我々はこのような情報を各コンセプトにまとめる。人名の場合では、"Apple In People Name"コンセプトと言い、会社名の場合では、"Apple in Enterprise Name"コンセプトと言う。おおまかに分類するならば、"Apple In Name"コンセプトと分けても良い。全体のなかで最近一週間に使ったものであれば、ショート・タームとする。もし、それが数週間にわたって使われたものならミディアム・ターム、かなり長い間使用されているものであれば、ロング・タームとする。

ショート・タームは、利用者が短期間においてわりとある目的を持って集中的に行われたものである。それに対して、ロング・タームは、利用者が大きな関心を持っているもので、例えば、利用者の仕事関係で、ある情報を二、三日に一回、若しくは毎日収集することがある。このような常に収集する情報は、ロング・タームコンセプトに入れる。

以上のショート・ターム、ミディアム・ターム、ロング・ターム以外に、Remarkable と Exceptional という二つ特殊なカテゴリがる。Remarkable は、このコンセプトが複数のタームに属しているときに用いられる。例えば、一ヶ月間にあるコンセプトに属しているリンクの利用が続いた場合、このコンセプトのリンクはショート、ミディアム両方に属している。我々はこのようなリンクを Remarkable と呼ぶ。Remarkable が存在する場合、優先的に提示する。

利用者がたまたまクリックしたコンセプトのリンクに対して、我々は Exceptional と呼ぶ。このようなリンクは、利用者は突然調べたが二度と調べないかもしれない。しかし、利用者にとって大事なものである可能性もある。Remarkable と Exceptional の提示方法について、例えば、Remarkable がある場合、Remarkable だけを提示し、Remarkable がない場合、Exceptional の中から、ランダムに選んで提示する。

前述のように、データ記録モデルはショート・ミディアム・ロングの階層で構成される。このモデルでは、利用者の情報アクセス行動の不確実性が適切に考えられている。このモデルを用いて、利用者情報アクセス行動の推定について、下記のように考えた。

4.2 逐次適応モデルによる確率推定

逐次適応推定モデルを動かすと、利用者のアクセス行動に基づいて、キーワード・コンセプト・タームごとの確率を算出し、Estimation Results に保存する。例えば、ショート・タームを一週間として、最近の一週間に利用者がよく"Apple"のキーワードを利用して、人名や会社名などの情報を調

べたとする。"Apple"のキーワードに関して、"Apple In Name"のコンセプトに属するリンクをクリックした回数をα_t、このコンセプト以外のリンクをクリックした回数をα_fとする。今回の情報アクセス行動で、利用者が"Apple"のキーワードを利用し、"Apple In Name"のコンセプトのリンクをクリックした回数をd_t、このコンセプト以外をクリックした回数をd_fとする。次回にこの利用者が"Apple"のキーワードを使って、"Apple In Name"コンセプトのリンクをクリックするショート・タームの確率は、式（5）を利用して算出できる。同様に、キーワード・コンセプト・タームごとにショート・タームに属する他のKeyword-Conceptの確率も算出できる。同じように、ミディアム、ロングに関する確率も算出できる。

ここでは、我々が事前に"Apple"のキーワードに関連あるコンテンツを図2のようなコンセプトで名付けた。しかし、クリックされていないサンプルは利用者にとっては無意味なサンプルであると考えられるので、それを記録から除去する。

もし利用者が"Apple In Name"に属するリンクApple Brookをクリックしたら、"Apple In Name"コンセプトのサンプル数をプラス1となる。もし利用者が"Apple In Technology"に属するリンクApple iPodをクリックしたら、コンセプト"Apple In Technology"のサンプル数をプラス1となる。しかし、クリックされていないリンクには何もしない。

例えば、三ヶ月以上利用者の情報アクセス行動を記録したとする。"Apple"に関しては、ロング・タームで、"Apple In Technology"コンセプトのリンクがよくクリックされた。ミディアム・タームでは、"Apple In Music"コンセプトのリンクがよくクリックされた、ショート・タームでは、"Apple In Name"コンセプトのリンクがよくクリックされた。

上記の結果に従って、利用者が"Apple"を検索キーとして、検索をかけると、Estimation Resultsに保存されている結果をチェックする。もし重複する結果が存在したら、このコンセプトに属するリンクをRemarkableとして優先的に洗い出して、他のコンセプトに属するリンクをランダムに混ぜてから、均等に一回目の推定を行う。重複する結果が存在しない場合、ロング、ミディアム、ショート、または、Exceptionalのリンクを各3リンク（合わせて12リンク）でランダム・均等に利用者に提示する。

次に、もし利用者がミディアムリンクのどれかをクリックしたとする。その操作によって、二回目の推定を行う。この際、一回目の推定と似たように、先ず重複する結果があるかをチェックする。重複の結果が存在する場合、Remarkableとして優先的に提示し、更に、ミディアムリンクを二番目の優先順位で提示し、その次、ショート、ロングの順でリンクを各一つずつ提示する。Remarkableが存在しない場合、ミディアムリンクを優先的に提示し（例えば、6つ。3回目以降もミディアムのコンセプトがクリックされた場合、9つ）、その後、ショート、ロング、Exceptionalの順で、各コンセプトのリンクを2つ（3回目以降は1つ）、ミディアムの後ろに提示する。

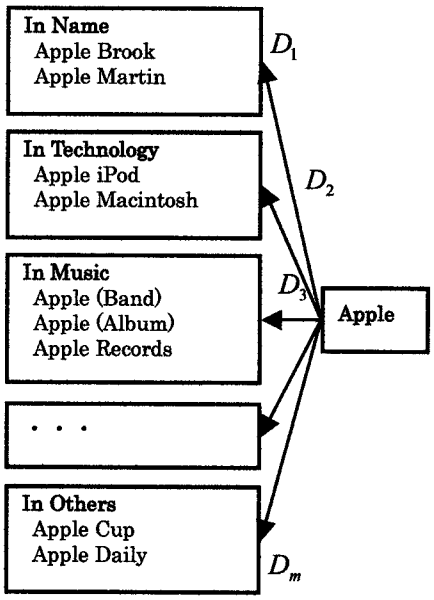

図2　Appleの検索結果のコンセプト

利用者の検索行動によって、あるコンセプトがショート、ミディアム、ロングの各タームとも属している可能性がある。この場合、我々は短い期間のほうに高い優先順を与えた。例えば、ショートは0.5で、ミディアムは0.3で、ロングは0.2の重みを付ける。そうすると、このコンセプトが複数のタームに属している場合、重みの高いタームに属させる。

その結果、ショートとミディアム両方に属している場合、このコンセプトをショート・タームとして提示する。同じように、ミディアムとロング両方とも属している場合、このコンセプトをミディアム・タームとして提示する。

4.3 複数キーワードでの検索の応用

上に記したのは、単一キーワードで検索する時の確率推定例である。実際の場合、複数のキーワードで検索するケースが多い。OneStat.comの調査によれば、一つのキーワードから七つのキーワードまで検索する人が約98.82%で、そのうち、二つのキーワードから七つのキーワードまでの人が

約 85.4%である[OneStat 2006]。

我々は複数のキーワードで検索するときの対応方法を考えた。例えば、"Apple"と"Jam"の２つのキーワードで検索する場合、クリックされたサンプル数を両方のサンプル数それぞれをプラス 1/2 にする。詳しく言うと、利用者がこの二つのキーワードで検索した後、もしプログラミングコンセプトのリンクをクリックしたら、"Apple"キーワードの"In Programming"コンセプトのサンプル数をプラス 1/2 にする、同時に、"Jam"キーワードの" In Programming"コンセプトのサンプル数もプラス 1/2 にする　(図 3)。

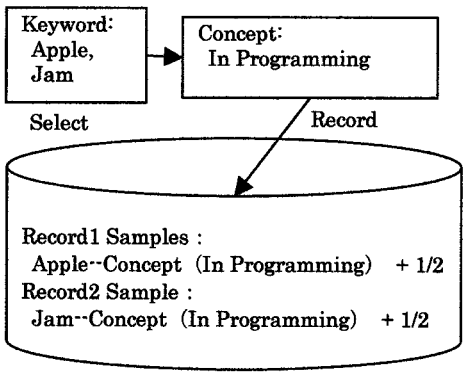

図 3. 複数キーのサンプル数記録

もし６つのキーワードが利用される場合、各キーワードの同じコンセプトのサンプル数をそれぞれプラス 1/6 にする。確率を算出する時、単一キーワードの場合と同じ、式（5）を利用して求めることができる。

複数のキーワードで検索を行った際、一部のキーワードが検索結果に含まれない場合がある。例えば、"Apple"と"Jam"で検索した結果の中に、"Apple" というキーワードが含まれない"Def Jam"があるとする。利用者が"Def Jam"をクリックした場合、"Apple"と関連するコンセプトのサンプル数が変わらないが、"Jam"と関連するコンセプトのサンプル数が一つ増すこととなる。

5　おわりに

本稿は利用者の個性化による不確実性を考慮し、学習機能を持つ完全ベイジアン推定を利用し、利用者の情報アクセス行動に応じて、逐次適応モデルを提案した。

この提案によって、利用者の興味の変化などに適応できる。さらに、利用者自身のアクセス記録に基づいて分析するため、協調フィルタリングの他人のデータによる情報推薦に比べて、より適切な情報を推薦できる。しかし、実際どのようにして利用者のアクセス記録をショート、ミディアム、ロングなどのタームに分けるか、また、どのような分け方でもっとも効果的かなどは、今後の課題となる。

さらに、GAM モデルは利用者の情報アクセス行動に基づいて、利用者の嗜好などの特性を分析できるため、利用者の他の特性を分析することも考えられる。今後の課題として、提案モデルを改善するとともに、実装し、評価を行っていきたい。

参考文献

[Kantor 2000] Paul B. Kantor, "Capturing Human Intelligence in the Net," Communications of the ACM, Vol. 43, No. 8, pp. 112-115.

[Hijikata 2004] 土方嘉徳, "情報推薦・情報フィルタリングのためのユーザプロファイリング技術", 人工知能学会論文誌, 19 巻 3 号.

[Kohara 2005] 小原恭介, 山田剛一, 絹川博之, 中川弘志, "Blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システム", The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 2005.

[Oku 2006] 奥健太, 中島伸介, 宮崎純, 植村俊亮, "Context-Aware SVN に基づく状況依存型情報推薦方式の提案", 日本データベース学会 Letters, Vol.5, No.1.

[Takama 2007] 高間康史, 難波弘樹, 岩瀬徳広, 服部俊一, 武藤優樹, 庄司俊寛, "テレビ視聴時の情報推薦に基づくヒューマン・ロボットコミュニケーション", The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007.

[Schein 2002] A. Schein, A. Popescul, L. Ungar, D. Pennock, "Methods and metrics for cold-start recommendations," Proc. 25th Annual ACM SIGIR Conference, 2002, pp. 253-260.

[Zhang 2006] Lianwen Zhang, Haipeng Guo, Introduction to Bayesian Networks (*in Chinese*), Science Press, 2006.

[Shikemasu 2006] 繁桝算男, 植野真臣, 木村陽一, ベイジアンネットワーク概説, 培風舘, 2006, pp.93-101.

[Kimura 2006] 木村陽一, 岩崎弘利, ベイジアンネットワーク技術, 東京電気大学出版局, 2006, pp.85-89.

[OneStat 2006] http://internet.watch.impress.co.jp/cda/news/ 2006/07/24/12755.html