

Web サイト訪問者のページ閲覧行動予測システム

近徳高^{*1} 高井秀人^{*2} 上里福美^{*3}

A Behavior Prediction System of the Web Site Visitors

Noritaka KON^{*1} Hideto TAKAI^{*2} Fukumi KOZATO^{*3}

Abstract: With the rapid increase of the number of web service receivers, the presentation style in the website has been becoming more important in terms of the profits to the website owners. However, in the investigation report about the web service receivers, they feel dissatisfaction of deriving necessary information from the target sites. In order for the owner of a website to expect the profitable returns from the service receivers, he needs to provide the information that is suitable for their potential demands. We show a method of supplying the information that appeals to the site visitors, i.e., inferring their potential demands from their behavior in the website. In this paper, we propose an inference system for a website called the Behavior Prediction System that suggests the pages suitable for the visitors to proceed next. The system based on data mining algorithms and a neural network method.

1. はじめに

「平成 14 年版情報通信白書」によれば、平成 13 年 12 月の時点における日本のインターネット利用者数は 5,593 万人(対前年比 18.8%増)と推計され、1 年間で 885 万人の増加を示した。ブロードバンド加入者数も急速に増加してきており、利用者は約 2,000 万世帯と推計されている。高速回線によるインターネット利用者の増加に伴って、web 上のコンテンツは、急激に増加してきている。

Web サイトにおける情報提供のあり方は、情報提供者の利益と深く結びつけて考えられるようになった。電子商取引や企業・団体の広報・宣伝活動においては、いかに訪問者を惹き付ける Web サイトを提供するかに関心が高まっている。しかし、Web サービス利用者に関する調査報告では、「必要な情報を取りだしにくい」といった不満が多く見られるのも事実である。これらの不満を解消することが出来れば、その Web サイトを訪問者にアピールできるチャンスは増え、サイト自体の活性化がなされることになるだろう。

しかし、訪問者の要求を抽出する為に訪問者自身の情報提供を不用意に求めると、訪問者はサイト運

営母体に対して警戒心や押しつけがましさを感じてしまうという弊害が現れる。結果として、その企業・団体に関してのマイナスイメージが植え付けられてしまう懸念がある。訪問者の興味を自然に惹き付け、かつ取り込むためには、訪問者が潜在的に求めている情報を訪問者に悟られることなく探り出し、提供しなければならない。訪問者の潜在的な要求をこちらから自然かつ明示的に提案することで、訪問者に強くアピールする情報提供が可能になると考える。

これまでにも、訪問者の関心に沿った情報だけを自動的に推薦する情報推薦・検索システムは報告されている 1),2)。これらは、訪問者が膨大な情報群からの確に情報を取りだすための負担軽減を目的としている。今回我々が提案するシステムの活用目的は、Web を通しての情報提供が、提供者側の利益と結びつくことである。訪問者のニーズを的確に推測し、企業・団体の活動展開に有利な形で提供できるシステムの開発を、目標到達地点と想定している。

従来より我々は、Web ページ訪問者の目的や関心に関する知識を得るための下地として、アクセス履歴から訪問者の行動を複数の予測手法により抽出して訪問者の行動を学習して推測する、「行動予測

^{*1} 東京工芸大学電子工学専攻

^{*3} 東京工芸大学工学部電子情報工学科講師

2004 年 9 月 7 日 受理

^{*2} 東京工芸大学工学部電子情報工学科 2002 年度卒業

システム」を提案してきた3)。データマイニング手法単体でアクセス履歴を解析し、その閲覧傾向を抽出するシステムには既に報告例がある4)が、我々のシステムでは、複数の手法を同時に利用して、複数の結果からより確からしい閲覧傾向を抽出することを目的としている。本報告ではこの行動予測システムの概略について解説し、実際の運用結果を示してシステムの性能を評価する。

2. 行動予測システム概要

行動予測システムとは、ある Web サイトへの訪問者が、その階層下のページを続けて閲覧する場合に、どのようなページ移動行動を取るかを予測するシステムである(図1)。ある時点で閲覧中のページの次にどのページを閲覧するかを、訪問者の傾向により予測をする。訪問者のページ移動行動をアクセス履歴として記録・分類して訪問者の行動指向に関する基礎知識として蓄え、その知識を元に行動の分析・予測を4種類の手法で行う。cookie 情報から訪問者個人の過去の訪問のアクセス履歴が特定できる場合には、訪問者の所属するグループの属性による分類以外に個人の行動特徴も利用して予測するが、訪問者に cookie 利用を要求することは、しないものとする。

4種類の手法からの分析・予測結果からいずれを採用するかは、階層型ニューラルネットワークによって決定する。アクセス履歴から訪問者個人の今回の訪問に関する情報を入力し、各訪問者に対してどの予測手法が適当なのかを抽出する。

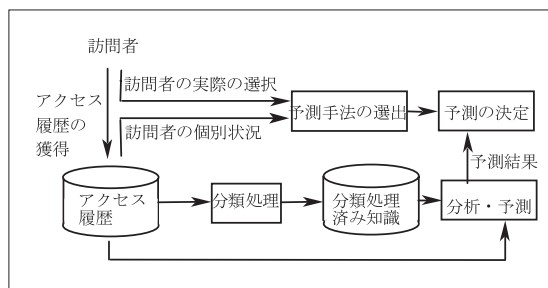


図1. システム構成

3. アクセス履歴の取得と分類

訪問者のアクセス履歴は、汎用スクリプト言語の PHP で記述したサーバ上の CGI で獲得し、データベース操作作用言語 SQL でデータベース管理を行う。訪問者のアクセス履歴を、訪問者の社会的立場や Web サービスとの関わりを示していると考えられる属性により分類し、匿名の訪問者の行動傾向を予測するための基本知識として準備する。本システムで採用した分類の為の属性は、訪問者の自覚・承諾なく取得が可能なドメイン名である。JP ドメイン種別8種とその他計9種類を、訪問者の属するグループとして設定した。

4. 分析・予測手法

アクセス履歴を分類して蓄えられたドメイン毎の基礎知識を用いて、各訪問者が次に閲覧するページの予測を行う。Apriori アルゴリズム5)、相関係数、アクセスランキング1及び2の4種類の手法で、複数の予測を立てる。

4.1. Apriori アルゴリズム

Apriori アルゴリズムはデータマイニング手法のひとつで、大規模データベースから高速に相関ルールを導出することを目的とする。対象間の相関関係の強さは、支持度、信頼度で表現される。今回の利用では、訪問者が1回のセッションで閲覧したページを1アイテムセット(1トランザクション)として相関ルールを求め、ドメイン知識毎に支持度、信頼度の高いページを抽出する。

4.2. 相関係数

ピアソンの積率相関係数により、全ページ間の相関係数を求める。訪問者の属するアクセス履歴のドメイン知識からアクセス回数を抽出し、各ページへのアクセス回数からページ間の相関係数をすべて求め、相関関係の強いページを抽出する。

4.3 アクセスランキング1

各ページから他の各ページへのアクセス数をカウントし、カウント数が多いページの上位5つを移動先の予測結果として出力する。訪問者がある一定

回数以上訪問したことが明らかな場合はその訪問者個人におけるランキングを予測結果を利用して予測するが、それ以外の場合は所属するドメイン ID におけるランキングを予測結果を抽出する。

4.4 アクセスランキング 2

1 と同様に、現在閲覧中のページからアクセスした実績の多いページの上位 5 位までを次に閲覧するページの予測結果とするが、このとき現在閲覧中のページには、いずれのページから移動してきたかの情報も考慮に入れて予測する。予測結果は、所属ドメイン ID ごとに抽出する。

5. 各予測結果の比較

東京工芸大学電子情報工学科の Web サイト（22 ページ構成）への訪問者に関して、2002 年 1 月から 2003 年 12 月末日までのアクセス履歴を収集し、訪問者の実際のページ間移動の状態と 4 種類の手法からの予測の一致率を比較する。各予測には、アクセス履歴収集開始後からその予測時点までの全ての履歴を利用する。

表 1 月毎の各手法の予測一致率

予測 期間	総予 測数	手法毎の予測の一致率 (%)			
		相関 係数	Apriori	アクセス ランキング 1	アクセス ランキング 2
2002 年 1-3 月	2534	51.4	34.7	42.7	39.8
2002 年 4-6 月	2747	54.2	36.6	38.6	45.3
2002 年 7-9 月	4333	54.1	39.0	43.3	47.7
2002 年 10-12 月	4236	59.3	41.6	44.0	50.2
2003 年 1-3 月	4426	47.0	31.5	40.6	40.0
2003 年 4-6 月	1691	54.5	24.9	34.8	43.8
2003 年 7-9 月	3907	60.2	34.2	38.8	47.4
2003 年 10-12 月	4289	61.9	37.6	40.1	50.4

※2003 年 5 月分は web サーバの故障によりデータ無し

各手法毎に、訪問者が閲覧しているページからサ

イト内で次に移動するページを 5 ページ予測し、訪問者が実際に移動したページがそのいずれかと一致した場合を、予測と行動が一致したものとした（表 1 参照）。この表より以下が読み取れる。

1. 3/6 割程度の一致を得た。
2. 圧倒的に高い一致率を示す手法はなかった。
3. アクセス履歴の蓄積に対し、一致率の顕著な増加は確認できない。

表 2 手法間の予測結果一致率% (2002.1-2003.12)

	相関 係数	Apriori	アクセス ランキング 1	アクセス ランキング 2
相関係数	-	43.1	29.7	37.8
Apriori	-	-	31.4	36.7
アクセス ランキング 1	-	-	-	30.7

また、手法間で比較してみたところ、予測結果が互いに一致するのは 3 割から 4 割半ば程度であることから、手法毎にかなり異なった予測を行なう傾向があることがわかる（表 2 参照）。

6. 予測法選出機能

手法ごとに正答一致率にばらつきは存在するが、圧倒的な精度で予測が行える手法は提案できていない。各手法でデータの利用方法が異なる為に予測結果にばらつきが生じ、場合によっては、一致率が最も低いと判断される手法で抽出された予測が正しい場合もある。属性の近い訪問者であっても、どの手法の予測が一致するかについては、訪問タイミングなど、訪問時によって異った傾向があるとも考えられるからである。

しかしながらこの傾向を解析することは、各訪問者の訪問時の状況とその行動傾向の非線形な相関関係を推定することに相当し、容易ではない。そこで、適切な予測手法の選出機能の実現のために、階層型ニューラルネットワークを導入した。これにより、閲覧行動予測システムの予測率を上昇させる。

訪問者の現在の傾向を手がかりとして、各予測時に最適な予測手法を選択する。学習法はバックプロパゲーション法を採用し、教師信号はその訪問者が実際に次にアクセスしたページを提案した予測法

とする。訪問者の現在の行動状況と、最適な予測法との関係をネットワークに学習させることによって、各訪問者に最適な予測法を決定することが出来るネットワークを構成する(図2)。

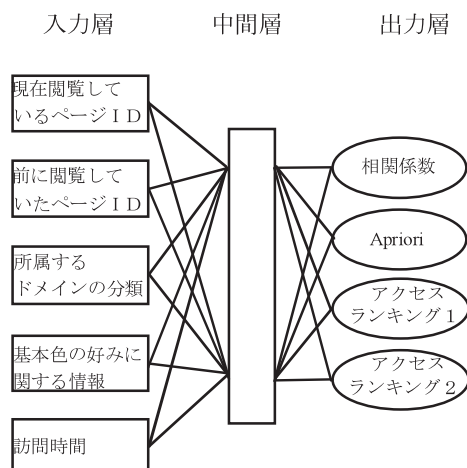


図2 ニューラルネットワークの構成

入力層は、訪問者が現在閲覧しているページID、一つ前に閲覧していたページID、所属するドメインの分類、その時点での色の好みの傾向に関する情報、訪問時間帯の五つに対応した素子群からなる。色情報の利用は、訪問者のサイト訪問時点での心理状態を示唆する可能性を期待しており、サイト訪問時に訪問者がページを構成する基本色の変更を何に希望したかによって判断している。出力層には各予測法に対応した素子を持っている。

7. 予測結果選出機能の評価

追加された機能が一致率に与えた影響について、2002年1月から2003年12月までの全予測結果を用いて検証を行った。ニューラルネットワークが選出する手法による予測の正答一致率を、3ヶ月ごとに区切って算出した(表3参照)。選出した手法の出した予測のいずれかが訪問者の行動と一致する場合を、正答一致率に反映させる。またネットワークの学習も、履歴データを3ヶ月分ずつ追加して行なうこととした。例えば、2002年1-6月分までのアクセス履歴で学習を終えているネットワークでは、2002年7-9月間の予測で使用する手法を選出し、さ

らにそのネットワークに2002年7-9月分のデータを追加してチューンナップ学習をしてから、2002年10-12月分の予測に使う手法の選出をする。

ニューラルネットワークが選出した手法による予測の一致率と、各手法を単独で用いた場合の予測一致率を3ヶ月毎に比較してみると(図3参照)、ネットワークを利用した結果が、各手法単独の一致率を上回っている。これは訪問者の状況に基づいた学習を行ったネットワークが、場合に応じて適当な予測法を選出している可能性を示している。また多少の上下動はあるものの、基本的には長期間に渡っての学習を経るほど、予測結果は向上に向かっていく。これらの傾向は、ニューラルネットワークによる予測手法選出の有効性を示していると見て出来るだろう。

図3に示すグラフでは、ある程度ネットワークの学習が進んだ後に、一時的に正答一致率の低下が見られる(2003年、2004年共に1-3月)。一時的なニューラルネットワークの性能低下は、この時期のアクセス履歴が普段とは異った傾向を示しているため、そこまでに蓄積されたデータで学習を進めたネットワークが有効に働いていないことを示していると考えられる。現在のネットワークの学習は、細かいアクセス履歴の傾向変化に敏感に反応しすぎないように、3ヶ月ごとのインターバルとしているが、このことが一時的な状況の変化には適応が難しいという側面を産み出している。正答率が大幅な低下を見た場合、一定短期間だけ学習インターバルを小さくする変則的な学習インターバルの採用で、急激な傾向変化にネットワークが対応出来るかについては、今後の検討課題である。また、さらに複数年のアクセス履歴による運用してみることによって、サイト訪問者の長期間に渡る変動の傾向をネットワークに学習させることができるかも検討したい。

表3 ニューラルネットワークの選出した予測法の一致率の3ヶ月毎の平均値

予測期間	総判定数	2002年1月からの学習累積期間別の予測の正答一致率(%)							
		2002年 3月ま で	2002年 6月ま で	2002年 9月ま で	2002年 12月ま で	2003年 3月ま で	2003年 6月ま で	2003年 9月ま で	2003年 12月ま で
2002年1-3月	2534								
2002年4-6月	2747	57.0							
2002年7-9月	4333	59.3	60.1						
2002年10-12月	4236	62.6	65.7	63.7					
2003年1-3月	4426	52.2	53.2	53.4	51.7				
2003年4-6月	1691	56.7	58.8	56.3	57.1	57.3			
2003年7-9月	3907	64.7	66.5	67.5	65.8	66.0	68.5		
2003年10-12月	4289	66.6	68.9	68.4	68.9	70.8	69.6	70.9	
2004年1-3月	2607	62.7	63.6	63.0	65.0	64.3	66.9	65.3	64.3

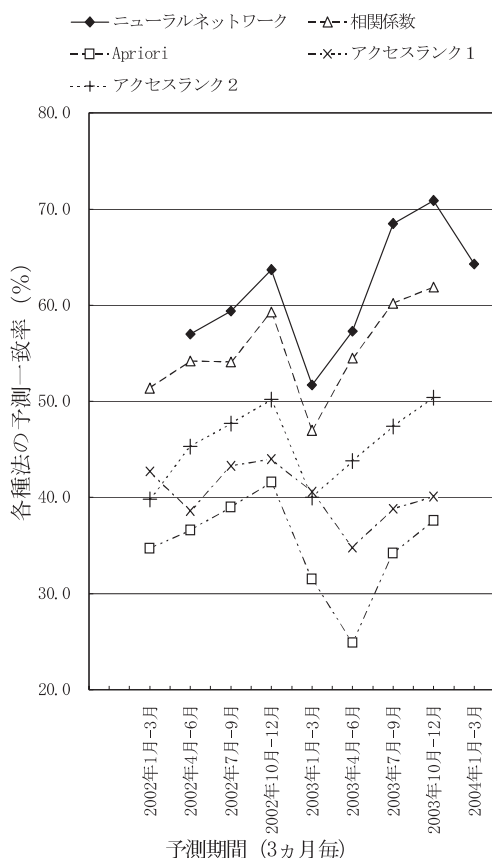


図3 ニューラルネットワークの未学習データに関する一致率の比較

8. 結論

ニューラルネットワークを介した閲覧行動予測

システムでは、訪問者の行動の5/7割程度を予測することが出来た。学習を重ねるにつれて一致率が上昇傾向を見せ、より適した予測法の選出が行われる傾向が示された。一方、単独手法の採用ではこれを越える好成績をあげることはなく、情報の蓄積による顕著な予測結果向上の傾向も見せなかった。このことから、訪問者の行動予測システムとして、ニューラルネットワークを介した複合システムの利用が有望である可能性を示すことが出来た。

今回の評価は単一サイトへの訪問者に限って行われている。サイトの特性が予測結果の一致率に影響を与える可能性は、見逃すことが出来ない。本システムの現実的有効性を検討するためには、更に様々なサイトでの更なる運用テストが必要である。

参考文献

- 1) 野美山浩他, 「個人適応型情報検索システム-個人の興味を学習する階層記憶モデルとその協調的フィルタリングへの適応」, 情報処理学会 情報学基礎 42-8, pp.49-56, 1996.
- 2) 有吉勇介, 「組み合わせフィルタリング方式を用いた情報推薦システム」, 情報処理学会 第60回全国大会論文集, pp.123-124, 1999.
- 3) 岩本善暁他, 「Apriori アルゴリズムによる Web ページ閲覧者の行動予測」第16回人工知能学会全国大会, 5.2002

- 4) 4) 加藤久慶他,「データマイニング手法を利用した Web アクセスログへの適用」, 第 15 回人工知能学会全国大会, 5.2001
- 5) 5) R.Agrawal&R.Skirant,"Fast Algorithms for Mining Association Rules",Proc.of the 11th Conference on Very Large Databases, pp.487-499, Santiago, Chile, 9.1994