研究成果総合報告書

自己組織化マップを用いた
適応型ウェブマーケティングシステムの
研究開発

平成 18 年 4 月

財団法人ファジィシステム研究所
株式会社キューブス

目 次

第1章　研究開発の概要

1.1　研究テーマ名

「自己組織化マップを用いた適応型ウェブマーケティングシステムの研究開発」

1.2　研究開発の背景・研究目的及び目標

1.2.1　研究開発の背景

インターネットの普及に従い，インターネットを利用した広告手法が広く用いられるようになっている．Ｗｅｂ広告は，その代表である．Ｗｅｂ広告とは，閲覧したホームページ上に自動的に広告が掲示されるものである．

このＷｅｂ広告には，閲覧者の興味を全く無視した広告を強制的に掲示するという問題がある．利用者が自分の興味のある項目を登録すれば，その興味に合致する広告が提示されるが，近時のプライバシー意識の高まりにより，このような自己に関する情報を登録してもらうことは困難になりつつある．また，閲覧者の興味は移り変わるものであり，一旦登録した興味ある項目が，常に閲覧者の嗜好に合致するわけでもない．

そこで，(1) 興味のある分野を事前に登録するという作業が不要であり，しかも (2) 嗜好が変化した場合にも，その嗜好の変化に対応して的確な広告を提示できるＷｅｂ広告配信システムが求められていた．

1.2.2　研究目的及び目標

本事業では，(1) 興味のある分野を事前に登録するという作業が不要であり，しかも (2) ユーザの嗜好の変化に自動的に追随し，更に (3) 得られたユーザの嗜好に関する情報をもとに，ユーザの嗜好を推測して，当該ユーザが興味あると思われる別のカテゴリの情報を自動的に提供する動画配信システムを開発することを目的とした．

具体的な目標は以下のとおり．

1. ネット利用者の嗜好調査
    - 嗜好調査に最適なカテゴリを選定すること．
    - ネット利用者の嗜好調査用のポータルサイトを構築すること．
    - サイトの中にあるカテゴリに注目し，そのカテゴリに対するサイト訪問者の嗜好を調査すること．
    - カテゴリ間の相関についての解析すること．

2. 自己組織化機能を持った広告配信アルゴリズムの開発

   - 自己組織化，学習機能をウェブマーケティングへ応用した事例の調査を行うこと．
   - 自己組織化機能を持った広告配信アルゴリズムを開発すること．
   - 実証実験を行い，候補手法の内から，最終的に採用する手法を決定すること．
   - 採用したアルゴリズムの改良を行うこと．

3. システム開発

   - 相関関係のあるカテゴリを提示するプログラムを作成すること．
   - ルーチン部を作成し，全体のシステムを完成させること．
   - 「自己組織化機能を持った広告配信アルゴリズム」の改良を行う都度，システムに実装すること
   - 動作テスト，デバッグ，プログラムの改良を行うこと．

4. 実証実験

   - ポータルサイトにおいて実証実験を行うこと．

1.3 成果概要

1.3.1 ネット利用者の嗜好調査

(a) カテゴリ選定

ポータルサイトを作成する際に，最も注意を払わなければならない作業のひとつに，カテゴリの選定がある．カテゴリの分類次第で，サイト訪問者の当該サイトに対する評価が大きく違ってくるので，その分類には，充分な注意を払う必要がある．

そこで，本事業では，まず，ポータルサイトに参加をしていただけそうな店舗の種類や，その数などを十分吟味したうえで，最適なカテゴリの選定を行った．

(b) データ収集用ポータルサイトの構築

ネット利用者の嗜好を調査するためのデータ収集用ポータルサイトの構築を行った．

このポータルサイトは，筑豊地域に密着したポータルサイトとし，筑豊地域の人々が地域の店舗情報を楽しく閲覧できるサイトにした．ポータルサイトの構築にあたっては，既存のサイトと連携することも考えられるが，現状では，適切なポータルサイトが見当たらないという問題があった．そこで，プロジェクト内で新規のポータルサイトを開設することとした．

また，ポータルサイトの構築にあたっては，コンテンツの収集が大きな課題となる．この部分に関しては，地域情報誌の発行機関と連携することで合意に達したので，充実したコンテンツを定期的に入手することが可能となった．

(c) ネット利用者の嗜好調査

構築したサイトにどのようにして多くのアクセス者を誘導するかが，本事業における非常に大きな課題であった．

そこで，地域外の人を可能な限りポータルサイトに誘導する工夫を行った．具体的には，既に大きなアクセス数を持っているサイトと広告契約を締結し，このサイトからデータ収集用ポータルサイトへの誘導を図ることとした．

このような準備を行った後，嗜好調査を実施した．

(d) カテゴリ間の相関についての解析

嗜好調査は，基本的にポータルサイト構築後，すべての期間において行った，これらのデータは，本実証実験における新アルゴリズムの開発おける大きなヒントを与えるのと同時に，その性能の改善検討にもおいても重要な資料となった．

特に，初年度に行った，アクセスデータの解析においては，ユーザの嗜好がクラスタ化されることがわかった．

1.3.2 自己組織化機能を持った広告配信アルゴリズムの開発

(a) 技術調査

自己組織化マップに関する国内の重要な会議に出席し，同分野の技術動向について調査を行った．

(b) アルゴリズム開発

- 平成１５年度は，解析的に求めた結果を基に，当該ユーザが現に最も興味をもっていると考えられるカテゴリを学習により求め，これを提示するアルゴリズムを開発した（以下「基本配信モデル」）．
- 平成１６年度は，これを更に発展させ，カテゴリ間の相関を動的に解析するアルゴリズムを考案した（以下「マルコフモデル」）．
- 基本配信モデルとマルコフモデルとを，様々な見地から比較検討した．その結果，両モデルともに良い結果が得らえることが判ったが，両モデルの優れた点には違いがあることも判った．
- そこで，最終製品に搭載するモデルを，どちらか一方に絞り込むのではなく，顧客の要望に応じて，両方のメニューを提供することに方針を決定した．
- マルコフモデルについては，更に改良の結果，単にランダムに広告を表示した場合と比較して，２倍以上の予測精度を得ることにも成功した．
- 両モデルの改良にあたっては，ウェブマーケティングシステムに応用する場合のマシン負荷も考慮した．

- 更に，モデル作成に必要な時系列データの長さ，モデル更新の頻度，未学習データに対する予測精度など，様々な観点から検討を重ねた．
- この結果，最終的な開発成果（アルゴリズム）では，70 %〜80 %という高い正解率でユーザの遷移を予測することができた．
- 事業化にあたり優位性を確保・維持するために，この開発成果は，特許出願を行った．

1.3.3 システム開発

開発したモデルをテストサーバに実装した．サーバに実装するに当たっては，新アルゴリズムを一つの大きなクラスの集合体として扱うこととし，今回開発したアルゴリズムを適切にシステムに実装することが出来た．

また，その一方で，アルゴリズムそのものについても必要な機能を洗い出し，細かなクラスとして構成することにした．更に，新アルゴリズムを汎用のサーバ機で稼動させるため，可能なかぎりスムーズに稼動するように調整を行った．

また，基本配信モデルとマルコフモデルのそれぞれの優位性を確かめるために，双方のアルゴリズムに対してフィールドテストを行った．そのなかで，それぞれのアルゴリズムの入れ替えをスムーズに行う必要が生じてきた．そこで，入れ替えをスムーズに行えるようにできるように，システムの構成を調整した．これは，前年度まで，システムサイドにおいて，機能ルーチンをクラス化していたために可能になったことである．

1.3.4 実証実験

平成１８年２月末までに，基本配信モデルとマルコフモデルの改良版について，１ヶ月程度の実証実験を延べ９ヶ月間行った．

実装されたシステムは，何のトラブルもなく稼動し続けており，実装当初に懸念されたサーバの資源不足という事態にも陥らないことがわかった．

1.4 事業化に向けた取り組み（今後の展望）

1.4.1 背景

次項で述べるように，本事業で開発した製品を投入するインターネット広告市場は拡大の一途を辿っている．

また，本製品の特徴である，地域情報誌とウェブマーケティングの組み合わせは極めてユニークである．加えて，本製品には，サイト訪問者の嗜好に応じた広告を提示できるという他にない機能を搭載しており，画一的な広告を配信する現在の製品に比較して，圧倒的に優位である．

そこで，本製品は，市場において一定の地位を獲得できると確信している．

1.4.2　市場規模

IT 専門調査会社の IDC Japan 株式会社が 2002 年に発表した国内におけるインターネット広告の市場についての調査によると，2001 年のインターネット広告市場規模は対前年比 29 %増の約 707 億円であった（出所：「IDC Japan 株式会社「国内インターネット広告市場動向：市場規模及び予測（IDC # J21505)」」2002 年 10 月）.

その 4 年後の 2005 年には，ネット広告全体の市場は 2,808 億円と，2001 年の 4 倍もの市場規模に拡大している（出所：電通「日本の広告費」）．また，近時ブロードバンドアクセスの普及が追い風となって，ストリーミングを利用した動画広告などが胎動を始めており，2005 年にはネット動画広告の市場規模も前年の 5 倍の 5 億円に達した（出所：インターネット広告推進協議会（JIAA)）.

このようにインターネット広告の広告形態別のシェアを激変させつつも，全体として市場は拡大基調が続いており，2006 年には 3,286 億円に達すると IDCJapan 株式会社では予測している.

1.4.3　事業化に向けた取り組み（今後の展望)

(a) 地域情報誌とのメディアミックスを前提としたパッケージ商品の開発

製品の性質上，単体での販売よりもそれに付随するあるいは連携する製品との複数商品での販売が主になると予想される．そのため，今後は，配信サーバに関連する周辺システムの拡充をおこなう．開発した周辺システムは，配信サーバとともに，製品群に組み込み，オプションという位置づけで販売を行う.

(b) 地域情報誌と Web ページを連動させた広告配信ビジネスモデルの確立

現在，ビジネスモデルとしては，広告代理店を担う会社がポイントになる．今回，このポジションに最も適している会社は，情報誌の発行業者などではないかと考えている.

また，何よりも重要なのは地域情報誌と連動することにより，地域の活性化が効率よく行えるという点にある．紙面による地域活性化にとどまらず，ウェブも利用したメディアミックスを実施することで，その効果がより大きくなると期待できる.

現在既に 2 つの情報誌との連携が決まっており，それらのウェブサイトも稼動している．今後は同じようなビジネスモデルで，他誌との連携を促進する.

(c) 配信サーバのタイプについて

開発した配信サーバは 2 種になる．基本配信モデルと，マルコフモデルがそれに当たる．これら 2 つはそれぞれに長所がある．基本配信モデルの長所は「計算負荷が低い」などが挙げられる．一方，マルコフモデルでは,「ユーザの 1 セッション内における動的な行動を予測できる」というような長所が上げられる.

今後は，これらの長所を明確にし，お客様がお客様のニーズに応じてどちらのモデルを選択するのかを選べるように製品を構成していく.

さらに，必要であれば，基本配信モデルの簡易版も準備し，より手軽な製品として市場に普及させるようにする.

(d) 販売先および販売活動について

配信サーバを導入されるお客様として想定されるのは,「広告配信業者」ないしは「ポータルサイト運営業者」になるものと推察される.

そこで,特に,ポータルサイト運営業者へのセールスを強化することを考えている.ポータルサイトへの導入・運用をかさね,ノウハウを蓄積した上で,広告配信業者へのアプローチを行う.

(e) 販売方法および販売対象・地域の多様化

販売については,当初は代理店などをおかずに直接販売のみを行うことにする.また,ニーズを調査してからのことにはなるが,自社で配信サーバを運営し,ASP 的にサービスを提供することも検討する.

販売地域については,現在は福岡県下,得に筑豊地域としているが,地域情報誌との連携モデルが確立すれば,これを広く展開できると考えている.

1.5 研究体制（研究開発責任者，研究組織・管理体制，研究者氏名，経理担当者氏名等）

1.5.1 研究開発責任者

財団法人ファジィシステム研究所　理事長　山川　烈

1.5.2 研究組織・管理体制

・株式会社キューブス
・財団法人ファジィシステム研究所

1.5.3 研究者氏名

・株式会社キューブス
　代表取締役社長　下野雅芳
　取締役 (SI 事業部長)　石村俊幸
・財団法人ファジィシステム研究所
　理事長　山川　烈
　副理事長・所長　内野英治
　主席研究員　森田博彦

1.5.4 経理担当者氏名

財団法人ファジィシステム研究所　事務局長　後藤英一

1.6　研究実施場所

・株式会社キューブス
　〒 820-0066　福岡県飯塚市幸袋 526-1-303
・財団法人ファジィシステム研究所
　〒 820-0067　福岡県飯塚市大字川津 680-41

1.7　知的財産の取得状況

特許出願中：１件
出願番号：特願 2006-068384
件名：「情報配信システム，情報配信方法及びプログラム」

1.8　対外発表の状況

・NET& COM2006
　平成 18 年 2 月 1 日（水）〜3 日（金）　東京ビッグサイト
・産学官連携フォーラム 2006
　平成 18 年 3 月 1 日（水）〜3 日（金）　福岡国際会議場
・NET& COM2005
　2005 年 02 月 2 日 (水)〜4 日 (金)　東京ビッグサイト
・NET& COM2004
　2005 年 02 月 4 日 (水)〜6 日 (金)　幕張メッセ

なお，特許出願を行うことから，これまで中核技術に関する対外発表は控えてきた．平成１８年３月に特許出願を終えたので，今後は様々な機会を捉えて，積極的な対外発表を行う予定である．

1.9　今後の当該プロジェクト連絡窓口

財団法人ファジィシステム研究所
事務局長　後藤英一
Tel：0948-24-2771　FAX：0948-24-3002
E-Mail：goto@flsi.cird.or.jp

第2章　本論

2.1　研究テーマ名

「自己組織化マップを用いた適応型ウェブマーケティングシステムの研究開発」

2.2　背景

現在インターネット環境は急速に発展しており，Web ページは新たな情報発信の媒体として活躍している．Web ページは世界中からアクセスが可能であるが，インターネット利用者が主体的に情報にアクセスしようとしなければ，その情報は存在さえ知られることはない．そこで，Web ページの存在を広く告知する必要がある．その手段の一つとして Web 広告があり，効果的な情報の告知方法として注目されている．

2.3　目的と目標

2.3.1　目的

Web 広告の利点の一つには，高度なターゲッティングができることが挙げられる．より細かく個々のユーザのニーズに応えることで効果の高い告知を実現することができる．しかしながら，従来の方法ではアクセス解析やインターネット利用の時間帯の解析等，静的な解析を行う程度であった．このような方法では，ユーザの細かいニーズに応えることができず，広告効果も一定以上は望めない．より効果の高い Web 広告配信を実現する為には，よりリアルタイムにユーザの趣向の変化を捉えることが必要となる．

そこで，我々は動的な Web 広告配信モデルを開発することを本事業の目的とした．具体的には，（1）興味のある分野を事前に登録するという作業が不要であり，しかも（2）ユーザの嗜好の変化に自動的に追随し，更に（3）得られたユーザの嗜好に関する情報をもとに，ユーザの嗜好を推測して，当該ユーザが興味あると思われる別のカテゴリの情報を自動的に提供する動画配信システムを開発することを目的とした．

2.3.2　目標

具体的な目標は以下のとおり．

1. ネット利用者の嗜好調査

- 嗜好調査に最適なカテゴリを選定すること．

- ネット利用者の嗜好調査用のポータルサイトを構築すること．
- サイトの中にあるカテゴリに注目し，そのカテゴリに対するサイト訪問者の嗜好を調査すること．
- カテゴリ間の相関についての解析すること．

2. 自己組織化機能を持った広告配信アルゴリズムの開発

- 自己組織化，学習機能をウェブマーケティングへ応用した事例の調査を行うこと．
- 自己組織化機能を持った広告配信アルゴリズムを開発すること．
- 実証実験を行い，候補手法の内から，最終的に採用する手法を決定すること．
- 採用したアルゴリズムの改良を行うこと．

3. システム開発

- 相関関係のあるカテゴリを提示するプログラムを作成すること．
- ルーチン部を作成し，全体のシステムを完成させること．
- 「自己組織化機能を持った広告配信アルゴリズム」の改良を行う都度，システムに実装すること．
- 動作テスト，デバッグ，プログラムの改良を行うこと．

4. 実証実験

- ポータルサイトにおいて実証実験を行うこと．

2.4 一般的な Web 広告配信モデル

Web 上の広告配信モデルでは，ユーザが閲覧する Web ページおよび配信する広告はいくつかのカテゴリに分類されているものとする．例えばカテゴリ名は，グルメ，ショッピング，ファッションといったものである．そして，Web 広告配信モデルはユーザの Web ページ訪問を基にして，ユーザが次に遷移する Web ページのカテゴリを予測することが目的となる (図 2.1 参照)．

Web 広告配信モデルはユーザの Web ページ閲覧の履歴情報を受け取り，ユーザの趣向を解析する．そして，ユーザがあるカテゴリの Web ページを閲覧しているときに，次にどのカテゴリの Web ページに移るのかを予測し，そのカテゴリの Web 広告を配信する．ユーザが次に移る Web ページのカテゴリは，その時点でユーザが興味を持っているカテゴリであるので予測カテゴリが一致していれば広告効果は高いといえる．逆に，予測カテゴリが全く違う場合には予測カテゴリに対するユーザの興味はないので広告効果は低いといえる．

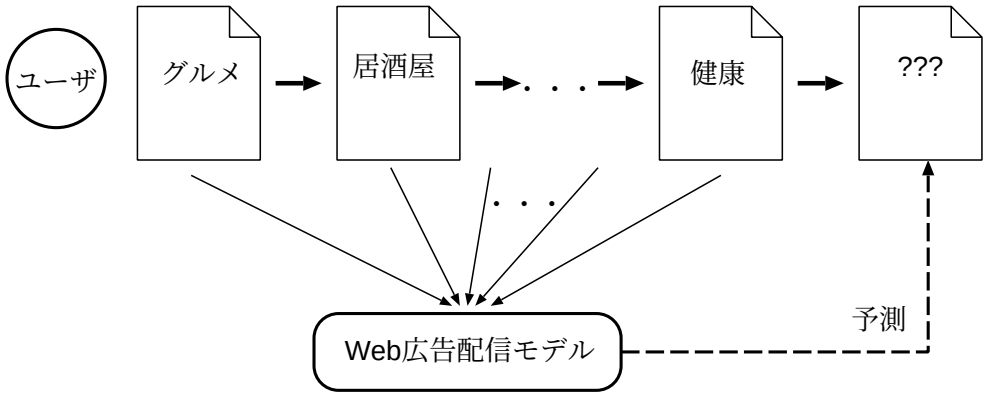

図 2.1: Web 広告配信モデル概念図

2.4.1 カテゴリ選定

ポータルサイトを作成する際に，最も注意を払わなければならない作業のひとつに，カテゴリの選定がある．カテゴリの分類次第で，サイト訪問者の当該サイトに対する評価が大きく違ってくるので，その分類には，充分な注意を払う必要がある．

そこで，本事業では，まず，ポータルサイトに参加をしていただけそうな店舗の種類や，その数などを十分吟味したうえで，最適なカテゴリの選定を行った．

2.4.2 データ収集用ポータルサイトの構築

ネット利用者の嗜好を調査するためのデータ収集用ポータルサイトの構築を行った．このポータルサイトは，筑豊地域に密着したポールサイトとし，筑豊地域の人々が地域の店舗情報を楽しく閲覧できるサイトにした．ポータルサイトの構築にあたっては，既存のサイトと連携することも考えられるが，現状では，適切なポータルサイトが見当たらないという問題があった．そこで，プロジェクト内で新規のポータルサイトを開設することとした．

データ収集用のポータルサイト用のシステム開発に関しては，その内容が 2 つに分類される．純粋なシステム部分（Java で記述）とコンテンツ部分（HTML 部分）がそれである．純粋なシステム部分については，データ収集用のポータルサイトの基幹部分であるので，全て開発した．

他方，コンテンツの HTML 化については，特殊な技術は必要とされないものの，短時間で大量のコンテンツ（原稿）を HTML 化することが要求される．そのため，HTML 部分の生成に関しては，充分な開発スピードを持つと判断される外部企業にこの作業を依頼し，その監修（サイトデザイン・サイトマップ・グラフィックデザイン・工程管理など）は株式会社キューブスが行った．

開発したサーバは，アクセスの向上と安全性確保の観点から，ハウジング契約を締結し，外部のサーバ管理業者に保管を委託した．

また，ポータルサイトの構築にあたっては，コンテンツの収集が大きな課題となる．この部分に関しては，地域情報誌の発行機関と連携することで合意に達したので，充実したコンテンツを定期的に入手することが可能となった．

2.4.3　ネット利用者の嗜好調査

ポータルサイトを構築した後には，構築したポータルサイトにアクセスする人々（サイト訪問者）を積極的に集める手段を構築しなければならない．構築したサイトに訪問してくる人が少なければ，ポータルサイトの意味をなさない．つまり，嗜好調査そのものが実施できなくなる．長いスパンで見ると，実証実験の意味をなさなくなる可能性さえはらむことになる．このように，構築したサイトにどのようにして多くのアクセス者を誘導するかが，本事業における非常に大きな課題であった．

そこで，地域外の人を可能な限りポータルサイトに誘導する工夫を行った．具体的には，既に大きなアクセス数を持っているサイトと広告契約を締結し，このサイトからデータ収集用ポータルサイトへの誘導を図ることとした．このような準備を行った後，嗜好調査を開始した．

2.4.4　カテゴリ間の相関についての解析

嗜好調査は，基本的にポータルサイト構築後，すべての期間において行った，これらのデータは，本実証実験における新アルゴリズムの開発おける大きなヒントを与えるのと同時に，その性能の改善検討においても重要な資料となった．特に，初年度に行った，アクセスデータの解析においては，ユーザの嗜好がクラスタ化されることがわかった．

具体的なカテゴリ間の相関についての解析については以下の方法でおこなった．各カテゴリごとに，当該カテゴリへのアクセス率が最も高いユーザ（訪問者）のデータを集計（合算）し，各カテゴリ間の相関の様子を検証した．

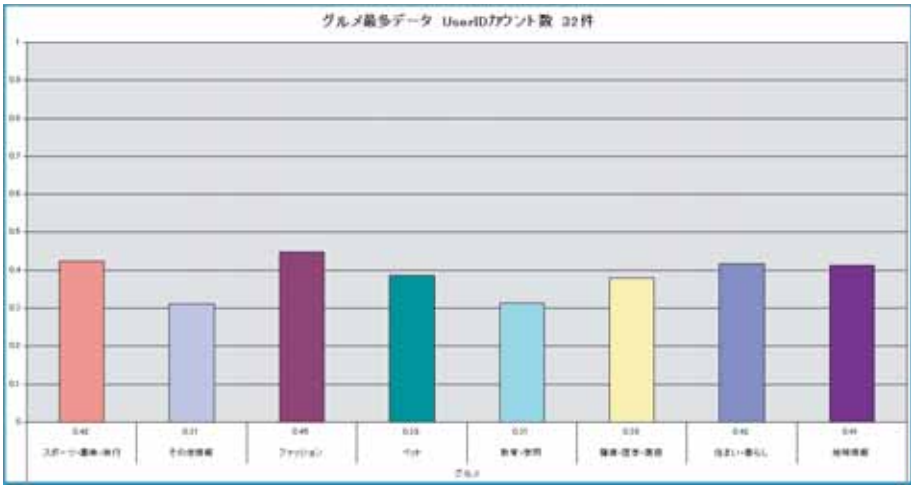

図 2.2: グルメに最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（グルメは 1.0 の値）

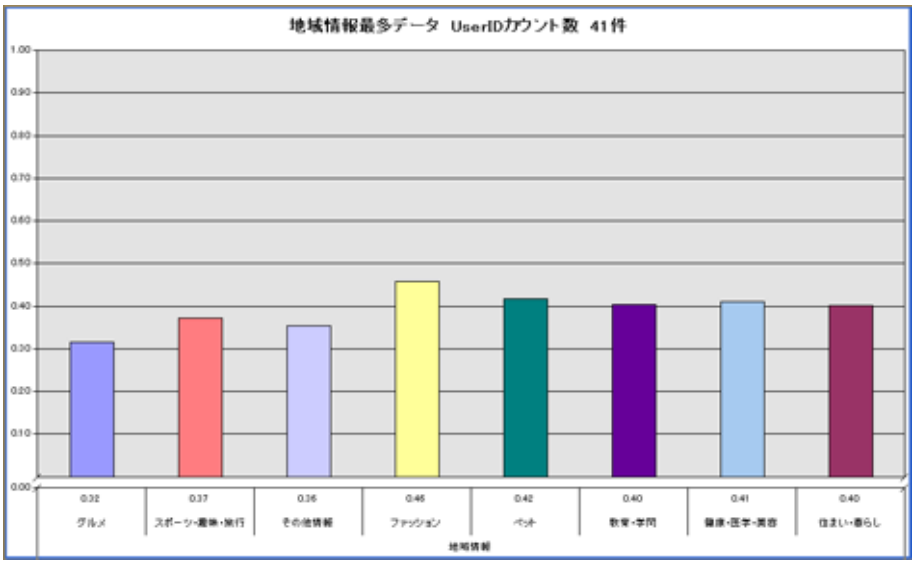

図 2.3: 地域情報に最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（地域情報は 1.0 の値）

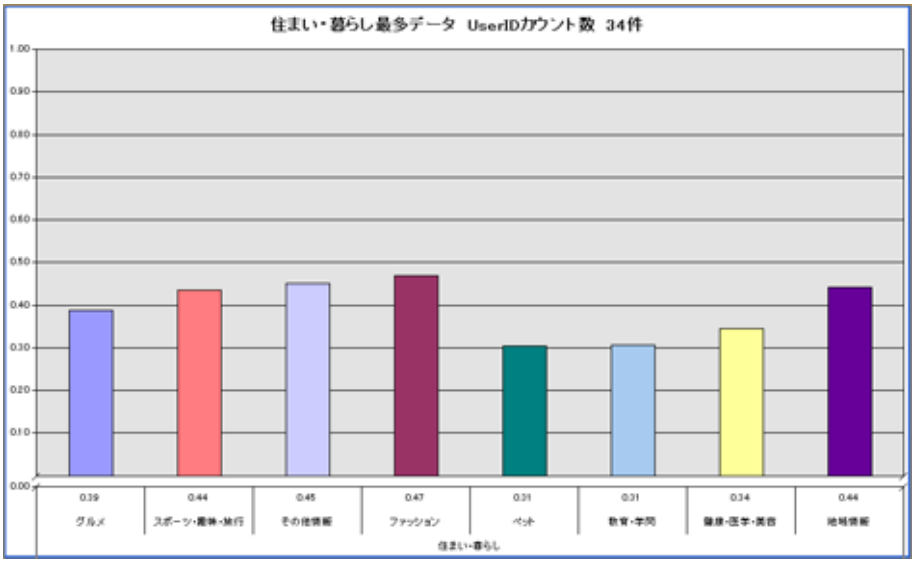

図 2.4: 住まい・暮らしに最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（住まい・暮らしは 1.0 の値）

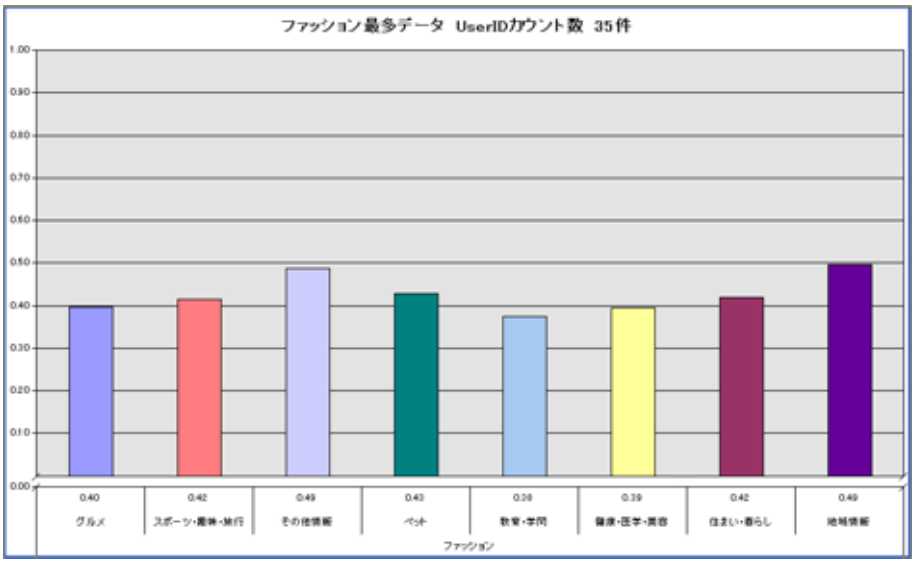

図 2.5: ファッションに最も興味を示したユーザのその他のカテゴリに対する興味の度合
（ファッションは 1.0 の値）

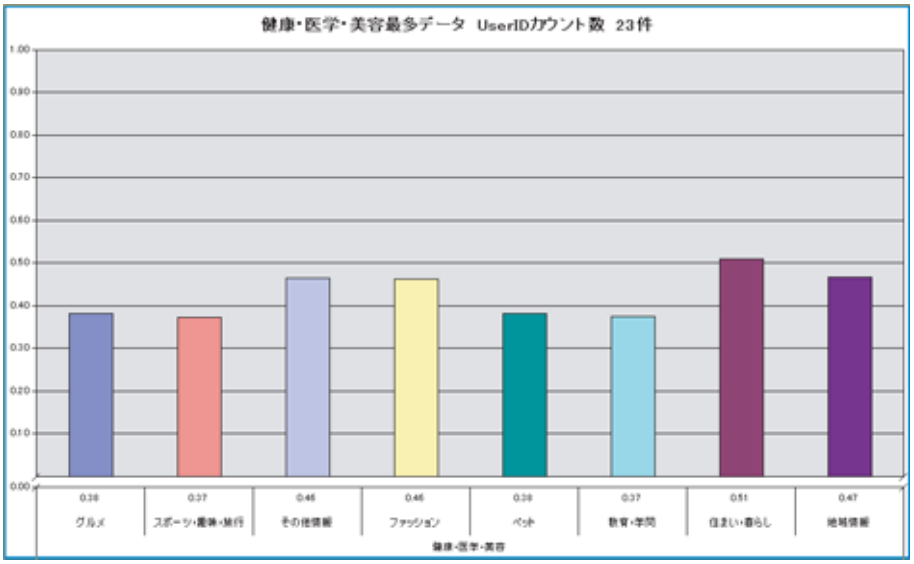

図 2.6: 健康・医学・美容に最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（健康・医学・美容は 1.0 の値）

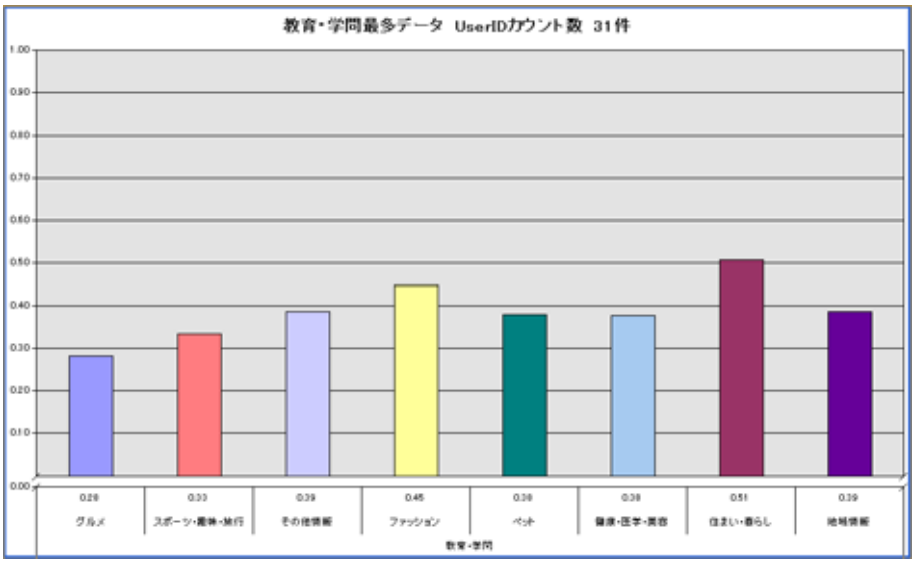

図 2.7: 教育・学問に最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（教育・学問は 1.0 の値）

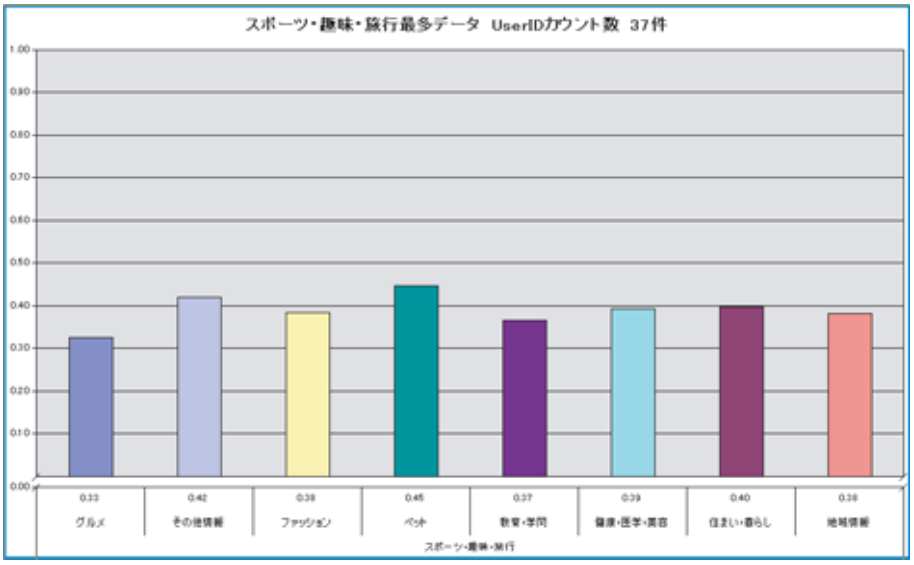

図 2.8: スポーツ・趣味・旅行に最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（スポーツ・趣味・旅行は 1.0 の値）

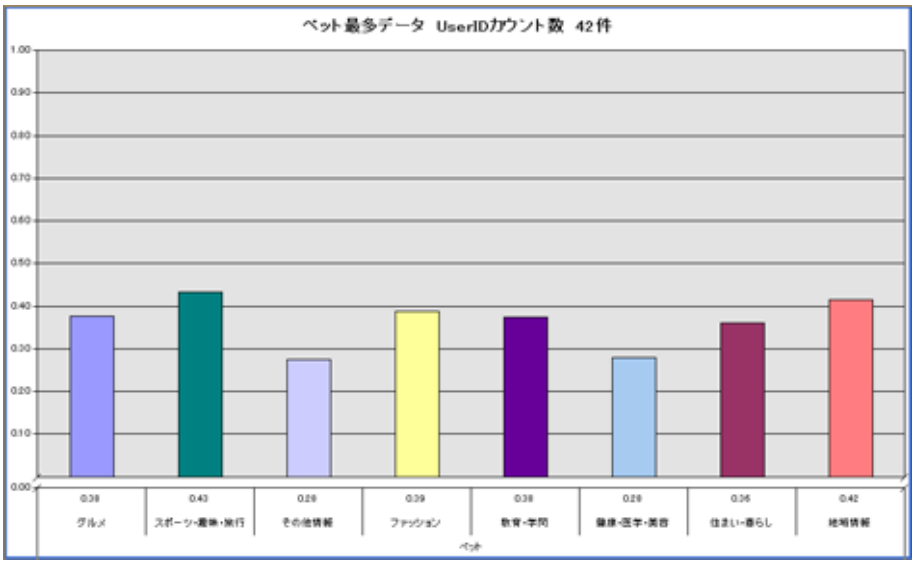

図 2.9: ペットに最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（ペットは 1.0 の値）

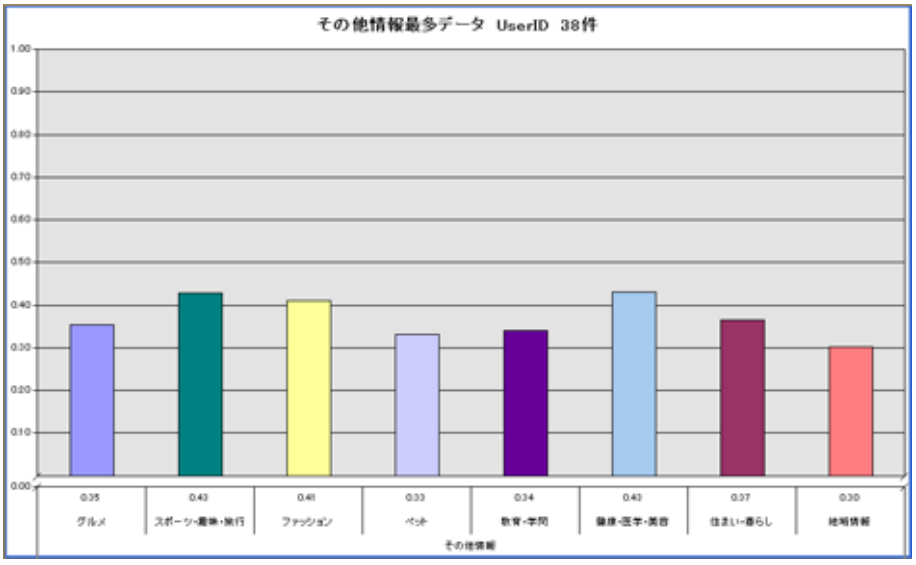

図 2.10: その他情報に最も興味を示したユーザのその他のカテゴリに対する興味の度合い
（その他は 1.0 の値）

確かに，カテゴリ間に相関があるのが確認された．ただ，興味深いのは，これらの相関は非対称であることである．たとえば,「グルメ」に興味のあるユーザは,「ファッション」にも興味がある．がしかし,「ファッション」に興味のあるユーザが「グルメ」に興味があるかというと，それほど高い興味を示していないということがわかる．ただ，対称な部分も存在している．たとえば,「ファッションに」興味のあるユーザは，地域情報に興味があるという結果が得られており，その逆に,「地域情報」に最も興味を示しているユーザも「ファッション」に高い関心を示している結果が得られている．これは,「ファッション」に興味のあるユーザが具体的な商品の購入場所を考えて，自身の生活圏内で自分の趣味に合うショップを探しているのではないかと推察される．

この結果から，

1. カテゴリ間の相関を考慮した学習を行えばより短時間で精度の良い学習が行える可能性がある．

2. カテゴリの遷移の履歴という点に着目して,「次に，どのカテゴリを閲覧する可能性が高いか？」ということを出力する新たなアルゴリズムの開発が必要である．

という 2 つのアプローチが必要であることが明確になった．１．は即座に反映され，２．は 2 年度目，3 年度目の主たる研究テーマとなり，マルコフモデルを用いた新アルゴリズムの研究の動機付けとなった．

2.5 基本配信モデル

基本配信モデルはユーザ全体の静的な趣向の解析に基づいた広告配信モデルである．基本配信モデルはどのカテゴリの Web ページに何回の訪問があったかを逐次記録する．具体的には，各カテゴリ i ごとに，そのカテゴリを訪問した回数 n_i を累積し，その累積度数分布に基づき配信する広告のカテゴリを決定する．各カテゴリの訪問回数を全てのカテゴリを訪問した全回数で割ると，各カテゴリに対する確率 (割合)p_i を求めることができる．

$$p_i = \frac{n_i}{\sum_{j=1}^{C} n_j} \tag{2.1}$$

ここで，C は広告カテゴリの総数を表す．基本配信モデルでは静的なユーザの趣向は捉えることができるが，細かくユーザのニーズに応えることまではできない．

また，確率 p_i に基づいて，確率的もしくは決定論的に配信する広告のカテゴリを決定することができる．

2.5.1 確率的な広告配信

各カテゴリの確率 p_i に基づいてランダムに広告配信のカテゴリを決定する．例えば，カテゴリ数 $C = 3$ で各カテゴリの確率が式 (2.2) の場合では．10 回の広告配信のうちカテゴリ 1 が 1 回，カテゴリ 2 が 3 回，カテゴリ 3 が 6 回の割合で各カテゴリの広告が配信されることになる．

$$
\begin{aligned}
p_1 &= 0.1 \\
p_2 &= 0.3 \\
p_3 &= 0.6
\end{aligned}
\tag{2.2}
$$

2.5.2 決定論的な広告配信

各カテゴリの確率 p_i の中で最大値を示すカテゴリの広告を配信する．

$$
c = \arg\max_i p_i \quad , (i = 1, 2, \ldots, C) \tag{2.3}
$$

例えば，カテゴリ数 $C = 3$ で式 (2.2) の場合では，10 回の広告配信のうち，全てカテゴリ 3 の広告を配信することになる．

2.6 マルコフモデルを用いた配信モデル

基本配信モデルを用いた配信モデルは静的な解析に基づく広告配信モデルであった．そのため，個々のユーザの趣向の変化に対しては対応することができなかった．しかし，マルコフモデルは動的なモデルであるので個々のユーザについてよりリアルタイムに趣向の変化に対応することができると予想される．

以降ではマルコフモデルについて述べ，その後マルコフモデルを用いた広告配信モデルによるシミュレーション結果について述べる．

2.6.1 マルコフモデル

ユーザが次に指定するカテゴリーは，それまでにユーザが指定したカテゴリー，すなわちユーザがそれまでに閲覧した情報に影響を受ける (図 2.11 参照)．

カテゴリ間の遷移は式 (2.4) に示される状態遷移確率行列 $\mathbf{P}$ に従う．

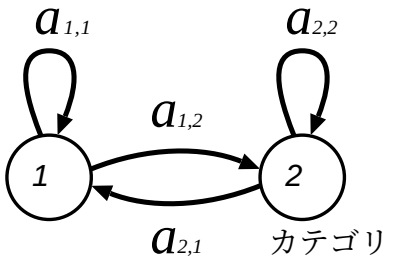

図 2.11: カテゴリ数 2 の単純マルコフモデル．

$$\mathbf{P} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{pmatrix} \tag{2.4}$$

$$a_{i,j} \geq 0, \quad i,j = 1,2,\cdots,m \tag{2.5}$$

$$\sum_{j=1}^{m} a_{i,j} = 1, \quad i = 1,2,\cdots,m \tag{2.6}$$

ここで，$a_{i,j}$ はカテゴリ i からカテゴリ j へ遷移する確率を表している．m は全カテゴリ数である．状態遷移確率 $a_{i,j}$ は確率の条件式 (2.5) と式 (2.6) を満たす．

例えば，Web ページのカテゴリを $1,2,3$ とし，遷移確率行列 $\mathbf{P}$ が式 (2.7) で表される場合を考える．この場合，カテゴリ 1 からカテゴリ 1 に遷移する確率が 0.3，カテゴリ 1 からカテゴリ 2 に遷移する確率が 0.6，カテゴリ 1 からカテゴリ 3 に遷移する確率が 0.1，カテゴリ 2 からカテゴリ 1 に遷移する確率が 0.5，$\cdots$ であることが示されている．

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0.5 & 0.2 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} \end{array} \tag{2.7}$$

マルコフモデル作成

マルコフモデルを作成するには，履歴データから状態遷移確率を推定しなければならない．状態遷移確率の推定には最尤推定法を用いる．

いま，時刻 0 から時刻 T までの状態遷移の履歴を $s_0, s_1, \cdots, s_T$ とする．s_t は時刻 t における状態である．m を全カテゴリ数，状態 $s_{t-1} = i$ から状態 $s_t = j$ への遷移の回数を $n_{i,j}$ とすると，履歴が $s_0, s_1, \cdots, s_T$ となる確率は，

$$a_{s_0,s_1} a_{s_1,s_2} \cdots a_{s_{T-1},s_T} = \prod_{i=1}^{m} \prod_{j=1}^{m} (a_{i,j})^{n_{i,j}} \tag{2.8}$$

である．ただし，初期状態は与えられているものとする．いま，遷移確率 $a_{i,j}$ $(i = 1,2,\cdots,m,\ j = 1,2,\cdots,m)$ は未知であるので，$n_{i,j}$ $(i = 1,2,\cdots,m,\ j = 1,2,\cdots,m)$ が与えられてもこの確率は計算できない．しかし，式 (2.8) を使えば，$a_{i,j}$ の推定値を求めることができる．以下にその方法を示す．

式 (2.8) を未知パラメータ $a_{i,j}$ $(i = 1,2,\cdots,m,\ j = 1,2,\cdots,m)$ の関数とみて $L(a_{1,1}, a_{1,2}, \cdots, a_{m,m})$ とする．この関数を尤度関数と呼び，遷移確率の値が $a_{i,j}$ であることの尤もらしさを表している．この尤度関数の値を最大にする $a_{i,j}$ の値は，未知遷移確率 $a_{i,j}$ の最尤推定値と呼ばれる．

最尤推定値を具体的に求めてみる．L が最大の場合には $\log L$ も最大になるので，

$$\log L = \sum_{i=1}^{m}\sum_{j=1}^{m} n_{i,j} \log a_{i,j} \tag{2.9}$$

を最大にする $a_{i,j}$ を求めればよい．ただし，この場合，$a_{i,j}$ は確率の条件式 (2.5) と式 (2.6) を満たしていなければならない．これは制約条件付き最大値問題であり，ラグランジュの未定乗数法を用いて解くことができる．ラグランジュ関数 F は，$\lambda_i (i = 1, 2, \cdots, m)$ をラグランジュ乗数とすると，

$$\begin{aligned} F &= \log L + \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{m} a_{i,j} - 1 \right) \\ &= \sum_{i=1}^{m} \left[\sum_{j=1}^{m} n_{i,j} \log a_{i,j} + \lambda_i \left(\sum_{j=1}^{m} a_{i,j} - 1 \right) \right] \end{aligned} \tag{2.10}$$

と与えられる．ただし，この段階では条件式 (2.5) は考慮しないでおく．$\log L$ を最大にする $a_{i,j}$ は，F を $a_{i,j}$ で偏微分し，それを 0 とおいた連立方程式を解いて求められる．さて，F を $a_{i,j}$ で偏微分すると，

$$\frac{\partial F}{\partial a_{i,j}} = \frac{n_{i,j}}{a_{i,j}} + \lambda_i \tag{2.11}$$

となるので，この右辺を 0 とおくと，

$$a_{i,j} = -\frac{n_{i,j}}{\lambda_i} \tag{2.12}$$

が得られる．乗数 λ_i は条件式 (2.6) に式 (2.12) を代入すれば，

$$1 = \sum_{j=1}^{m} a_{i,j} = -\frac{1}{\lambda_i} \sum_{j=1}^{m} n_{i,j} = -\frac{n_i}{\lambda_i} \tag{2.13}$$

となり，$\lambda_i = -n_i$ と求められる．ここで，$n_i = \sum_{j=1}^{m} n_{i,j}$ である．これと，式 (2.12) から推定値

$$a_{i,j} = \frac{n_{i,j}}{n_i} \tag{2.14}$$

が得られる．この解は条件式 (2.5) も同時に満たしているので，求める最尤推定値である．この推定値は，状態 i にいた回数に対する次の時刻に状態 j へ遷移した回数の割合を表しており，我々の直観ともよく一致する．

例として，状態数 $5, 10, 15$ の確率過程に対し，遷移確率の推定を行い，用いた履歴データ数に対する遷移確率の推定誤差を図 2.12 に示す．図より，推定誤差が安定するには，$400 \sim 800$ 個のデータ数が必要なことがわかる．

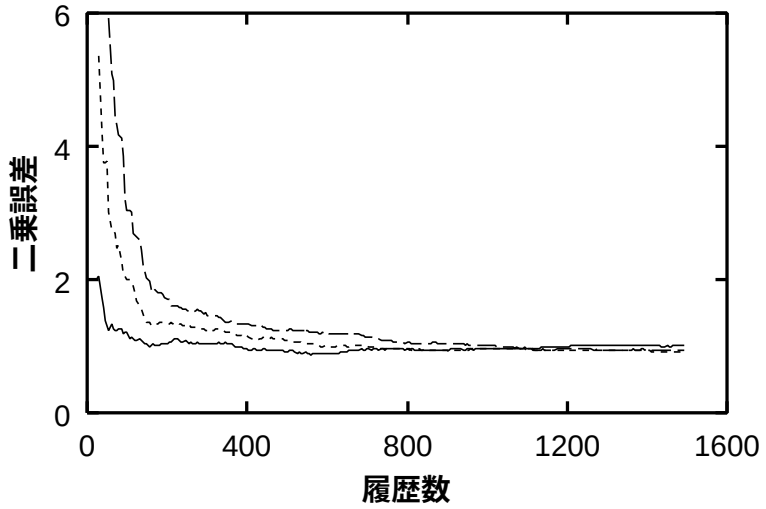

図 2.12: 用いた履歴データ数に対する遷移確率の推定誤差.
—：5 状態マルコフモデル, ...：10 状態マルコフモデル，- - -：15 状態マルコフモデル.

2.6.2 マルコフモデルを用いた広告配信方法

具体的な広告配信方法としては，これまで同様に確率的な広告配信方法と決定論的な広告配信方法の 2 通りの方法が考えられる.

まず決定論的な広告配信方法としては，遷移確率が最も高いカテゴリの広告を配信する方法である．例えば，式 (2.7) の遷移確率行列が与えられたとする．現在，ユーザがカテゴリ 2 を訪問していれば，次に遷移するカテゴリの中で，最も高い遷移確率 (0.5) を持つカテゴリ 1 の広告を配信する.

一方確率的な広告配信方法は，遷移確率に基づいて乱数を発生させ，次に遷移すると予想されるカテゴリの広告を配信する方法である．例えば，式 (2.7) の遷移確率行列が与えられたとする．この方法では，現在，ユーザがカテゴリ 2 を訪問していれば，0.5 の確率でカテゴリ 1，0.2 の確率でカテゴリ 2，0.3 の確率でカテゴリ 3 の広告が配信される．即ち，前者の配信方法と異なり，カテゴリ 2 やカテゴリ 3 の広告を配信する可能性がある.

2.6.3 計算機シミュレーション

まず実際の遷移履歴データを基に遷移確率を計算し，得られた遷移確率を用いて前述の 2 通りの配信方法で次時刻に配信する広告を予測した．なお，初期状態のカテゴリには履歴データ (真値) を与えた．配信した広告のカテゴリが履歴データと一致した場合は 1 とし，一致しなかった場合は 0 として的中率を求めた．時系列で的中率を求めた結果を図 2.13〜図 2.15 に示す．この結果は，ユーザ番号が 288 のデータを用いたときの結果である.

また，予測したカテゴリの遷移系列と実際のカテゴリの遷移系列の比較を図 2.16〜図 2.18 に示す．それぞれ 3 つの配信方法で配信した広告のカテゴリと実際に遷移したカテゴリが一致した数 (正解数) を表 2.1 にまとめた．なお，ユーザ番号が 328, 354, 873 のデータに対して行ったシミュレーション結果も合わせてまとめた.

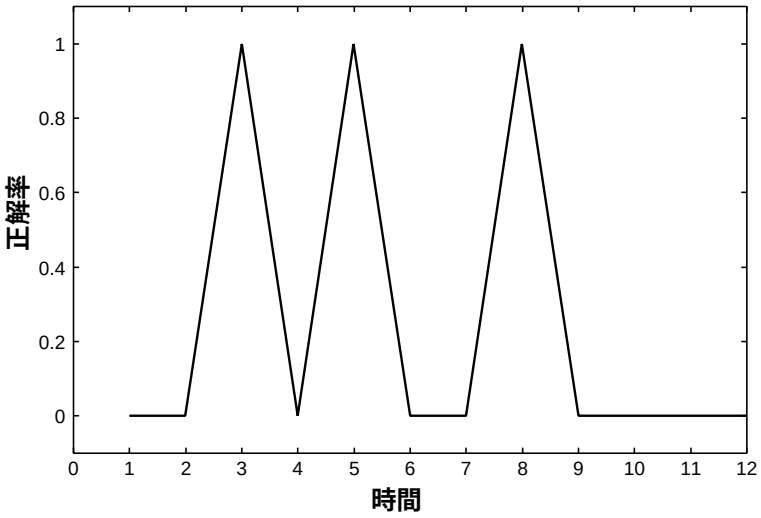

図 2.13: 基本配信モデルの的中率.

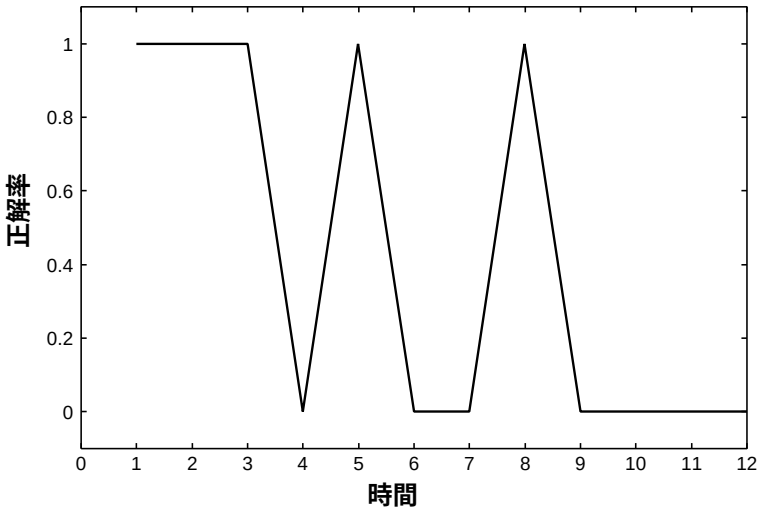

図 2.14: マルコフモデルを用いた配信モデル (遷移確率の最も高いカテゴリの広告を配信) の的中率.

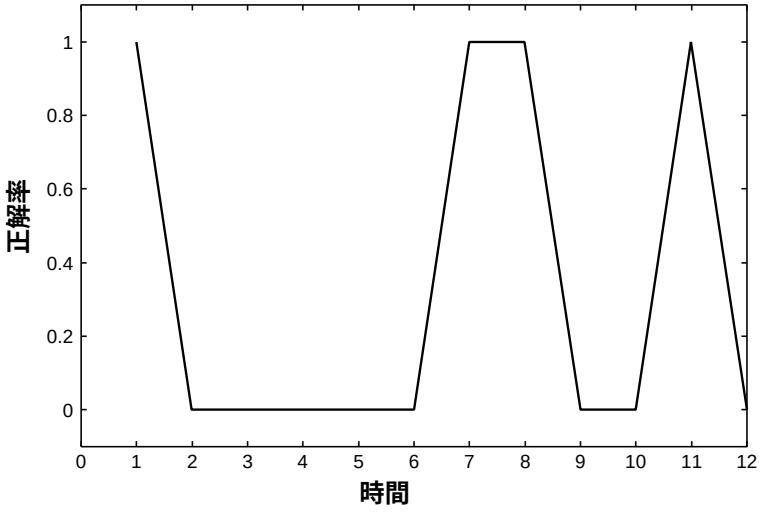

図 2.15: マルコフモデルを用いた配信モデル (遷移確率に基づいて広告を確率的に配信) の的中率.

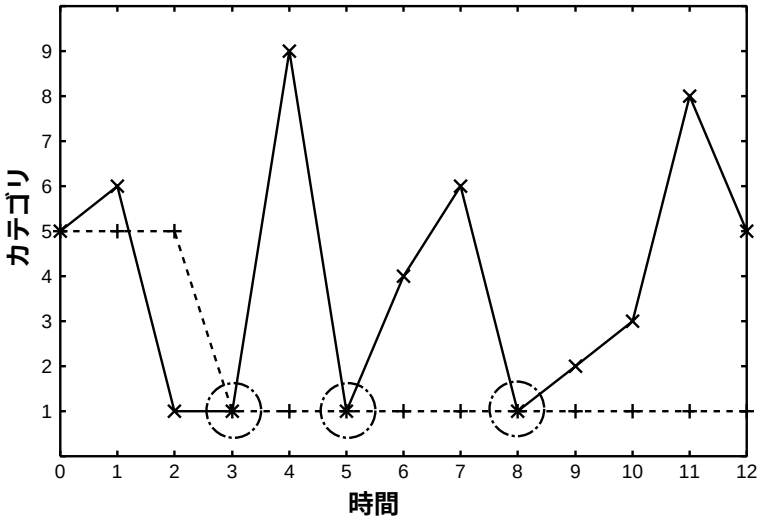

図 2.16: 基本配信モデルによるカテゴリの遷移.
___：実際のカテゴリの遷移, ---：基本配信モデルにより配信した広告のカテゴリの遷移,
◌：モデルにより配信した広告のカテゴリと実際に訪問したカテゴリが一致した部分.

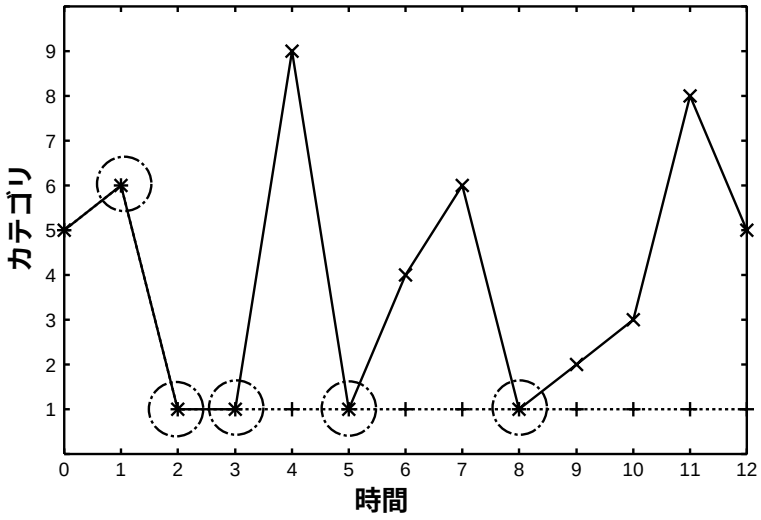

図 2.17: マルコフモデルを用いた配信モデル (遷移確率が最も高いカテゴリの広告を配信) によるカテゴリの遷移.
___ : 実際のカテゴリの遷移, - - - : マルコフモデルを用いた配信モデル (遷移確率の最も高いカテゴリの広告を配信) の遷移, ◌ : モデルにより配信した広告のカテゴリと実際に訪問したカテゴリが一致した部分.

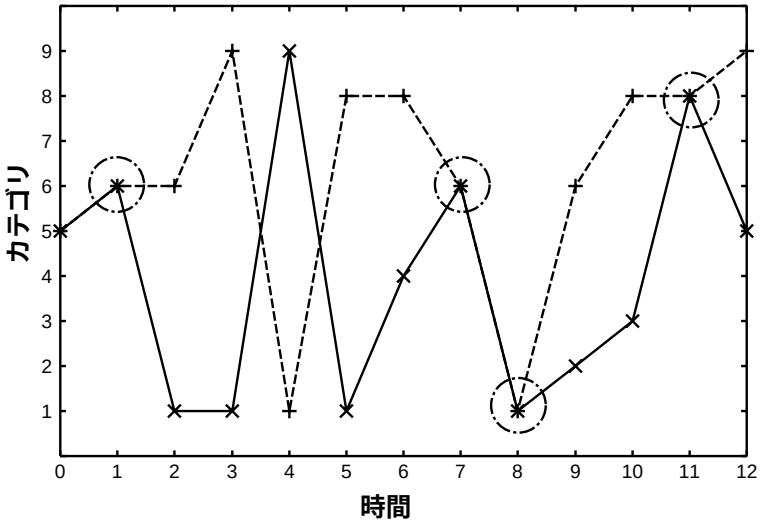

図 2.18: マルコフモデルを用いた配信モデル (遷移確率に基づいて広告を確率的に配信) によるカテゴリの遷移.
___ : 実際のカテゴリの遷移, - - - : マルコフモデルを用いた配信モデル (遷移確率に基づいて広告を確率的に配信) の遷移, ◌ : モデルにより配信した広告のカテゴリと実際に訪問したカテゴリが一致した部分.

表 2.1: それぞれの予測方法での正解数.

ユーザ番号 (履歴数)	288 (13)	328 (38)	354 (69)	873 (94)
基本配信モデル	3	2	7	12
マルコフモデルを用いた配信モデル (遷移確率が最も高いカテゴリの広告を配信)	5	5	12	17
マルコフモデルを用いた配信モデル (遷移確率に基づいて広告を確率的に配信)	4	4	9	15

この結果より，基本配信モデルよりもマルコフモデルを用いた配信モデルの方が，予測した広告カテゴリと実際に遷移したカテゴリが一致した数が多く，予測が精度良く行われていることがわかる.

2.7 重みつきマルコフモデルを用いた配信モデル

前節でマルコフモデルを用いた配信モデルについて述べたが，マルコフモデルはモデル作成に多数のデータを必要とすることが分かった．そこで，本章ではモデル作成に費す時間を軽減する方法として重みつきマルコフモデルを用いた広告配信システムを提案する.

重みつきマルコフモデルはあらかじめ用意しておいた数個のマルコフモデルを線形加算することであらたなマルコフモデルを作成する手法である.

重みつきマルコフモデルのシステムは２つに大別でき，１つは代表モデルと呼ばれる数個のマルコフモデルを作成する部分である．もう１つは，用意した代表マルコフモデルを用いて新たなマルコフモデルを作成し，次に遷移する状態の予測を行う部分である．以降で重みつきマルコフモデルのアルゴリズムについて述べる.

2.7.1 代表マルコフモデルの作成

Web 広告配信において，マルコフモデルは各ユーザの趣向を反映したモデルと考えられ，一方，代表マルコフモデルは多くのユーザの趣向を反映したモデルといえる.

代表マルコフモデルの作成にはまず，多数の履歴データ (ユーザのカテゴリ間の遷移についての情報) からユーザごとにマルコフモデルを作成する．このとき履歴データをユーザのログインからログアウト (１セッション) の間で区切り，区切られた区間で一つのマルコフモデルを作成する．そして，作成したモデルの遷移確率行列 $\mathbf{P}$ をベクトル量子化する．ベクトル量子化後に得られるコードベクトルが代表モデルの状態遷移確率行列を表す．代表マルコフモデル作成の手順を図 2.19 に示す.

ここで，ログインからログアウトのセッションごとに履歴データを区切る基準には日付を用いた．これは履歴データにセッション開始 (もしくは終了) を判断する情報が含まれていないためである．実際の状況を考えると，ユーザが日付をまたいで継続して Web ページを閲覧することは考えにくい．このような理由から，履歴データの日付が同じであるデータを一回のセッション中でのデータとした．ログイン (セッション開始) からログアウト (セッ

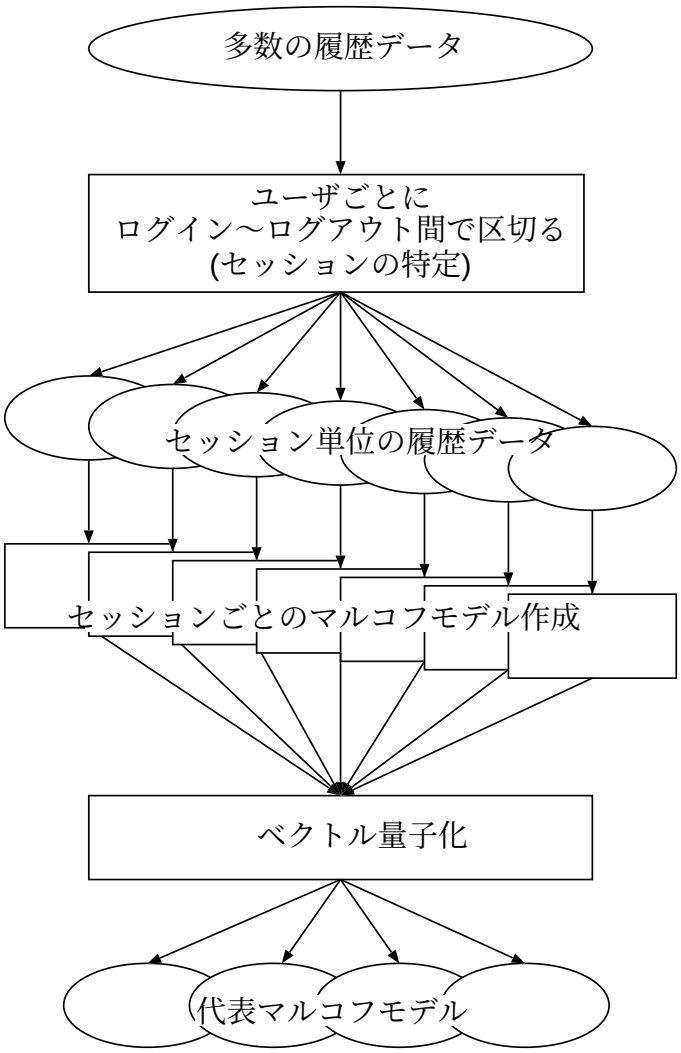

図 2.19: 代表マルコフモデル作成手順

ション終了) までの系列の長さ L の分布 ($1 \leq L \leq 50$) を図 2.20 に示す.

また，ベクトル量子化後の状態遷移確率行列 (コードベクトル) の各行において確率の総和が 1 になるように正規化する.

ベクトル量子化

ベクトル量子化とは，特徴ベクトル空間をカテゴリを代表するいくつかの領域に分割することをいう (図 2.21)．Web 広告配信に関していえば，ユーザの趣向を表現した多数のマルコフモデルを，似た趣向をもつモデル同士でグルーピングすることである．各カテゴリを代表するコードベクトルは，多数のユーザモデルに見られる代表的な趣向を表現することになる.

代表モデル作成に用いるベクトル量子化の手法には特に条件があるわけではないが，で

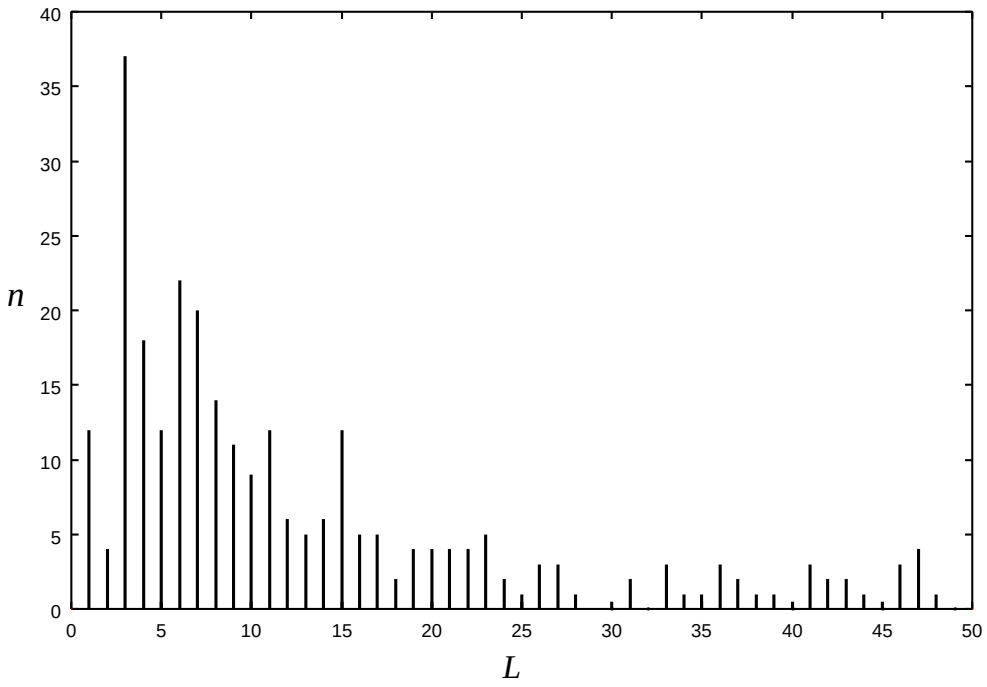

図 2.20: 系列の長さ L の分布 $(1 \leq L \leq 50)$.

きるだけ量子化誤差の抑えられる手法を選ぶのが望ましい．本研究では kMER[1](付録参照)を用いた.

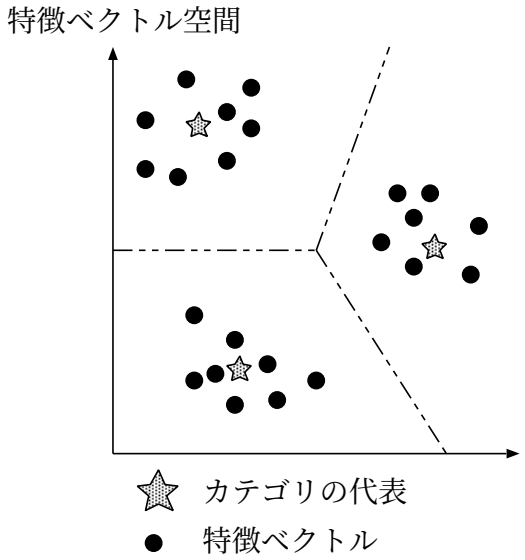

図 2.21: ベクトル量子化の概念図.

[1]参考文献：「自己組織化マップ－理論・設計・応用」マーク M. ヴァンフッレ 著，徳高平蔵・藤村喜久郎 監訳

2.7.2 ユーザーモデルの推定

Web 広告配信において，ユーザのカテゴリ間の遷移を予測するためにはユーザの趣向を反映したユーザモデルが必要となる．ユーザモデルは，少ない履歴データから作成された近似マルコフモデルと各代表マルコフモデルとの類似度を計算し，その類似度をもとに各代表マルコフモデルの線形加算により推定する．

類似度の計算

さて，少ない履歴から求めた近似マルコフモデルの状態遷移確率行列を $\mathbf{A}$，代表マルコフモデル k の状態遷移確率行列を $\mathbf{R}_k$ とすると，$\mathbf{A}$ の $\mathbf{R}_k$ に対する類似度 w_k を以下の式で定義する．

$$w_k = \exp\left(-\frac{||\mathbf{R}_k - \mathbf{A}||^2}{d^2}\right) \tag{2.15}$$

$$||\mathbf{A} - \mathbf{B}|| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(a_{i,j} - b_{i,j})^2} \tag{2.16}$$

ここで，$||\cdot||$ はノルム，d はガウス関数の標準偏差を表す．d には適当な値を用いる (0.2〜0.4)．この類似度をもとに，ユーザモデルの状態遷移確率行列 $\hat{\mathbf{P}}$ を以下のように求める．

$$\hat{\mathbf{P}} = \sum_{k=1}^{n}\left(\frac{w_k}{\sum_{l=1}^{n} w_l}\mathbf{R}_k\right) \tag{2.17}$$

ここで，n は代表マルコフモデルの個数である．ユーザモデルの推定手順を図 2.22 に示す．

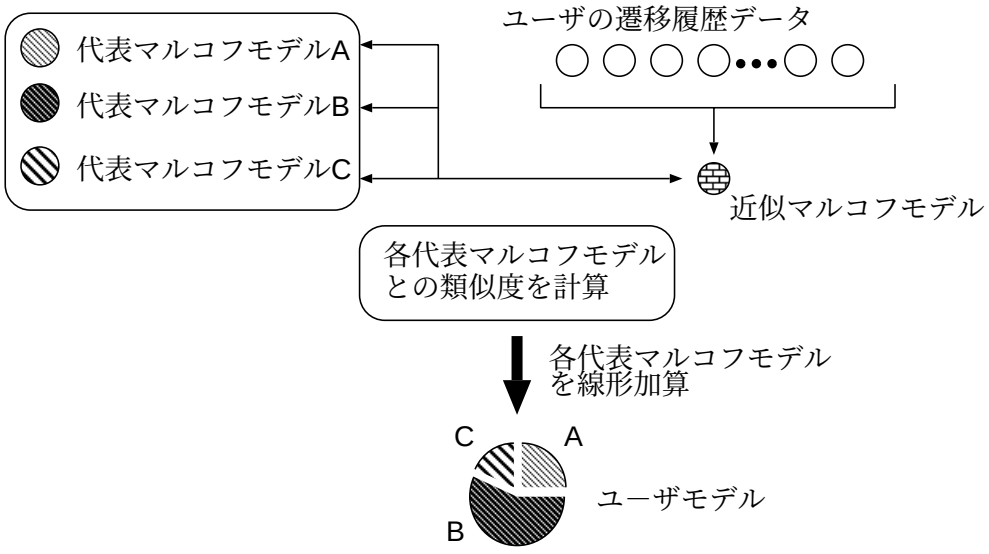

図 2.22: 代表マルコフモデルからユーザモデルの推定

2.7.3 ユーザモデルを用いた次状態の予測

推定されたユーザモデルを用いて，ユーザが次の時刻にどのカテゴリに遷移するかを予測する．予測方法は，現在ユーザが閲覧している Web ページのカテゴリを c_n，ユーザモデルの状態遷移確率を $a_{i,j}$ とすると，予測される次の遷移カテゴリ $\hat{c}_{n+1}$ を以下の式で求める．

$$\hat{c}_{n+1} = \arg\max_j a_{c_n,j} \tag{2.18}$$

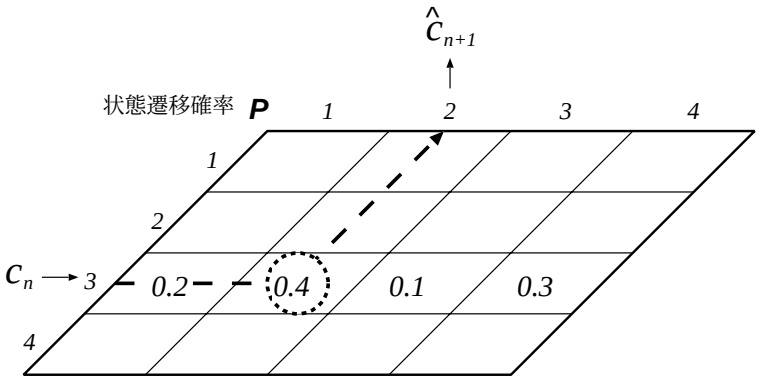

図 2.23: 状態遷移確率に基づいた予測

2.7.4 Web 広告配信への応用

さて，重みつきマルコフモデルを Web 広告配信に応用することを考える．Web ページ上にはさまざまな広告が存在し，ユーザはいろいろな広告の中で自分の趣向に合った広告をクリックし，情報を得ようとする．したがって，Web 広告配信システムでは，ユーザの趣向に合った広告を配信することが重要である．さらに，ユーザの趣向は時間と共に変化していると考えられるので，その時々で各ユーザに応じた広告を配信しなければならない．

具体的に，提案手法を応用した Web 広告配信の方法を図 2.24 に示す．まず，ユーザの Web 上の遷移履歴を提案システムに入力し，提案システムにより次にユーザが遷移するカテゴリを予測する．広告配信システムは，予測された次の遷移カテゴリに属している広告をユーザに配信する．

2.7.5 評価尺度

的中率

一般に物事の予測の評価は，予測結果が実際の結果と一致したかで行われる．的中率は，ユーザが時刻 k において閲覧した Web ページのカテゴリを $C(k)$ とし，代表マルコフモデルが予測し，その結果として配信した広告のカテゴリを $\hat{C}(k)$ とすると以下の式で表される．

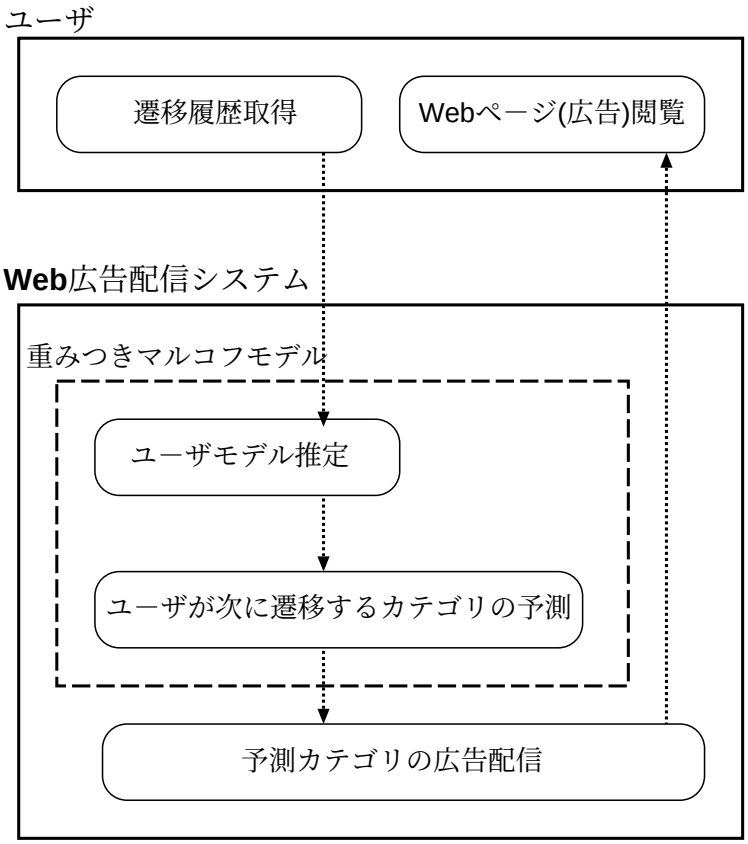

図 2.24: 重みつきマルコフモデルを応用した Web 広告配信手順

$$R = \frac{\sum_{k=1}^{N} \delta(C(k), \hat{C}(k))}{N} \tag{2.19}$$

$$\delta(C(k), \hat{C}(k)) = \begin{cases} 1 & (C(k) = \hat{C}(k)), \\ 0 & (\text{otherwise}). \end{cases} \tag{2.20}$$

ここで，N は広告を配信した最終時刻である．

的中率では，予測結果と実際の結果が完全に一致した場合でなければ評価は上がらない．しかし，ユーザの趣向が完全に一つのカテゴリに絞られることは考えにくい．したがって，推定したユーザモデルの状態遷移確率がどれだけユーザの趣向を良く捉えているかで評価を下すことが必要になってくる．

正解率

マルコフモデルの状態遷移確率行列 $\mathbf{P}$ はユーザの趣向を反映しており，状態遷移確率 $a_{i,j}$ が i 行の中で最高値でなくとも，ユーザはカテゴリ j に対して興味があるといえる．したがって，的中率による評価は必ずしも代表マルコフモデルに適した評価ではない．そこで，

新たに評価尺度として正解率を導入する．正解率 D を，カテゴリ総数 m としたとき，以下の式で定義する．

$$D = \frac{\sum_{k=1}^{N} \acute{\delta}(C(k), \hat{C}(k))}{N} \tag{2.21}$$

$$\acute{\delta}(C(k), \hat{C}(k)) = \frac{a_{C(k-1),C(k)}}{a_{C(k-1),\hat{C}(k)}} \tag{2.22}$$

ここで，N は広告を配信した最終時刻である．

この評価尺度では，予測結果が合っている (1)，合っていない (0) の 2 値ではなく，どの程度合っているかを評価することができる．

2.7.6 計算機シミュレーション

他配信モデルとの比較

重みつきマルコフモデルと他の広告配信モデルとで的中率の比較を行った．比較モデルは基本配信モデルとランダム配信モデルである．また，重みつきマルコフモデルにおいて代表マルコフモデル個数は 5 個，また式 (2.15) の $d = 0.2$ とした．シミュレーションには，2002 年 12 月 4 日から 2003 年 11 月 25 日までの期間に得られたもので，履歴数が 101 から 999 までのデータを用いた．各モデルの的中率を表 2.2 に示す．

表 2.2: モデルごとの的中率の比較.

基本配信モデル	ランダム配信モデル	重みつきマルコフモデル
0.110068765	0.112769388	0.112267034

表 2.2 を見る限り，どのモデルでも的中率はあまり変わらない結果となった．しかし，重みつきマルコフモデルにおいて代表マルコフモデル数が少なかったことが原因で的中率が変わらなかったと考えられる．次節では代表マルコフモデル数の変化に対する的中率の変化を調べる．

代表マルコフモデル数に対する的中率の変化

重みつきマルコフモデルの予測精度をみるために，代表マルコフモデル数を 3 から 84 個まで変化させた場合の予測の的中率を調べた．使用したデータは，2002 年 12 月 4 日から 2003 年 11 月 25 日までの Web 遷移履歴 113333 個である．このデータを用いて代表モデルを作成し，作成した代表モデルを用いて各ユーザごとにカテゴリ間の遷移を予測した．その結果を図 2.25 に示す.

図 2.25 から，代表マルコフモデル数を増やすことで予測の的中率は上がるが，0.3 程度の的中率が限界であると予想される．代表マルコフモデル数を増やすと，それだけシステムの負荷は大きくなるのでシステムの耐え得る範囲で代表マルコフモデル数を決定する必要がある．

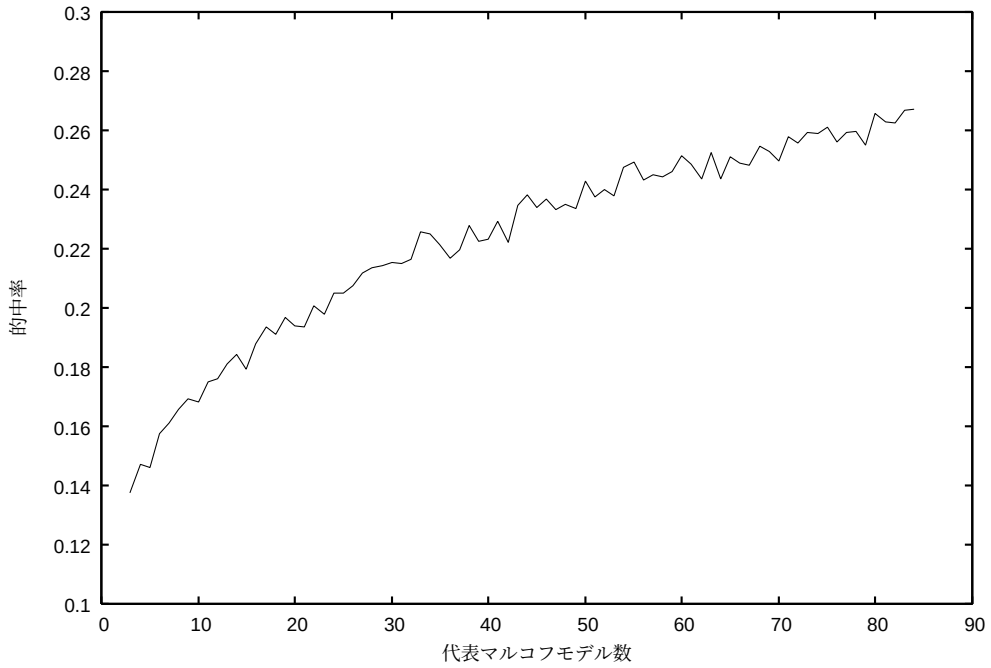

図 2.25: 代表マルコフモデル数に対する的中率の変化.

代表マルコフモデルの更新期間

代表マルコフモデルは多数のユーザモデルをもとに作成することを前述したが，これは代表マルコフモデルが大衆の流行を反映していることに他ならない．しかしながら，流行は時と共に変化すると考えられ，それに合わせて代表マルコフモデルも作り替えることが望ましい.

そこで，7 つの区間で代表マルコフモデル (10 個) を作成し，各区間で作成した代表マルコフモデルを用いて検証用データに対する予測を行った．予測性能は的中率と正解率で評価した．この実験では 2002 年 12 月から 2003 年 11 月までの遷移履歴データを用いた．代表マルコフモデルを作成した 7 つの区間 A,B,C,D,E,F,G と，検証用データの区間 V を表 2.3 に示す．A から G の各区間における遷移履歴データ数は 30,000 件である．各区間で作成した代表マルコフモデルを用いて区間 V を予測した的中率と正解率を図 2.26 に示す. 図 2.26 より本システムが 70-80 %の正解率でユーザの遷移を予測しているのがわかる.

表 2.3: 代表マルコフモデルを作成した７通りの区間，および検証用データの区間.

区間	開始	終了
A	2002 年 12 月 4 日	2003 年 4 月 21 日
B	2003 年 2 月 3 日	2003 年 5 月 24 日
C	2003 年 3 月 13 日	2003 年 7 月 20 日
D	2003 年 4 月 21 日	2003 年 8 月 27 日
E	2003 年 5 月 24 日	2003 年 9 月 25 日
F	2003 年 7 月 20 日	2003 年 10 月 10 日
G	2003 年 8 月 27 日	2003 年 11 月 5 日
V	2003 年 11 月 5 日	2003 年 11 月 21 日

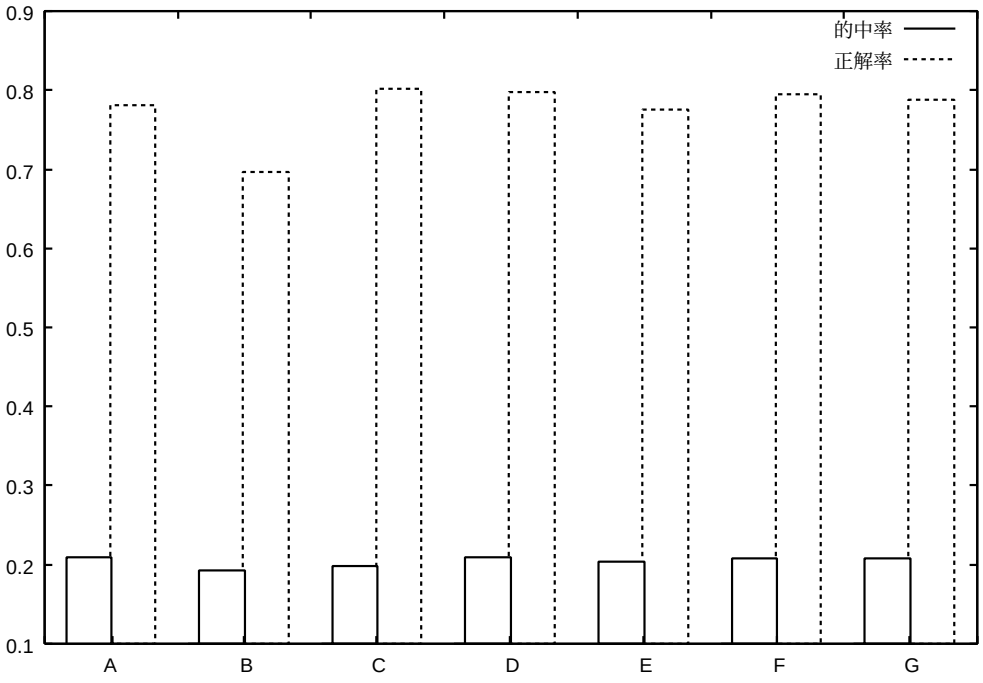

図 2.26: 各区間の代表マルコフモデルを用いた区間 V の予測の的中率および正解度.

2.8 広告配信モデルの評価方法

2.8.1 静的な評価方法

重みつきマルコフモデルは動的なモデルであるが，基本配信モデルは静的なモデルである．静的な評価方法として,「ユーザの定常確率分布」と「3 つのモデルによって配信された広告のカテゴリの割合」との比較を行った．ここで 3 つのモデルとは，基本配信モデル，ランダム配信モデル，重みつきマルコフモデルのことである．分布の比較を図 2.27 に示す.

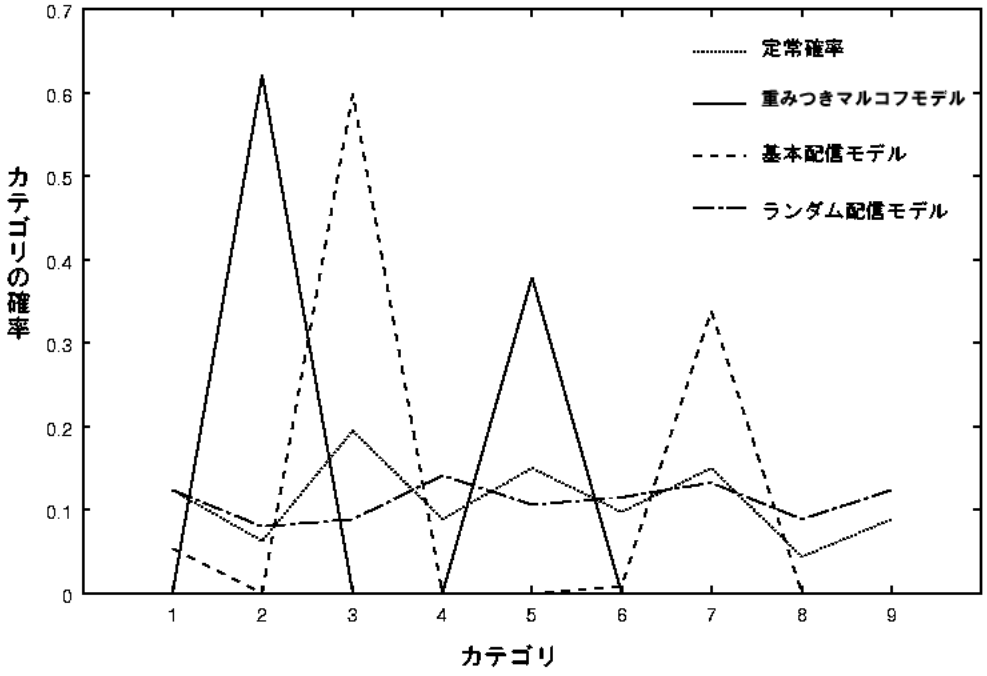

図 2.27: 各モデルについて配信広告のカテゴリごとの分布.

他の配信モデルと比べて，基本配信モデルが最も定常確率分布の特徴を捉えていることが分かる.

2.8.2 動的な評価方法

これまでの広告配信モデルの評価方法は，ユーザの k 時刻に遷移したカテゴリ $C_u(k)$ とモデルの配信した広告のカテゴリ $C_d(k)$ が一致している割合 (的中率) R を求め，モデルの比較を行っていた.

しかし，ユーザはログイン直後に訪問したカテゴリに興味が高いと考えられるので，評価についてもログイン直後の数回を重視するべきである．この評価方法については,「ログイン直後 n 時刻」というように評価の対象とする範囲を完全に限定する方法と，ログインからの経過時刻によって重みを付ける方法の 2 通りで行った．時間経過に伴う重み w_k には以下の式を用いた.

$$w_k = \exp(-k) \tag{2.23}$$

ここで，k はログインからの経過時刻を表す．また，式 (2.23) を図 2.28 に示す．

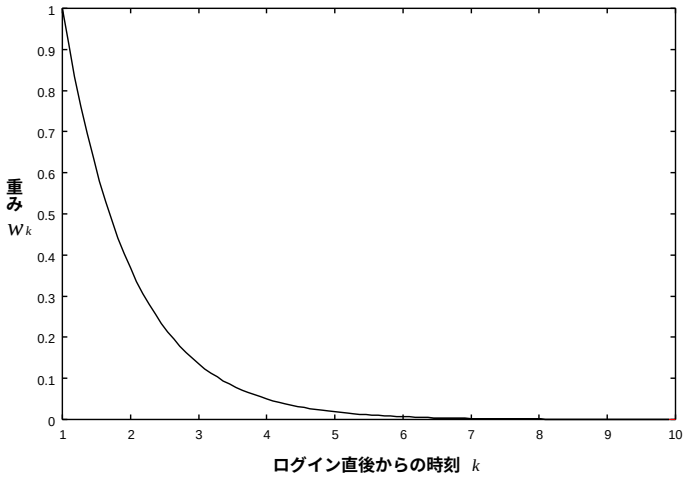

図 2.28: $exp(-k)$ の変化.

重みつき評価値 R_w を以下のように定義する．

$$R_w = \frac{\sum_{k=1}^{M} w_k \delta(C_u(k), C_d(k))}{\sum_{k=1}^{M} w_k} \tag{2.24}$$

また，ログイン直後 n 時刻までの的中率 R_n を以下の式で定義する．

$$R_n = \frac{\sum_{k=1}^{n} \delta(C_u(k), C_d(k))}{n} \tag{2.25}$$

それぞれの評価値での各モデルの比較を表 2.4 に示す．

表 2.4: 評価値の比較.

	R	R_w	R_2	R_3	R_4
ランダム配信モデル	0.1086	0.0985	0.1002	0.0992	0.1076
基本配信モデル	0.1113	0.1222	0.1304	0.1240	0.1241
重みつきマルコフモデル	0.1413	0.1394	0.1307	0.1410	0.1597

2.9 システム開発

2.9.1 実装

開発したモデルをテストサーバに実装した．サーバに実装するに当たっては，アルゴリズムを一つの大きなクラスの集合体として扱うこととし，今回開発したアルゴリズムを適

切にシステムに実装することが出来た．

また，その一方で，アルゴリズムそのものについても必要な機能を洗い出し，細かなクラスとして構成することにした．更に，新アルゴリズムを汎用のサーバ機で稼動させるため，可能なかぎりスムーズに稼動するように調整を行った．

また，基本配信モデルとマルコフモデルのそれぞれの優位性を確かめるために，双方のアルゴリズムに対してフィールドテストを行った．そのなかで，それぞれのアルゴリズムの入れ替えをスムーズに行う必要が生じてきた．そこで，入れ替えをスムーズに行えるように，システムの構成を調整した．これは，前年度まで，システムサイドにおいて，機能ルーチンをクラス化していたために可能になったことである．

2.9.2 システム構成

情報配信システムのネットワーク構成

図 2.29 は，情報配信システムのネットワーク構成を示す図である．この図に示すように今回の情報配信システムは，インターネットなどの通信ネットワークに接続されたウェブサイト，配信サーバ，およびユーザが利用するブラウザの含まれたパーソナルコンピュータから構成されることになる．ここでいうウェブサイトは公知のウェブサイトを指す，たとえば Yahoo 等のようなポータルサイトがそれにあたる．また，配信サーバは，今回の研究で構築されたものをさす．

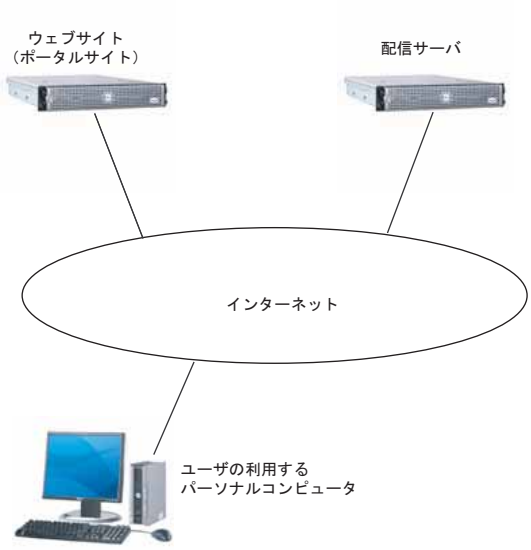

図 2.29: 情報配信システムのネットワーク構成

情報画面

図 2.30 は，ユーザが利用するパーソナルコンピュータに表示される情報画面の例を示している．この情報画面はブラウザを介して表示される．この図に示すように，情報画面に

は，ウェブサイトから送られてくる情報が表示され，かつ配信サーバから送られてくる情報が表示されることになる．

配信サーバから送られてくる情報として，今回は，広告とコンテンツ（ポータルサイト内で分類されているカテゴリ名称の表示順）の両方に対応できるようなものを考える．

配信された広告をクリックした場合，その広告に関連付けられたウェブサイトへ画面が遷移することが好ましい．この仕組みを実現するに当たっては，配信する広告を FLASH で作成し，その FLASH にクリック時の遷移先情報を持たせることで対応することにした．

図 2.30: ユーザのパーソナルコンピュータに表示される情報画面の例

シーケンス図

図 2.31 は，今回のシステムにおけるユーザのパーソナルコンピュータ，ウェブサイト（ポータルサイト）および配信サーバの通信手順を示すシーケンス図である．この図に示すように，ユーザのリクエストがすべての基点となる．まず，ユーザが公知のウェブサイト（ポータルサイト）にアクセスする．すると，ウェブサイトはそのリクエストに応じてウェブサイト内の HTML コンテンツをユーザのパーソナルコンピュータに渡す．HTML コンテンツを渡されたユーザのパーソナルコンピュータはブラウザの機能を使って，HTML を解釈し，それを画面に表示する．この際，解釈される HTML 内に，配信サーバへのリクエストを記述しておくことで，ブラウザは，配信サーバへのアクセスを行う．アクセスされた配信サーバは，ユーザのパーソナルコンピュータのクッキーを参照し，そのユーザの嗜好を検出する．検出結果に応じたコンテンツ（広告およびサイトないコンテンツのカテゴリ表示の適切な順序表示）をユーザのパーソナルコンピュータに配信する．最後に，ユーザのパーソナルコンピュータは，配信サーバから受け取ったコンテンツを、ブラウザを介して画面に表示する．

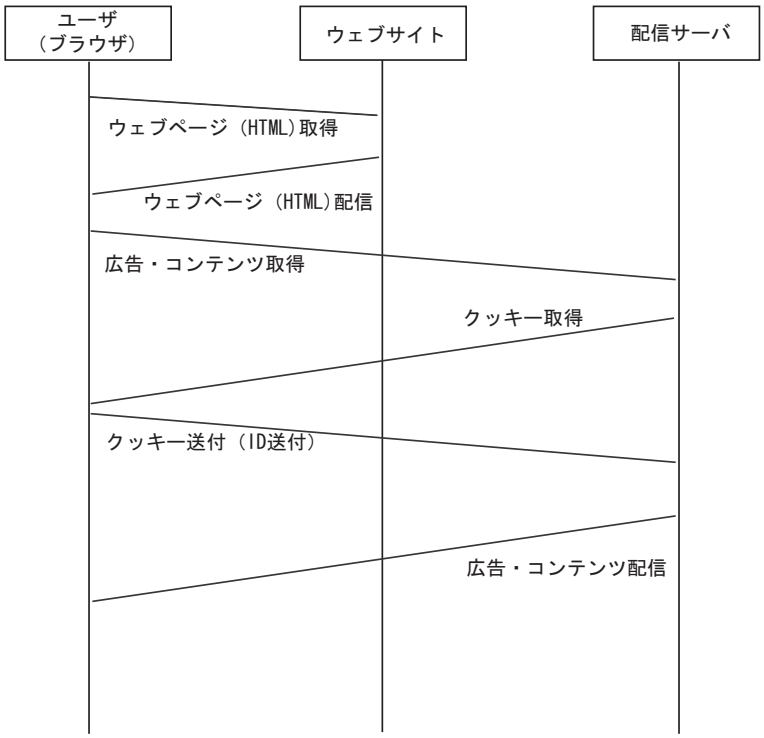

図 2.31: シーケンス図

これら一連の動作が，ユーザが対象となるウェブサイトでページの遷移を行うたびに実施されることになる．また，このことは，配信サーバに常にあるユーザがどのページを閲覧しているかを伝えることにもなる．配信サーバは，この情報もつど利用して，ユーザがどのようなコンテンツに関心があるのかを検知することになる．

配信サーバでは，ユーザのクッキー情報から，そのユーザが過去にどのようなコンテンツ（ページ）を閲覧したかを知る仕組みが実現されている．具体的には，配信サーバは，各

ユーザが始めてそのサイトに訪れた際に，ユニークな ID 番号が記述されているクッキーを配布する．2 度目以降のユーザのサイト訪問時には，配信サーバは，そのクッキーを参照し，どのユーザがどのページにアクセスしているかを検知する．配信サーバでは，どのページにアクセスしたかの情報を閲覧履歴として DB に格納する．

2.10 成果

全ての実施項目について，以下のとおり設定した目標を達成することができた．

2.10.1 ネット利用者の嗜好調査

1. 嗜好調査に最適なカテゴリを選定できた．
2. ネット利用者の嗜好調査用のポータルサイトを構築できた．
3. サイトの中にあるカテゴリに注目し，そのカテゴリに対するサイト訪問者の嗜好を調査した．
4. カテゴリ間の相関を導き出すことができた．

2.10.2 自己組織化機能を持った広告配信アルゴリズムの開発

1. 自己組織化，学習機能をウェブマーケティングへ応用した事例の調査を行った．
2. 自己組織化機能を持った複数の広告配信アルゴリズムを開発した．
3. 実証実験を行い，開発した候補手法の内から，最終的に採用する手法を,「基本配信モデル」,「マルコフモデル」の 2 つのアルゴリズムに決定した．
4. 採用した両アルゴリズムの改良を行った．

両アルゴリズムの特徴は以下のとおり．

(1) 基本配信モデル

- アルゴリズムの特性上，計算負荷が非常に低いという特徴が確認できた．
- ユーザの嗜好を非常に短いデータとして蓄えることが可能なため，これをユーザの PC にクッキーとして蓄えることができる．結果として，サーバのストレージが肥大化しないという特性がある．
- ユーザの嗜好を適応的に学習し続けるため，ユーザの嗜好の変化にも対応した広告配信が可能である．
- カテゴリ間の相関を学習に反映させることで，より即応性のある学習が可能となった．

(2) マルコフモデル

- 70〜80 %の正解率でユーザの遷移を予測することに成功した．

- 基本配信モデルよりも，1 セッション内におけるユーザの閲覧ページの遷移に追従した広告配信が実現できる事が確かめられた．
- 重みつきマルコフモデルを用いることでマルコフモデル作成に要する時間を短縮することができるようになったので，実装上問題無い事が確認された．
- あらかじめ用意しておく代表マルコフモデルの数を増やすことで予測精度の向上が可能となった．

2.10.3 システム開発

1. 相関関係のあるカテゴリを提示するプログラムを作成した．
2. ルーチン部を作成し，全体のシステムを完成させた．
3. 「自己組織化機能を持った広告配信アルゴリズム」の改良を行う都度，システムに実装した．

   サーバに実装するに当たっては，新アルゴリズムを一つの大きなクラスの集合体として扱うこととし，今回開発したアルゴリズムを適切にシステムに実装することが出来た．
   また，その一方で，アルゴリズムそのものについても必要な機能を洗い出し，細かなクラスとして構成することにした．更に，新アルゴリズムを汎用のサーバ機で稼動させるため，可能なかぎりスムーズに稼動するように調整を行った．
   実装されたシステムは，何のトラブルもなく稼動し続けており，実装当初に懸念されたサーバの資源不足という事態にも陥らないことがわかった．
4. 動作テスト，デバッグ，プログラムの改良を行った．

2.10.4 実証実験

平成 18 年 2 月末までに，基本配信モデルとマルコフモデルの改良版について，1ヶ月程度の実証実験を延べ 9ヶ月間行った．

2.11 今後の課題

代表マルコフモデル数に比例して計算コストは増大する．従って，システムに実装する際にはシステムのパフォーマンスとシステムにかかる負荷を考慮した上で代表マルコフモデル数を決める必要がある．その他，類似度の計算に使用したガウス関数の分散の値も最適値を探す必要性がある．

しかしながら，開発したシステムは，現在十分実用に耐える状態で稼働を続けており，これらの課題は，現時点では大きな問題ではない．今後，次章で述べる事業展開を行う中で，これらの課題を解決していくことにしたい．

2.12　今後の取り組み (事業展開)

今後は，以下のとおり事業化を進めていく予定である．

地域情報誌とのメディアミックスを前提としたパッケージ商品の開発

製品の性質上，単体での販売よりもそれに付随するあるいは連携する製品との複数商品での販売が主になると予想される．たとえば，あるポータルサイトに今回開発した配信サーバを販売するとした場合，配信サーバ単体の導入にとどまるよりも，配信サーバと連動したメール配信システム（メルマガ）や，あるいは携帯サイトとの連動システムなどを市場が欲するであろうということが容易に予測できる．

そのため，今後は，配信サーバに関連する周辺システムの拡充をおこなう．現状，ポータルサイトへの導入を考慮して，サイトメンテナンスに必要な入稿ツール，および携帯サイト構築のための自動ページ作成ツールなどの準備は整っているので，今後はこれらのようなツールを必要な順序で開発する．

開発された周辺システムは，配信サーバとともに，製品群に組み込み，オプションという位置づけで販売を行う．

地域情報誌と Web ページを連動させた広告配信ビジネスモデルの確立

現在，ビジネスモデルとしては，図１のようなものを検討している．図１において，広告代理店を担う会社がポイントになる．今回，このポジションに最も適している会社は，情報誌の発行業者などではないかと考えている．また，この場合，管理会社およびポータル運営会社ともに情報誌の発行業者が担うとスムーズであると考える．

また，何よりも重要なのは地域情報誌と連動することにより，地域の活性化が効率よく行えるという点にある．紙面による地域活性化にとどまらず，ウェブも利用したメディアミックスを実施することで，その効果がより大きくなると期待できる．

現在すでに 2 つの情報誌との連携が決まっており，それらのウェブサイトも稼動している．今後は同じようなビジネスモデルで，他誌との連携を促進する．

図 2.32: ビジネスモデル (イメージ)

配信サーバのタイプについて

今回の実証実験で開発した配信サーバは 2 種になる．基本配信モデルと，マルコフモデルがそれに当たる．これら 2 つはそれぞれに長所がある．基本配信モデルの長所は「計算負荷が低い」などが挙げられる．一方，マルコフモデルでは,「ユーザの 1 セッション内における動的な行動を予測できる」というような長所が上げられる．今後は，これらの長所を明確にし，お客様がお客様のニーズに応じてどちらのモデルを選択するのかを選べるように製品を構成していく．

さらに，必要であれば，基本配信モデルの簡易版も準備し，より手軽な製品として市場に普及させるようにする．

販売先および販売活動について

配信サーバを導入されるお客様として想定されるのは,「広告配信業者」ないしは「ポータルサイト運営業者」になるものと推察される.「広告配信業者」の利用する広告配信サーバは，非常に高負荷のかかる環境で稼動している．ここに割って入るためには，より厳密な負荷テストが必要になるものと思われる．そこで，特に，ポータルサイト運営業者へのセールスを強化することを考えている．ポータルサイトへの導入・運用をかさね，ノウハウを蓄積した上で，広告配信業者へのアプローチを行う．

ポータルサイト運営業者へのアプローチとしては，展示会などでの出展を積極的に行うとともに，自社での営業活動も行うものとする．とりわけ，地域情報誌との連携を進める．

地域情報誌は，より高い広告効果が望まれる紙面である．このことはその情報誌のウェブサイトでも同様のことが言える．そのため，地域情報誌のウェブサイト運営業は，今回の配信サーバを販売するのに最もよい会社のひとつとなる．

販売方法および販売対象・地域の多様化

販売については，当初は代理店などをおかずに直接販売のみを行うことにする．また，ニーズを調査した後のことだが，自社で配信サーバを運営し，ASP 的にサービスを提供することも検討する．

販売地域については，現在は福岡県下，得に筑豊地域としているが，地域情報誌との連携モデルが確立すれば，これを広く展開できると考えている．

付 録A

A.1 kMER(カーネルベース最大エントロピー学習)

各ユニットは荷重ベクトル $\mathbf{w}_i$ に加えて, $\mathbf{w}_i$ を中心とする局所 RF (Receptive Field : 受容野) カーネル K を持つ. 簡単のため, K の形状はガウス型のように, その中心の周りで放射状に対称的であるとする (図 A.1 参照) . RF カーネルが放射状に対称的であるから, あるしきい値での断面積は, 入力空間 V における半径 σ_i の円形 (一般には超球形) の RF 領域 S_i を定義することができる. 入力 $\mathbf{v}$ が S_i で張られる領域内に位置する場合はしきい値が引き上げられ, S_i は領域を狭められることになる. このとき, 同時に荷重ベクトル $\mathbf{w}_i$ も入力ベクトル $\mathbf{v}$ 方向に更新される. 入力 $\mathbf{v}$ が S_i で張られる領域内にない場合は, しきい値が引き下げられ, S_i は領域が拡大される (図 A.2 参照) . このような事象を一般化するために, 個々の S_i を次のようなコードメンバーシップ関数 $\xi_i(\mathbf{v})$ に関連づける.

$$\xi_i(\mathbf{v}) = \begin{cases} 1 & \mathbf{v} \in S_i \\ 0 & \mathbf{v} \notin S_i \end{cases} \tag{A.1}$$

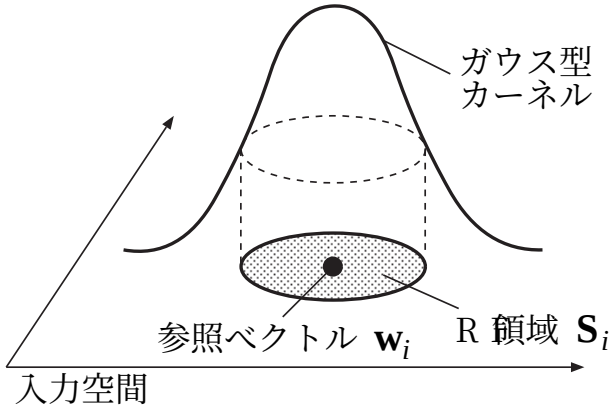

図 A.1: ガウス型カーネル及び, RF 領域

A.1.1 荷重ベクトルの更新

学習段階においては元の SOM アルゴリズムとは異なる型の競合の手法を導入する. まず, 次のようなファジィコードメンバーシップ関数 $\Xi_i(\mathbf{v})$ を定義する.

$$\Xi_i(\mathbf{v}) = \frac{\xi_i(\mathbf{v})}{\sum_k \xi_k(\mathbf{v})} \tag{A.2}$$

したがって, $0 \leq \Xi_i(\mathbf{v}) \leq 1$ および $\sum_i \Xi_i(\mathbf{v}) = 1$ である． Ξ_i に比例して, 一般的には $\mathbf{v}$ の方向に荷重ベクトル $\mathbf{w}_i$ を更新する． したがって, ある入力を複数のユニットで共有して活性化している場合には, 荷重ベクトルの更新量は小さくなる．

“バッチ” モードにおいて, N 個のユニットと M 個の入力サンプルが与えられたときの荷重ベクトル更新量 $\Delta\mathbf{w}_i$ は次式のようになる． $\Lambda(i,j,t)$ は (??) 式で表される近傍関数である． $\mathtt{Sgn}()$ は符号関数を表しており, ここではベクトルの各要素に適用している．

$$\Delta\mathbf{w}_i = \eta \sum_{\mu}^{M} \sum_{j}^{N} \Lambda(i,j,t)\Xi_j(\mathbf{v}_\mu)\mathtt{Sgn}(\mathbf{v}_\mu - \mathbf{w}_i) \tag{A.3}$$

$$\mathtt{Sgn}(v) = \begin{cases} 1 & v > 0 \\ 0 & v = 0 \\ -1 & v < 0 \end{cases} \tag{A.4}$$

A.1.2 RF 半径の更新

各ユニットのもつ RF 領域 S_i の中心である荷重ベクトル $\mathbf{w}_i$ の更新に続いて, RF 領域 S_i の半径 σ_i も更新する． 考え方としては, 位相マップが収束した時点での各ユニットの活性化確率が, 適当な係数 ρ を用いて, $P(\xi_i(\mathbf{v}) \neq 0) = \frac{\rho}{N}$ で与えられるように調節するというものである． RF 中心および半径を, 満足できる方法で展開するには, 経験的に各ユニットごとに 30 個の入力データを必要とすることが知られている． 全体で M 個の入力データがあるとすると, 以下の式で ρ に関する経験値が得られる．

$$\rho = \max\left(1, \frac{N30}{M}\right) \tag{A.5}$$

(A.5) 式で求めた ρ の値を用いて, “バッチ” モードにおける RF 半径の更新量 $\Delta\sigma_i$ は次式のようになる．

$$\Delta\sigma_i = \eta \sum_{\mu}^{M} \left(\frac{\rho}{N}(1 - \xi_i(\mathbf{v}_\mu)) - \xi_i(\mathbf{v}_\mu) \right) \tag{A.6}$$

A.1.3 最適化アルゴリズム

学習機構において式 (A.3) を用いる場合, 近傍関数 $\Lambda(i,j,t)$ の中心は個々の活性化ニューロンの周りに置かれ, これは全てのニューロンに応用される． このことにより, マップがもつれていたりする場合には, 初期段階においての計算に大きな苦労を要する． 上記の問題を解決するために, 計算的により有効な学習機構を展開することができる． このアルゴリズムでは 2 つの簡略化が行われている．

一つは, 荷重ベクトルの更新を (A.3) 式ではなく, 非活性ユニット j , 活性ユニット i それぞれについてに以下のように更新する． i^* は入力に対する最近傍ニューロンを表す．

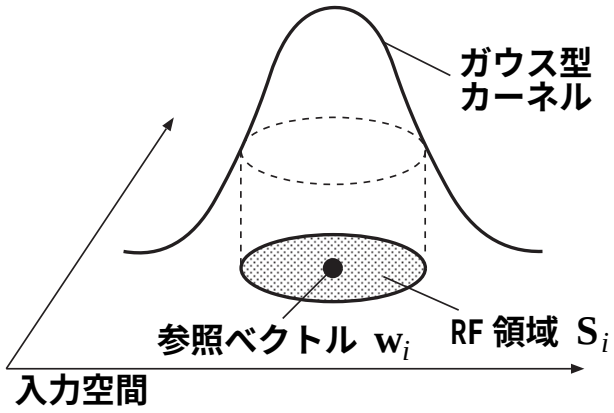

図 A.2: kMER におけるカーネル半径と荷重ベクトルの更新

$$\Delta \mathbf{w}_j = \eta \sum_{\mu}^{M} \Lambda(j, i^*, t) \Xi_{i^*}(\mathbf{v}_\mu) Sgn(\mathbf{v}_\mu - \mathbf{w}_j) \tag{A.7}$$

$$\Delta \mathbf{w}_i = \eta \sum_{\mu}^{M} \Xi_i(\mathbf{v}_\mu) Sgn(\mathbf{v}_\mu - \mathbf{w}_i) \tag{A.8}$$

二つめに, ある入力 $\mathbf{v}$ に対する活性化ユニットが全く存在しない場合でも, 最近傍ユニット i^* の RF 領域の半径 σ_i を常に更新すればよい.

以降の kMER による学習は, この “バッチ版の最適化 kMER アルゴリズム” による学習を指す.

A.1.4 学習例

kMER による学習の例として, 2 次元空間におけるガウス分布データを用いた学習結果を図 A.3 に示す. ユニット数が $N = 25$ の 2 次元 SOM (5×5) である. データ数は 900 (クラスごとのデータ数は 300 づつである.) , 学習回数 $t_{max} = 30000$ 回 , 学習係数 $\eta = 0.0005$, 係数 $\rho = \max(1, \frac{25*30}{900}) = 1$ で学習を行った. SOM と同様に, 初期の格子がもつれた状態から, 学習の過程で徐々に位相の整理がされる. 各クラスターに対して複数のユニットで捕えていく様子が分かる.

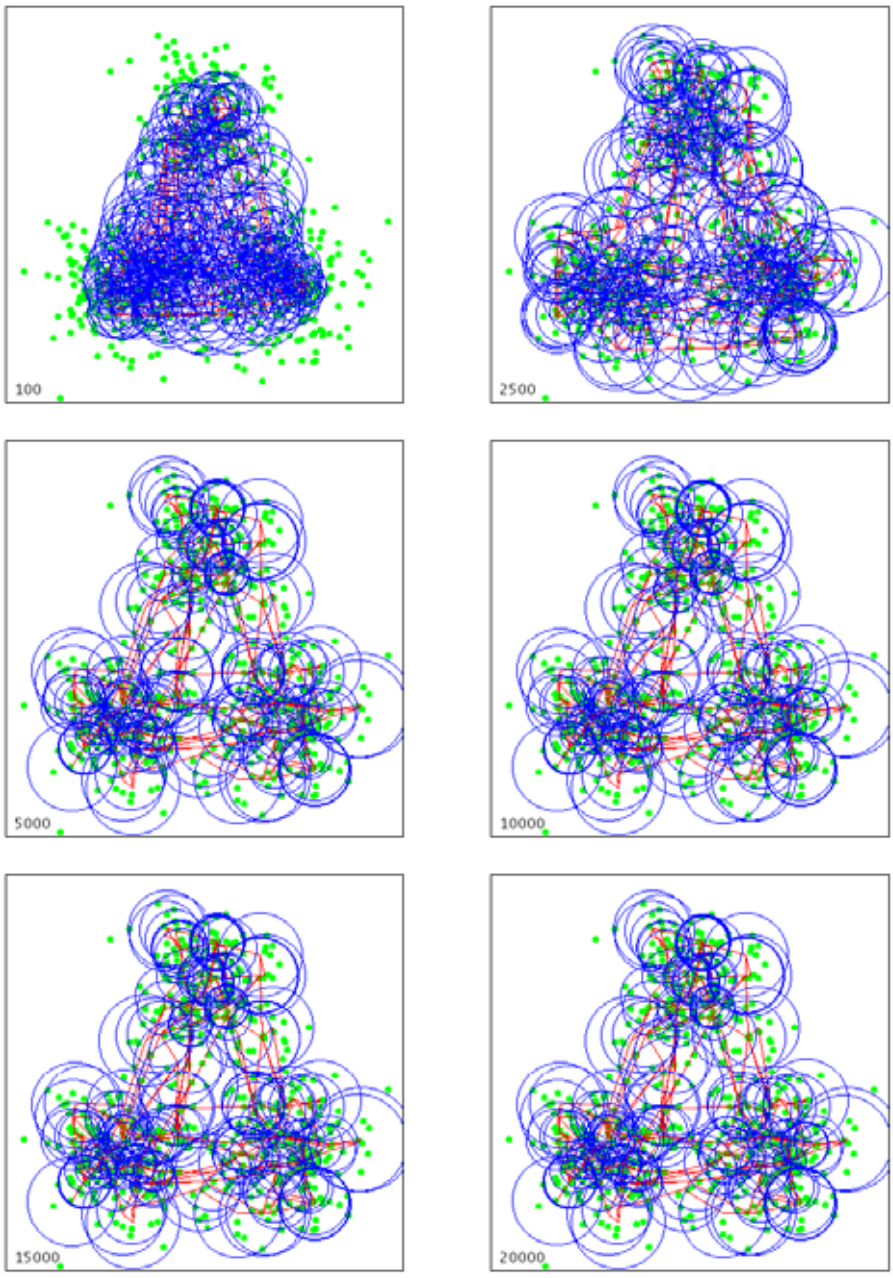

図 A.3: kMER による 2 次元ガウス分布データの学習過程
各円は荷重ベクトル位置を中心とした半径 σ_i の RF 領域を表す． 各円の中心を結ぶ直線は, 競合層上でのユニットの隣接関係を表す． 図中の多数の点は入力データを表す． 左下の数字は学習のステップ数を示す.