4W-4 Webアクセスログを利用したユーザモデルに基づく推薦システム

中嶋 敏行† 水原 徳洋† 太田 学‡ 石川 博‡
†東京都立大学工学部電子情報工学科
‡東京都立大学大学院工学研究科

1 はじめに

膨大な数のページを所有する大規模サイトの増加とともに、ユーザがたやすく目的のページにたどりつけるような支援システム[1]の需要が高まっている。本研究では、サーバのアクセスログを利用し、ＩＰやアクセス時間などのページ閲覧履歴を基にユーザの分類を行い、ユーザモデルに基づいて適切なページを推薦する Log-Based Recommendation System（L-R system）の構築法を提案する。

2 L-R system

推薦システムを構築するのに、多大のコストと時間がかかる現状を考慮し、本研究では、従来のサイトの内容を書き換えることなく、どんなサイトにも適応可能な推薦システムの提案を行う。この推薦システムは、アクセスログのみに基づいて推薦を行え、さらに推薦方法を自由に選択し、うまく組み合わせて利用することができるものである。

2.1 ユーザのページ閲覧履歴の抽出

ロボットなどの不必要なログの削除を行った後、図１のようにブラウザのバックボタンによる閲覧履歴を修正することにより、ユーザの正確なページ閲覧履歴を抽出し、ページ間の遷移確率を計算した。

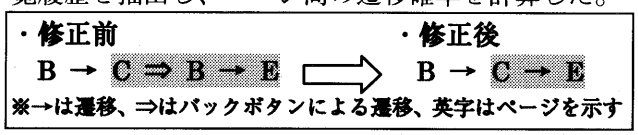

図１ アクセスログ修正

2.2 ユーザの分類

ユーザの分類を行う方法[2]として、IP、ホスト、アクセス時刻（時間、曜日、月など）、OS、ブラウザ、リファー、検索キーワードなどに基づいた方法がある。例えば、昼と夜にアクセスしてくるユーザには昼は会社員や主婦、夜は学生が多いなど明らかに異なった傾向が見られる。そのため、昼と夜でユーザを分類することでユーザの傾向にあわせた推薦を行うことができる。このように、ユーザを分類し、推薦を行うことは有用であるといえる。

Recommendation system using user models based on Web access logs
Toshiyuki Nakajima† Tokuyo Mizuhara† Manabu Ohta‡ Hiroshi Ishikawa‡
†Faculty of Engineering, Tokyo Metropolitan University
‡Graduate school of Engineering, Tokyo Metropolitan University

2.3 推薦

- **遷移確率のみを用いた推薦**

① リンク先読み推薦

現在閲覧中のページの遷移確率とそのページからリンクの張られている先のページの遷移確率を掛け算し、確率の高いページを推薦する方法である。この推薦は、３つ４つさらに先まで先読みして推薦することも可能である。

② リンクなし推薦

ユーザの閲覧履歴の抽出を行う際に、実際にはページ間にリンクは存在しないが、遷移は存在する場合がある。その遷移の数を集計し、数の多いものを推薦しようというものである。サイト構築者はこの推薦結果を基にリンクを張りかえ、ユーザの望むサイトに作り変えることも可能である。

③ 過去の履歴を利用した推薦

ユーザの現在閲覧中のページとそれより１つ前に閲覧していたページを利用して行う推薦である。A→B→C という遷移確率が高かったとすると、ユーザがＡページからＢページへ遷移したときに、次に閲覧する確率の高いＣページを推薦するものである。このようにすることにより、ユーザが同じページを閲覧しても毎回異なった推薦を提供することが出来る。この推薦は２つ、３つそれ以上の閲覧履歴を利用することも可能である。

④ パス推薦

ユーザのよく閲覧するページのパスを解析し、確率の高いパスを推薦する。D→E→F と遷移するユーザが多かった場合、D ページを閲覧した時にユーザに E、F ページを順にみるように推薦する。説明書や小説、利用規約など順番に読み進めていくようなページを持つサイトに有効である。

- **遷移確率にウェイトを付加した推薦**

① 履歴ウェイト推薦

ユーザの閲覧履歴において、最後に閲覧したページを目的のページと見立て、そこまで閲覧してきたページ全てから、最終ページを推薦するものである。集計方法は、A→B→C→D という遷移の場合、A、B、C ページにおける D ページへのウェイトを１増やすものである。最終ページに限る必要はなく、例えば、EC サイトでは商品の詳細ページを推薦対象ページに指定することも可能で

ある。同様の方法でユーザが一番長く閲覧したページを推薦対象とする滞在時間推薦なども考えられるが、アクセスログからでは、ユーザが最後に閲覧したページの滞在時間をとることができないため、ユーザが外部サーバに出ていく時に中間ページを用意するなどの工夫が必要である。

② リンクウェイト推薦

あるページが同じサーバ内のｎ個のページから指されている場合、ウェイトをｎとして遷移確率と掛け合わせ、その結果を基に推薦を行う。重要なページにはリンクが多く張られており、ユーザが目的とする内容である場合が多いため有効であると思われる。

以上説明した推薦方法をうまく組み合わせて利用することにより、各サイトに最適な推薦システムを構築することができる。

3 評価実験

本研究の推薦システムの評価を行うために、東京都立大学工学部電気・電子情報工学科のサーバを利用し、アクセスログもそこから入手した。

3.1 実験データと評価方法

アクセスログ：384941 件
ページ数：約 170 ページ　HTML 文書のみ
ログ収集期間：2000 年 6 月～12 月
被験者：大学内部の人 5 名と外部の人 5 名が参加
設定課題：「ネットワーク理論の授業内容を知りたい」など、課題を 5 つ準備
評価方法：被験者が課題の答えを探し当てるまでのクリック数と所要時間を集計し、それに基づいて各推薦方法の評価を行った。

図２　推薦システム表示形式

3.2 実験推薦システム

実験サイトで有用な推薦であると思われる、リンク先読み推薦、リンクウェイト推薦を実装した。比較対照として推薦を行わないもの（推薦なし）、バックボタンによる修正を行っていないログに基づいた遷移確率推薦（修正なし推薦）、2.1 で抽出した遷移確率による推薦（遷移確率推薦）も加えている。また、大学内部ユーザと外部ユーザによる分類も行っている。推薦システムの構築方法は図２のように、フレームによる分割を行い、上のフレームは元のサイト、下のフレームは推薦ページに設定している。上のフレーム内のページの移動に従い、推薦内容が変化するようになっている。推薦ページの表示は推薦するページの URL とタイトル、お勧め度（確率）を表示している。１ページにつきおよそ５件を推薦するように、お勧め度のしきい値を設定している。

3.3 評価と考察

結果を図３に示す。

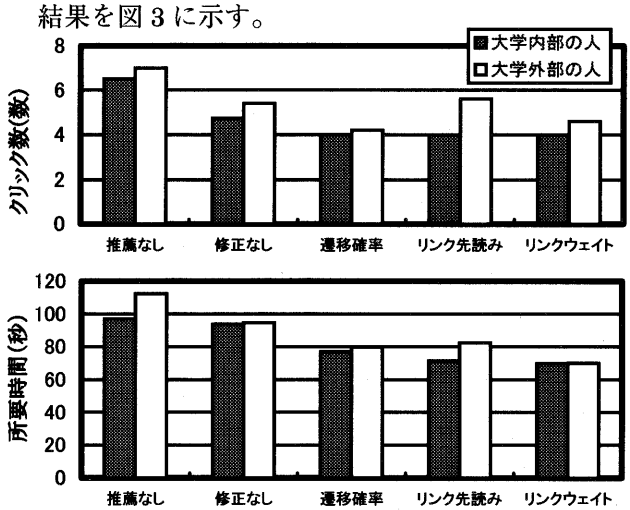

図3　クリック数と所要時間の増減表

今回の実験で推薦を利用することにより、クリック数と所要時間ともに減少させることができた。遷移確率推薦の所要時間が修正なし推薦より減少していることから、ユーザのより正確な閲覧履歴を抽出できているといえる。リンク先読み推薦のクリック数が遷移確率推薦より増加しているのは、目的のページを通り越すなどの悪い影響が表れていると考えられる。リンクウェイト推薦は遷移確率推薦よりクリック数ではわずかの増加がみられるものの、所要時間では減少しており、最もよい結果となった。また推薦なしでは内部ユーザと外部ユーザの差が大きかったが、リンク先読み推薦を除けば、推薦を利用することにより、所要時間の差を縮めることができた。本システムは大学外部ユーザのようなサイトに対して知識を持たない人により効果的であった。

4 おわりに

今回の研究では、L-R system の構築と評価を行い、有効性を示すことができた。今後の課題としては、課題や被験者の数を増やし、評価の信頼性を高めていきたい。さらに別のサイトでの評価実験も行っていきたい。また、HTML 以外のコンテンツへの応用も検討中である。

参考文献

[1]B.Mobasher, R.Cooley, J.Srivastava, "Automatic Personalization Based on Web Usage Mining", Communications of the ACM,Vol.43,No8,pp.142-151,2000.

[2]神田秀昭,"Web マーケティング 90 の鉄則",アスキー,2000