ユーザ特性を考慮したアクセスログに基づく情報推薦に関する研究

中村健二† 田中成典‡ 吉村智史† 細畠啓史† 北野光一† 田中裕一‡

関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

1．はじめに

近年，大規模なショッピングサイトのように膨大な情報を含む Web サイトが増加している．それに伴い，ユーザは雑多な情報の中から自分が求める情報の取捨選択を強いられている．こうした背景から，ユーザが求める情報の取得を支援する情報推薦システムへの要求が高まっている．情報推薦システムを実現する既存手法には，ユーザが過去に閲覧したページの内容を基に構築したユーザの興味モデルに基づく推薦手法 [1]-[3] やアクセスログから抽出した LCS(Longest Common Subsequence)[4]に基づく推薦手法[5]がある．LCS とは，Web ページを訪問したユーザが辿ったアクセス順序をセッションとしたとき，複数のセッションに共通して現れるアクセス順序である．しかし，前者は，事前に学習した興味モデルに合致する情報の推薦しか行えないため，ユーザの求める情報をリアルタイムに判断できないという問題がある．後者は，Web サイトを閲覧中のユーザのアクセス順序をアクティブセッションとしたとき，LCS とアクティブセッションを比較することでリアルタイムな情報の推薦を可能とするが，比較する LCS の数に比例して，推薦する Web ページの選出に要する時間も増大するという問題がある．そこで，本研究では，ユーザの興味モデルを構築し，推薦対象のユーザと類似する興味モデルを持つユーザ群の LCS のみを用いることで，計算時間を削減したリアルタイムな Web ページの推薦手法を実現する．

2．システムの概要

本研究では，アクセスログを用いてユーザが求める情報の推薦をリアルタイムに行うことを目的とする．本システムは図 1 に示すようにユーザの Web サイトに対するリクエストを入力とし，推薦ページの URL を出力とする．本システムは，1）アクセス傾向学習機能，2）ユーザの興味に則した LCS 抽出機能，3）ページ推薦機能の 3 つの機能により構成される．

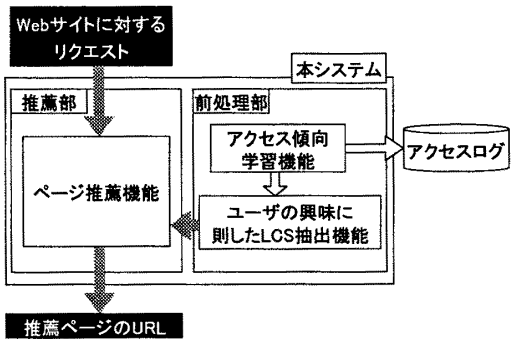

図 1　システムの流れ

2．1　アクセス傾向学習機能

本機能では，ユーザの興味モデルとして，過去に Web サイトにアクセスした全てのユーザに対して，ベクトル空間モデルで表現したアクセス傾向を抽出する．本研究で用いるベクトル空間モデルは，アクセスログより抽出した LCS を素性とし，その出現頻度を特徴量とする．そして，抽出したアクセス傾向を一定期間ごとに更新し，アクセス傾向の学習を行う．

2．2　ユーザの興味に則した LCS 抽出機能

本機能では，アクセス傾向学習機能で学習したアクセス傾向間の類似度を算出し，推薦対象のユーザと類似したアクセス傾向を持つユーザ群を抽出する．そして，抽出したユーザ群のアクセスログを対象に LCS を抽出し，ユーザの興味に即した LCS 群を作成する．

2．3　ページ推薦機能

本機能では，ユーザの Web サイトに対するリクエストを監視し，ユーザが現在閲覧している Web ページに辿り着くまでのアクセス順序をアクティブセッションとして記録する．そして，記録したアクティブセッションとユーザの興味に則した LCS 抽出機能で選出したデータを比較し，推薦する Web ページを選出する．

Research for Recommending Information based on Web Access Logs Regarding User Characteristics

†Kenji Nakamura, Satoshi Yoshimura, Hirofumi Hosohata, Koichi Kitano

Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

‡Shigenori Tanaka, Yuichi Tanaka

Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

3．システムの実証実験と考察

本システムの有用性を実証するため，ユーザの興味に即した LCS のみを用いて推薦を行う提案手法と従来の LCS に基づいて推薦を行う手法との推薦精度の比較実験を行った．

3．1　実証実験

本実験に利用する実験データは，既存手法[5]と同じ条件下で比較実験を行うため，"The Internet Traffic Archive"が公開している NASA の Web サイトに対するアクセスログを用いた．本実験では，アクセスログから抽出した 27,000 件のセッションの内，75%を学習データ，残りの 25%を実験データとして使用した．そして，実験データの各セッションの３ページまでをアクティブセッションとみなして推薦を行い，推薦する Web ページの選出に要する時間と比較件数を計測した．また，実験データ中でアクティブセッションに続いてアクセスされたページを正解ページの集合とし，F 値にて推薦精度の評価を行った．F 値は，推薦精度の総合的な評価指標であり，適合率と再現率の調和平均で表現される．適合率は，正解ページ集合をどの程度網羅しているかの指標である．再現率は，推薦の正確性の指標である．

3．2　結果と考察

実証実験の結果を表１と表２に示す．表１より，推薦する Web ページの選出に要する平均時間が既存手法の約 30%に削減できたことが確認できる．これは，ユーザの興味に即した LCS 以外を排除することで，推薦ページ選出に要する平均比較件数が約 30%を絞り込んだことに比例して削減されたと考えられる．また，表２より，推薦精度について，F 値を確認したところ，提案手法が従来手法と同程度の値を示すことがわかる．このことから，ユーザの興味に即した LCS のみを用いて推薦を行う提案手法が，推薦ページ選出に要する時間の削減に有効であることが実証された．適合率が従来手法より優れている理由は，ユーザの興味に即した LCS を用いることで適切に不要なデータを排除できたからであると考えられる．一方で，再現率が従来手法より低い値を示している理由は，推薦ページの選出にユーザの興味に即した LCS 以外を排除したため，ユーザが興味モデルと全く関連のない行動をとった場合に適切な推薦が行えなかったためと考えられる．また，従来手法と提案手法は共に適合率において低い値を示している．これは，推薦対象のユーザが推薦時に閲覧しているページとリンクしているページが推薦される場合が多いため，リンクは張られていないが関連の強いページを推薦できなかったことが原因であると考えられる．

表１　推薦に要する平均時間と比較件数

	従来手法	提案手法
平均時間(ms)	16,632	4,669
平均比較件数	43,650	12,802

表２　推薦精度

	従来手法	提案手法
適合率	0.2240	0.2528
再現率	0.5244	0.4660
F値	0.3292	0.3278

4．おわりに

本研究では，ユーザの興味に即した LCS のみを用いて推薦に要する計算時間を削減するリアルタイムな情報推薦手法を提案した．そして，システムの実証実験の結果からその有用性を実証した．今後は，本提案手法を実際のショッピングサイト等の商用サイトに適用する．そして，本提案手法を稼働中のショッピングサイトのアクセスログに適応し，Web サイトの構造の違いによるアクセス順序とアクセス傾向の相違点について詳細に調査を行う予定である．

参考文献

[1] Naren, R. : Web Personalization by Partial Evaluation, IEEE Internet Computing, IEEE, Vol.4, No.6, pp.21-31, 2000.7.

[2] 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，人工知能学会誌，人工知能学会，Vol.19, No.3, pp.669-701, 2004.5.

[3] 臼井大介，塚本享治：確率的手法を用いた Web ページ推薦システム，情報システムと社会環境研究会研究報告，情報処理学会，Vol.2006, No.27, pp.25-32, 2006.3.

[4] Wu, S., Manber, U., Myers, G., and Miller, W. : An O(NP) Sequence Comparison Algorithm , Information Processing Letters, Elsevier Science, Vol.35, No.6, pp.317-323, 1990.10.

[5] 山元理恵，小林大，吉原朋宏，小林隆志，横田治夫：アクセスログに基づく Web ページ推薦における LCS の利用とその解析，情報処理学会論文誌，情報処理学会，Vol.48, No.34, pp.38-48, 2007.6.