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

2006－DBS－140 (I) (19)
2006／7／12

WebアクセスログのLCSを用いたWebページの推薦手法

山元理絵[†] 小林大[†,††]
小林隆志[†††] 横田治夫[†††,†]

近年，Webサイトによる情報発信の重要性から，WWWは拡大を続け，ユーザのニーズに適したサイト構築や情報提供の要求が高まってきている．そこで，本稿ではLCS(Longest Common Subsequences)を用いたアクセスログ解析を行い，そこから発見されるパターンからユーザのアクセス行動を予測し推薦する方法について検討する．ユーザの初期のアクセス情報から，以降にアクセスされるページを推薦するため，多種多様なユーザのアクセスパターンの中で共通する傾向を抜き出し，アクセスの全体的な順序を保存するというLCSの持つ特徴を活かし，適切なページ推薦を行う手法を提案する．実際のアクセスログに対し本手法を適用して相関ルールを用いた従来手法との比較を行い，提案手法の優位を示すことでその性能を考察する．

Web Page Recommendation Using LCS Extracted from Web Log Analysis

Rie Yamamoto,[†] Dai Kobayashi,[†,††] Takashi Kobayashi[†††] and Haruo Yokota[†††,†]

Nowadays, the importance of the information transmission via websites leads to the explosive growth of the WWW, and therefore, development of the information techniques meeting users' needs is required. In this paper, we propose the method to recommend proper web pages to users based on the web access log analysis using LCS (Longest Common Subsequences). Our proposal evaluates the similarity between LCSs and the website visitor's behavior at an early stage, then predict subsequent patterns. We compare the proposed method with the technique using association rule mining by applying them to the real web access log data. Through experimental evaluation , we show that the proposed method can achieve better recommendation.

1. はじめに

近年，WWWの規模は拡大を続け，Webサイトの構造は複雑性を増している．それに伴い，ユーザがWebのナビゲーションを通して欲する情報を得るまでに多大な時間を要したり，もしくは必要な情報が得られないといった問題が生じてきている．一方，WWWは今日では広報・広告・マーケティングの手段として極めて重要な位置を占めているため，情報の発信者である企業は，顧客のニーズを的確に掴み，ユーザビリティの向上や顧客の関心を引き付けるようなサービスの提供を行う必要がある．

ユーザのナビゲーション行動の傾向を判断する材料として適しているのが，Webアクセスログである．アクセスログに含まれる情報には，ユーザのIPアドレス，URL，アクセス日時や，Cookie等に入力された内容などがある．これらの解析によってサイト管理者は，ページ別，時間帯別のアクセス状況や自サイトへのアクセス元，さらにはサイト全体における巡回の傾向を知ることができる．

アクセスログの解析手法に関しては，相関ルールを用いた頻出共起ページの組の抽出[1),2)]やバックトラックポイントの発見[3)]，利用頻度に基づくリンクの接続性の評価[4)]等，様々な手法が提案されてきた．以前は，ページ構造の改良によるサイト管理者のコスト軽減や，事前に次にアクセスされそうなページの読み込みを行うことでユーザビリティの向上を図ることを目的とするものが多かった．

近年では，ビジネスの場としてのWebの役割の増大から，Web personalizationが注目され，中でも特にユーザの嗜好に合ったページ推薦手法がさかんに研究されている[5)]．

ページ推薦手法としては，ユーザによる過去のアク

† 東京工業大学大学院情報理工学研究科計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
†† 日本学術振興会特別研究員 DC
Research Fellow (DC), Japan Society for the Promotion of Science
††† 東京工業大学 学術国際情報センター
Global Scientific Information and Computing Center, Tokyo Institute of Technology

セスの履歴と現在のユーザのアクセスパスを比較し，各々のページ間の遷移確率から，次にアクセスされるページを予測するものがある[6]．履歴として保管しておくパスの長さが長いほど推薦精度が上がるという結果が示されており，保管に必要となるスペースとのトレードオフに関する評価がなされている．しかし，特にサイト構造が不明瞭な時やその規模が大きい時には，目的のページに到達するまでに後戻りや遠回りを含んだり，目標が複数存在することで，多種多様なパスを取りうる．したがって，このようなアクセスパスのマッチングに基づく推薦では，少しでも頻出パターンから外れたユーザには正確な推薦ができず，さらに現在までのアクセス履歴から推薦されるのは直後のページのみに限られるという問題もある．

[7]では，apriori アルゴリズムを用いて 1 トランザクション中で頻繁に共起するページの組の集合及びページアクセスにおける相関ルールを作成し，そこからページ推薦を行うための手法が提案されている．この手法では，頻出アイテムセットに加えるための最小 support 値の設定に，サイト内における配置を考慮して変化させる手法[8]を採用しており，また，履歴として用いるユーザのアクティブセッションの長さの閾値を推薦ページが見つかるまで下げていくという方法で精度を上げている．アクセスパスそのものを扱うのではなく，共起頻度の高いページ同士の関係を見て推薦を行うため，前述の問題を解消することができる．しかし，共起率の高いページを推薦するこの手法では，セッションに対しページアクセスの順序情報を含まないため，ページ参照の順序に特徴的な傾向がある場合など，ページアクセスの順序によってはすでにアクセスしたページを推薦したり，もしくはユーザにとってもはや不要となったページを推薦してしまう可能性がある．

これらの欠点を補う手段として，LCS（Longest Common Subsequence）によるパターン抽出がある．これは Web サイト内における全てのセッションの URL の推移をルートとして抽出し，各ルートに対して，全てのルートそれぞれとの LCS（最長共通部分列）を求め，頻出アクセスパターンを発見するものである．これを用いることで，ユーザ同士のアクセスパスが完全に一致しない場合でも全体のアクセスの傾向を表現することが可能になり，順序情報を保持したまま非常に多様なユーザの Web サイト内における行動の理解に役立てることができる．現在までに我々は，LCS を用いたアクセスログの解析手法とサイト構成の改善手法の提案[9]，さらに LCS 抽出の効率化を行ってきた[10]．

そこで本稿では次のステップとして，ユーザのニーズに合うページ推薦を目的とし，アクセスログ解析によって発見された LCS を用いて現在までのアクセス履歴から次にアクセスされるページを予測して推薦する手法を提案する．さらに，この手法を使って実際に Web サイトのアクセスログを解析することで本手法の有効性を検証する．

2. Web アクセスログ解析への LCS の適用

本節では，我々が[9],[10]で提案したアクセスログから LCS を抽出する方法について述べる．LCS を抽出するためには，まずアクセスログからユーザセッションを切り出し，データを精錬する必要がある．その後各セッションを比較することで LCS を抽出し，その頻度を集計する．

2.1 ユーザセッションの抽出

アクセスシーケンス解析を行う際，蓄積されている未加工のアクセスログを精錬してマイニングに必要なデータのみを取りだし，ユーザのセッションを抽出する必要がある．以下では，その手法について説明する．

2.1.1 セッション ID のトラッキング

各訪問者の解析対象となるサイト内におけるページ間の移動情報を得るため，セッションごとのアクセス URL の集合をアクセスシーケンスとして整理する．

アクセスログに含まれる IP アドレスには，同一 IP から複数のユーザがアクセスしている場合，それらを個々のユーザ毎のセッションとして区別するために，Cookie を用いて各セッションに一意なセッション ID を割り当てる方法が有効である．特に，ユーザのナビゲーションに合わせて，リアルタイムに適切なページ推薦を行うことを目的とする本研究においては，このような手段を用いることが好ましい．しかし，Cookie はユーザの了承を必要とするため，Cookie 情報が利用できない場合は，正確性は下がるが IP アドレスの情報などで代用する．

セッション ID ごとに整理された URL の集合を時系列順に並べることで，各セッションで訪問者が行ったアクセスの URL シーケンスを得ることができる．

2.1.2 アクセスログの精錬

アクセスログの中には解析処理を行う際に不必要な情報や，目的に合わない情報が多く含まれる．ここでは，実際の解析を行う前に，それらの情報を取り除くアクセスログの精錬について述べる．

[11]にある標準的な前処理により，Web ページ間の移動情報に関係のない HTML ファイル以外のファイルへのリクエストを取り除き，検索エンジンのロボッ

トの情報等も除去する．

さらに，訪問者が対象サイト内の1ページのみを見てセッションを終了した場合，目的とするサイト内での移動情報を得ることができないため，シーケンス情報として利用することは難しいと判断し，このような情報も取り除くこととする．

2.2 LCSの抽出と頻度集計

まずLCSの定義とその特徴について簡単に説明し，次にLCSの抽出手法と頻度集計について述べる．

2.2.1 Longest Common Subsequences

リストxの部分列x_aとリストyの部分列y_bの中で両方のリストに含まれるものを共通部分列という．共通部分列の中で最も長いものを最長共通部分列 (Longest Common Subsequences) と呼び，LCSと略す．

二つのリストの中に同じ要素が同じ順序で出現したものが共通部分列なので，LCSが長いということは二つのリストの類似性が高いことを表す．

以下にリストX，YのLCSを抽出した例を示す．

$$X = (\underline{A}, F, \underline{B}, \underline{D}, E)$$
$$Y = (\underline{A}, \underline{B}, C, \underline{D}, G)$$
$$LCS(X, Y) = (A, B, D)$$

これをURLシーケンスに適用することで，アクセスシーケンスが完全に一致しない場合でも，寄り道等の余分な情報を取り除くことにより各々のシーケンス間の類似性を発見することができると考えられる．

2.2.2 LCSの抽出

二つのシーケンスからLCSを求めるには，一般的に動的計画法が用いられる．これにより長さMの文字列Xと長さNの文字列YのLCSを求める場合，$O(MN)$時間でLCSを求めることができる[12]．

ここで，LCSを求める問題と等価である，SED(Shortest Edit Distance)を求める問題に関して，効率化された手法が提案されている[13]．この手法では，比較する二つの文字列の差異が小さいほど必要とする時間計算量が小さくなるため，実際のデータに適用すると，多くの場合で$O(MN)$よりも大幅に小さい計算量でLCSの計算が可能になる．

2.2.3 LCSの頻度解析

Webアクセスログから得られたアクセスルートに関して，前節で示した手法により解析を行う．URLシーケンスをURLを各要素に持つシーケンスとみなす．各ルートについて総当たり的にLCSを抽出する．ここから各LCSの集計を行い，高頻度で出現するLCSパターンを発見する．

[9]では，LCS抽出のための計算において，全てのアクセスシーケンスに対して総当りで求めるため，セッション数が大きくなるとその二乗に比例して時間計算量が増加してしまい，計算コストが大きいという問題があるが，本研究では，[10]で提案されているハッシュを用いたアクセスシーケンスのフィルタリング手法やインクリメンタルなLCS抽出手法，並列計算のためのアルゴリズムを用いることで，LCS抽出にかかわる計算量をさらに抑えることとする．

3. LCSを用いたページ推薦

1節で述べたように，本研究では，前節で説明した手順でアクセスログから抽出したLCSを用いて，ユーザにページを推薦する手法を提案する．

3.1 推薦候補ページの選出と重み付け

提案手法では，前節のようにしてアクセスログから抽出したLCSのそれぞれと，現在のユーザのセッション (アクティブセッション) とのマッチングを行い，頻出LCSの中で，ユーザの現在位置以降に現れているページを推薦する．

ある1つのLCS lcs_iに対し，全セッション中において数え上げられた回数c_iが閾値*min.Count*を満たし，かつ長さが*min.Length*以上であるLCSの集合を$LCSs = \{lcs_1, lcs_2, \cdots lcs_k\}$で表す．ここで，*min.Count*と*min.Length*は，Webサイトの持つ特性に合わせて設定するパラメータである．このとき，nページのアクティブセッションact_nからそれに続くユーザのページアクセスを予測する場合を考える．

LCSの持つ，2つのセッションから共通部分を抜き出すことでそれらの傾向を表すという特性から，lcs_iとアクティブセッションを比較する際，それぞれが完全に一致する必要はなく，共通要素が多く存在する場合にlcs_iはそのアクセス傾向の特徴を表現していると捉えることができる．したがって，lcs_iとact_nの間で共通しているページを調べ，lcs_iの後半部分の中でまだアクセスされていないページがあれば，そのページはその後にアクセスされる可能性が高いといえる．

また，推薦順を決定するための推薦候補のページへの得点の重み付けに際しては，以下のような点を考慮する必要がある．まず，2節で述べたLCSの抽出方法からわかるように，高いc_iを持つlcs_iは，多くのセッションにおいて頻繁にナビゲートされたアクセスパスの部分シーケンスであり，これは出現頻度が高い事を意味するため重視すべきである．また上述のように，lcs_iと一致の度合いが高いact_nは同じ傾向を持つ可能性が高いため高い得点を付加すべきである．以上を考慮し，ページp得点の算出方法として次の式を用いる．ただし，ページpは候補ページの集合に含まれ

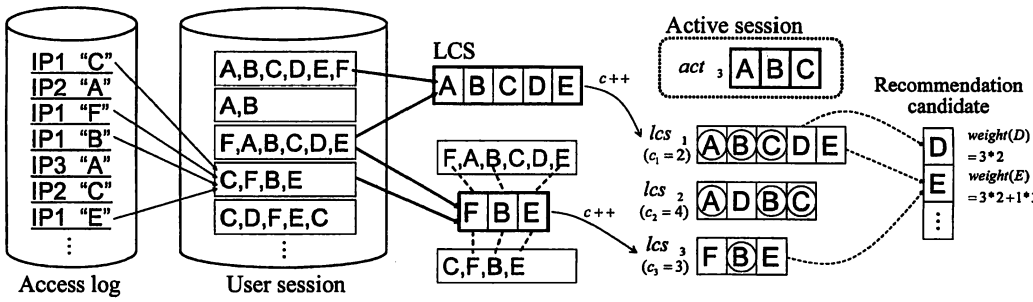

図 1 LCS を用いた推薦

るとする.

$$weight(p) = \sum_{lcs_i \in LCS\,s} |lcs_i \cap act_n| \cdot c_i \qquad (1)$$

3.2 ページ推薦手法

以上の点を踏まえて，本研究では，以下の手順で推薦ページの集合を作成する.

(1) lcs_i と act_n の間で共通するページを抜き出す.

(2) lcs_i より，一番目から共通部分の最後までの要素全てを除去する.

(3) 残ったページを推薦ページの候補とし，得点を付加する.

(4) LCS の中で閾値を満たす全ての要素対して (1) ~ (3) を行い，候補ページの中で得点の総和が上位のページを推薦する.

例えば図 1 のように $act_n = (A,B,C)$ が与えられた時，$lcs_i = (A,B,C,D,E)$ と一致する部分は (A,B,C) であり，それに続くページ D,E が推薦の候補ページに加えられる．また $lcs_i = (A,D,B,C)$ については，同様に (A,B,C) が一致するものの，共通部分の最後のページ C 以降に続くページはないため，ここから推薦候補に加えられるページはない．したがって，この例では推薦されるページの候補は $\{D,E\}$ となる.

この例では，ページ D が 2 つの LCS で推薦候補となっている．したがって，各 LCS から付加された得点の和がページ D の得点となる.

4. 評価実験

本研究で提案する手法の有効性を確認するため，実際の Web サイトのアクセスログに対し，本手法と，1 節で触れた Mobasher らが提案している apriori アルゴリズムによって生成される相関ルールを用いた推薦[7] 手法を簡単化した手法を実装した実験システムを用いて評価し，比較を行った.

4.1 準備と評価指標の定義

対象としたアクセスログは，"The Internet Traffic Archive" (http://ita.ee.lbl.gov/index.html) で配布されているいくつかの Web サイトにおけるアクセスログの内，NASA の Web サイトでの 1995 年 8 月 1 日から 8 月 31 日までの Web サーバへのリクエストに対するアクセスログを用いた．実際に推薦を行う場合には Cookie などによりリアルタイムに一意的なユーザの特定を行うことが好ましいが，今回使用したデータには Cookie の情報が含まれていなかったため，同一 IP アドレスからのアクセスを同一ユーザからのアクセスとみなし，ページアクセスの間隔が 1,200 秒以上のときにセッションを分割した．ログ全体の中に出現した固有 URL 数は 1,276 で，総セッション数は 39,900 であった．ここで，各 URL のログへの出現頻度には大きな偏りがあったため，各セッション中の出現割合が 0.5%に満たない URL を取り除き，さらに推薦の評価に利用できないため長さが 3 以下のセッションも除外した結果，URL 数は 174，総セッション数は 23,663 となった.

全セッションの内，時期が早い方の約 75%を学習セットとしてそこから LCS, 相関ルールを用いて頻出パターンを抽出し，残りの約 25%の新しいセッションの集合をテストセットとみなしてページ推薦を行い，その評価を行った．テストセットに含まれるテストセッションの長さの分布を図 2 に示す.

比較対象として用いる[7] では，n ページのユーザのアクセス履歴にが与えられると要素数 $n+1$ の頻出アイテムセットを探索し，アクセスされた n 個のページを全て含むアイテムセットから，差分のページを推薦するという手法が提案されている．これは，ページの組 {A,B}，{A,B,C} がそれぞれ頻出アイテムセットに含まれる時，{A,B,C} の共起頻度が高ければ，{A,B} へのページアクセスに引き続いて C へのアクセスが起こる確率も高くなるという性質を利用している．この手法では，前述のように履歴として用いるユーザのアクティブセッションの長さの閾値を推薦ページが見つ

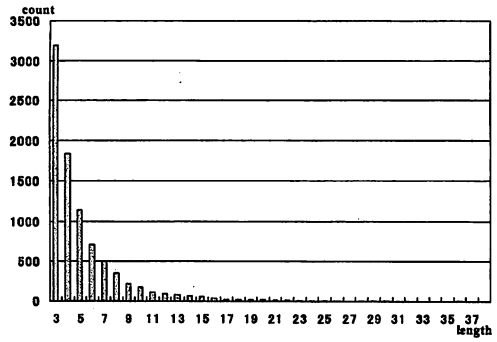

図 2 テストセッション長の分布

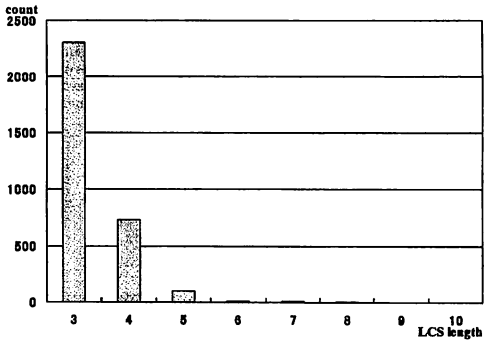

図 3 LCS 長の分布

かるまで下げていくという方法を取っているが，今回は，簡単化するため n ページのアクティブセッションからは要素数 $n+1$ のアイテムセットのみを用いて推薦することとする．

実験に用いたパラメータは，比較手法に関しては一定の最小サポート値 0.008 を用い，要素数が 1 から 4 の頻出アイテムセットを生成した．また LCS については，*min.Count* を 150，*min.Length* を 3 として *LCS s* を作成した．

評価に用いる指標として，以下に定義される *precision* と *coverage* を用いる．

$$precision(Recom_n) = \frac{|Recom_n \cap eval_n|}{|Recom_n|} \quad (2)$$

$$coverage(Recom_n) = \frac{|Recom_n \cap eval_n|}{|eval_n|} \quad (3)$$

ここで，$Recom_n$，$eval_n$ はそれぞれ n ページのアクティブセッションから導かれた推薦ページの組，アクティブセッションに引き続いて実際にアクセスされたページの組 (評価セット) を表す．precision は推薦の正確性の指標であり，推薦されるページ数に対する正解ページ数の割合で表現される．また，coverage は評価セットをどれだけ網羅しているかの指標であり，評価セットのページ数に対する正解ページ数の割合で表現される．

4.2 実験結果

ページ推薦を行うために，テストセットの各セッション (テストセッション) のはじめの n ページをアクティブセッションとみなし，そこから LCS を使った提案手法と頻出アイテムセットを使った比較手法のそれぞれでページ推薦を行い，そのセッションの残りのページと比較することで precision と coverage を求めた．

図における各点は，推薦する上位ページの数 $|Recom_n|$ を 1 から 14 の間で変化させた場合に対応しており，右の点ほど $|Recom_n|$ が大きい場合を表す．

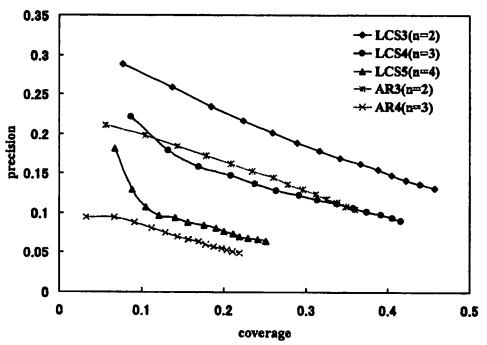

図 4 提案手法と相関ルールを用いた手法の比較

実験の前段階として，2 節の手順で LCS を抽出した結果，生成された LCS の長さの分布は図 3 のようになった．

n ページのアクティブセッションからページ推薦を行う際，比較手法では要素数 $n+1$ の頻出アイテムセットのみを用いて計算することとする．提案手法についても同様に長さ $n+1$ の LCS のみを用いて推薦を行い，それぞれについて推薦するページ数 ($|Recom_n|$) を変化させた時の結果が図 4 である．

図中の LCSx，ARx はそれぞれ，$x-1$ ページのアクティブセッションから長さが x の LCS，要素数が x のアイテムセットを用いて推薦を行った際の結果を表している．図 4 から，同じ長さのアクティブセッションに対するページ推薦の精度について，提案手法が比較手法に比べて良い結果が得られていることがわかる．

また，提案手法においては，アクティブセッションと LCS が完全に一致しない場合でも，一致する要素数の割合で重み付けをしているためページ推薦を行うことができる．そこで，n ページのアクティブセッションから推薦を行う際に，使用する LCS 長を限定せず

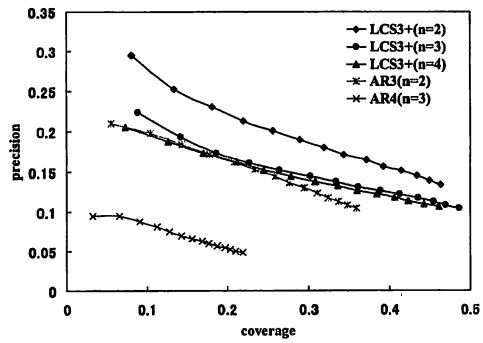

図5 n ページからの推薦に全ての LCS を用いたときの比較

に *LCS s* 内の全ての LCS を用いた場合についても実験を行った．結果を図5に示す．図中の各 LCS3+については，長さが n のアクティブセッションから生成された全ての LCS を用いて推薦を行った時の値を表している．

4.3 考　　察

両図を通してほぼ右下がりの曲線となっていることから，全体として $|Recom_n|$ の値を大きくするほど coverage が上がり，また precision が下がる傾向が見て取れる．

図4より，同じ長さのアクティブセッションから推薦を行った場合，相関ルールによる推薦に比べ提案手法が優れていることがわかる．

一般に，n の値を大きくするほど履歴から抽出されたパターンとの一致の度合いが大きくなり，precision は改善するはずであるが，図4，5において n の値が小さい方が良い結果が得られている．これは，今回対象としたサイトではサイト構造が階層化，細分化されており，目的のページへの短いアクセスパスでナビゲーションを終えるユーザが大半を占めることが原因であると考える．図2からも明らかであるように，セッション長が長くなるほどテストセッションそのものの存在数が少なくなっているため，n を大きくすると正解の存在数，すなわち $|eval|$ も小さくなる．それに従い式(2)より precision の値も低下する．これは図4における提案手法（n=4）の左4点（$|Recom_4| = 1 \sim 4$）に顕著に現れている．

一方，このようなセッション長の偏りに起因して図3のようにログから抽出される LCS 長にも大きな偏りが生じ，n の値が大きくなるほど推薦ページの決定に用いることができるパターンの数が減少してしまう．これによって，n の値が大きいときに精度が低くなっているが，3節において示した LCS による推薦の特徴から，用いる LCS の長さに幅を持たせたところ，長さを固定した場合に比べ，図5のように n=3，4の場合で改善されている．しかし，n=2 の場合については若干の悪化が見られた．これは，LCS 長が3と4以上の LCS の要素数に大きな差があり，LCS 長が4以上である，サンプル数の少ないものも推薦時の判断材料とすることで，精度が下がったためであると考える．

今回，重み付けの計算について，式(1)以外にも様々な計算方法を用いて同様の実験を行ったが，c_i の重みを極端に低くした場合に精度が落ちることを除いて，結果に大きな変化は見られなかった．各 lcs_i の出現回数 c_i の分布を調べたところ，偏りが非常に大きく，また，*min.Count* の値を大幅に上げ，*LCS s* から LCS の大半を取り除いても結果の変化は小さかった．以上から，今回対象とした Web サイトでのユーザのアクセス傾向に対しては，c_i が上位であるごく少数の lcs_i に含まれる URL が推薦結果に大きく影響を及ぼしていることが原因であると考える．したがって，今回用いたサイトのように URL 毎のログへの出現率の偏りが大きい場合，上位に含まれる LCS のみを重み付けに用いることで，精度を下げずに計算量を削減することができると考える．

また，今回対象としたアクセスログ以外にも，我々の研究室の Web サイトのアクセスログに対しても本手法を適用したが，アクセスパターンに多くの異なる傾向が見られた．これは，研究室の Web サイトは規模が小さく，厳密な階層構造を取っていないことが原因であると考える．

5. おわりに

本稿では，アクセスログ解析によって抽出された LCS を用いて，ユーザの過去のアクセス行動からそれに続くアクセスページを予測する手法を提案した．また，実際のアクセスログにそれを適用し，相関ルールを用いた手法と比較することでその有効性を検証した．

本研究では，LCS が，アクセスパスが完全に一致しない場合でもユーザ同士の傾向の類似性を表現可能であり，また，アクセスパスの部分シーケンスであるため完全なパスでなくてもアクセスの順序情報を保存するという特徴を持つことに着目し，アクティブセッションとのマッチングにおいても，LCS との一致の割合を考えることで，推薦精度の向上を図った．

実際のアクセスログを用いて，その一部から LCS を抽出し，残りのログに対して提案手法による推薦のシミュレーションを行った．apriori アルゴリズムによって生成された相関ルールを利用した従来手法と比較し，

提案手法の優位性を実証した．

考察で述べたように，実験で用いたサイトのようにURL毎のログへの出現率の偏りが大きい場合，出現率が高いLCSの影響が大きい．したがって，上位に含まれるLCSのみを重み付けに用いることで，推薦時の計算量が大幅に削減でき，オンライン処理である推薦ページ決定時の計算時間を減少できることが確認できた．一方で，アクセス頻度が均一で，頻出URLページが多数存在するようなWebサイトにおいては，抽出されるLCSの長さや種類が増えるはずであり，検証が必要である．

今回の結果では重み付けの式による精度の変化はあまり見られなかった．しかし，階層化された商業サイト等でコンテンツページを優先的に推薦する場合には，サイト構造における深さなどを考慮すべきであり，それに合わせて重み付けの式にも変更が必要である．

考察でも述べたように，規模やサイト構造の違いによりユーザのアクセスパターンに異なる傾向がある場合，それらを考慮して適切な推薦を行う必要がある．サイトの特徴と推薦精度の関連性の調査が今後の課題である．

今回用いたアクセスログではセッション長が短いユーザのアクセス傾向と長いユーザのアクセス傾向に違いが見受けられた．このことから，同じサイトの中でも，複数の異なる推薦アルゴリズムを用いることが有効であると考える．したがって，複数の推薦アルゴリズムを利用し，リアルタイムに推薦アルゴリズムを切り替える方法を検討することも今後の課題である．

謝辞 本研究の一部は，文部科学省科学研究費補助金特定領域研究 (18049026)，東京工業大学21世紀COEプログラム「大規模知識資源の体系化と活用基盤構築」および独立行政法人科学技術振興機構戦略的創造研究推進事業CRESTの助成により行なわれた．

参 考 文 献

1) R.Agrawal and R.Srikant: Fast Algorighms for Mining Association Rules, *Proc. of 20th Intl. Conf. on Very Large Data Bases*, pp.487–499 (1994).
2) S.Sarawagi, S.Thomas and R.Agrawal: Integrating association rule mining with relational database systems: alternatives and implications, *Proc. of ACM SIGMOD Intl. Conf. on Managgement of data*, pp. 343–354 (1998).
3) Srikant, R. and Yang, Y.: Mining web logs to improve website organization, *Proc. 10th Intl. Conf. on WWW*, pp.430–437 (2001).
4) B.Mobasher, R.Cooley and J.Srivastava: Creating Adaptive Web Site Through Usage-Based Clustering of URLs, *Proc. of the 99 Workshop on Knowledge and Data Engineering Exchange*, p.19 (1999).
5) Eirinaki, M. and Vazirgiannis, M.: Web mining for web personalization, *ACM TOIT*, pp.1–27 (2003).
6) J.Pitkow and P.Pirolli: Mining longest repeating subsequences to predict WWW, *Proc. of the 1999 USENIX Annual Technical Conf.* (1999).
7) Mobasher, B., Dai, H., Luo, T. and Nakagawa, M.: Effective perssonalization based on association rule discovery from web usage data, *Proc. 3rd Intl. Workshop on Web information and data management* (2001).
8) Liu, B., Hsu, W. and Ma, Y.: Mining association rules with multiple minimum supports, *Proc. of the ACM SIGKDD intl. Conference on Knowledge Discovery and Data Mining*, pp.337–341 (1999).
9) 宇根田純治, 横田治夫：Webログの共通シーケンス解析，信学技法DE2002-2，電子情報通信学会 (2002).
10) 戸田誠二, 横田治夫：LCSを用いたアクセスログ解析の並列処理による性能向上，第13回データ工学ワークショップ論文集, DEWS2004 7-B-5 (2004).
11) Banerjee, A. and Ghosh, J.: Concept-based clustering of clickstream data, *Proc. 3rd Intl. Conf. on Information Technology*, pp.145–150 (2000).
12) Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C.: *Introduction to Algorithms*, MIT Press (1990).
13) Wu, S., Manber, U., Myers, G. and Miller, W.: An O(NP) sequence comparison algorithm, Information Processing Letters,35:317–323 (1990).