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

山元 理絵[†] 小林 大[†,††] 吉原 朋宏[†]
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，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．Webアクセスログを Webページ推薦に用いる方法は，クライアント側に手を加える必要がなく有用であるが，これまで提案されている手法では，頻出アクセスパターンとわずかでも外れると適切な推薦ができない，あるいは順序を考慮できないといった問題点があった．我々は，それらの問題を解決するために，Webアクセスログから LCS（Longest Common Subsequences）を抽出してページ推薦に利用する手法である WRAPL を提案している．本稿では，実際の Webアクセスログを用いた実験を通して WRAPL の効果を詳細に解析し，その実験結果から得られた知見を基に優先順位付け手法に対して改良を行い，その有効性を示す．

Analyses of the Effects of Utilizing Web Access Log LCS for Web Page Recommendation

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

Sophisticated websites satisfying users' requirement becomes much more important to propagate information via websites, nowadays. Web page recommendation methods using web access logs are useful for them because they need no modification in client-side applications to meet the requirement. However, traditional methods have problems of insufficient recommendation precision caused by strict matching of access patterns or neglect of access sequences. To solve the problems, we are proposing WRAPL as a method of extracting LCSs (Longest Common Subsequences) from web access logs and using them to recommend web pages for an active session. In this paper, we analyze the effects of WRAPL using actual web access logs and propose an enhanced weighting method for it to improve the precision based on the analyses.

1. はじめに

近年，ビジネスの場としての Web の役割と情報量の増大から，コンテンツ推薦やサイト構造の動的な変更に代表される，Web パーソナライゼーションが注目され[1)]，特にユーザの嗜好に合った Web ページをシステムがユーザに推薦し提示する Web ページ推薦が求められている．

† 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

Web パーソナライゼーションは一般的に，(i) データの収集，(ii) 前処理，(iii) 解析，(iv) 推薦や動的なサイト構造変更等のアクションの手順で行われ，(i) のデータの収集では，ユーザからの入力による情報や Web アクセスログ等が利用される[1)]．後者はユーザの行動パターンや傾向を抽出することが可能であるため，さまざまな分野，目的で研究が行われている[2)]．また，Web パーソナライゼーションを行う際に最も重要な要因となる (iii) の方法は，以下のように分類される[1)]．

- 内容に基づくフィルタリング
- 協調フィルタリング
- ルールに基づくフィルタリング
- Web 利用マイニング

この中で，Web 利用マイニングに基づく Web パーソナライゼーションは，基本的にユーザのアクセスログのみで解析を行うことが可能であり，ユーザによる情報の入力や，ページ・コンテンツの評価を必要とし

ない．また，クライアントサイドに特別なプログラムも必要とならないため，ユーザの環境に依存せず拡張性が高い．そこで，本研究では Web 利用マイニングに基づく Web パーソナライゼーションに焦点を当てる．利用されるデータマイニング手法の例としては，相関ルール発見，シーケンシャルパターン発見，クラスタリング，クラシフィケーションがあげられている[1]．

このうち，クラスタリングやクラシフィケーションを用いた手法は，はじめにクラスタや分類を作成する際の設定によってパーソナライゼーションの結果に大きなばらつきが生じるという問題があると我々は考えるため，相関ルールとシーケンシャルパターンに注目する．

相関ルールはアイテム間の共起性に注目し，1 セッション内でともにアクセスされたページ間関係を抽出する．シーケンシャルパターンは，相関ルールの拡張であり，セッション間の順序情報を考慮する．

しかし，これらのパターンを利用した手法[3)~6)]では，各セッション中のページアクセスの順序が考慮されないため，アクセス順序に特徴的な傾向がある場合には適切なパーソナライゼーションが行えない可能性がある．

そこで我々は，初期段階の特別な設定を必要とせず，またアクセスの順序情報を保存できる Web 利用マイニング手法として LCS（Longest Common Subsequences）に着目し，アクセスログ中のシーケンスの LCS を用いることで，アクセスパターンのぶれを吸収した概括的なアクセス順序を利用して推薦精度を向上させる手法を提案している．

これまでに，LCS を用いたアクセスログの解析手法とその解析に基づくサイト構成の改善手法[7]，その場合の LCS 抽出の効率化手法[8]を提案し，その有効性を示してきた．Web サイト内におけるすべてのセッションの URL の推移をシーケンスとして抽出し，そのそれぞれに対して他のすべてのシーケンスとの LCS を求め，頻出アクセスパターンを発見するものである．

さらに，本稿の事前検討として，上述のアクセスログ解析によって抽出された 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に関して，研究会で報告を行った[9]．LCS を用いることで，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の表現が可能になるとともに，順序情報を保持することができるため，実際の Web アクセスログを用いた実験において，相関ルールを用いる手法と比較して推薦精度が向上したという実験結果が得られている．しかし，事前検討であったため実験内容とその解析が必ずしも十分でなく，予想に反する結果も含まれていた．そこで，本稿では，アクセスログに基づく Web ページ推薦における LCS の利用効果をより詳細に解析するために，条件および設定を変更したいくつかの実験を行うと同時に，それらの実験結果から得られた知見を基に手法の改良を行い，その効果を調べる．

2. 関 連 研 究

本章では，まず，Web 利用マイニングに基づく Web パーソナライゼーションに関する既存研究について述べる．次に，本研究とは異なるアプローチで解析を行っている研究を紹介する．

教育システムのアクセスログからの相関ルールやシーケンシャルパターンの抽出を用いた学習項目推薦システム[3]では，それぞれのパターンから求めた推薦順位を融合して最終的な順位を決定し推薦を行う．しかし，その最終的な順位決定の詳細な方法については言及されていない．また，学生の学習レベル等，アクセスログ以外からの情報も利用している．

相関ルールを用いた Web ページ（URL）推薦手法[4),5)]では，アクティブユーザの現在までのアクセスページと共起頻度の高いページが推薦される．文献 4) では，パターン中で未出現のページが読み込まれるとその場でルールを作成するため，新規のページアクセスに対しても推薦が可能となるが，そのコストは大きく，またユーザは最初にブックマークの情報を提供する必要がある．文献 5) では，効率的に推薦を行うための構造が提案されているが，推薦に用いるルールは必ず最後のアクセスページを含む必要があるため，その頻度が低い場合には適切な推薦が行えない．推薦精度の評価においては，前者では模擬のデータを使用し，あらかじめカテゴライズされた URL に対し想定したカテゴリに含まれる URL を推薦できたときに “正解” として評価を行う．一方，後者では実際のアクセスログを使用し，現在までのアクセスページと推薦されたページを比較することで，その正確性と包含度合を計算している．

文献 6) では，相関ルールとシーケンシャルパターン，さらにその派生である連続シーケンシャルパターンを用いた Web ページ推薦において，精度を比較している．相関ルールを用いた手法[5]におけるマイニングのフェーズを他の 2 つのパターンに置き換えて実験を行った結果として，Web prefetching 等のタスクには連続シーケンシャルパターンが最適であり，一般的なアプリケーションには相関ルールとシーケンシャル

パターンを用いるのが適していると結論付けている．

他の Web 利用マイニング手法を採用しているものとして，クラスタリングを用いた Web ページ推薦手法[10] がある．この手法では，遺伝的アルゴリズムを用いることにより，非常に少ないパラメータで初期段階のクラスタ作成を行うことができる．しかし，類似度やメンバシップ等の計算で複数の方法があげられており，そのそれぞれにトレードオフ関係があるため，対象とするデータによって適切な方法の選択が難しいと考える．

次に，アプローチの異なる解析手法として，内容に基づくフィルタリングのためのマイニング技術である TextExtractor[11] や，ユーザの Web ページ閲覧行動の支援システムである Letizia [12] がある．TextExtractor では，ユーザが Web ページ閲覧中に行う，なぞり読みやリンククリック等の特徴的なマウス操作を利用し，ユーザプロファイルを作成する．ユーザの嗜好を評価するためのフィードバックとして，ページ内のテキスト部分を文や行の単位で利用できるが，クライアント側にプログラムを埋め込む必要がある．一方，Letizia は，Web ブラウザから得られる，ユーザがたどったリンクやブックマークの情報を基にユーザプロファイルを作成する．Web ページ（リンク）の推薦においては，その根拠となるキーワードが同時に提示される．しかし，Letizia は Web ブラウザと協調的に情報取得，推薦を行うため，この手法においても，クライアント側の特別な設定が必要となる．

協調フィルタリングに基づく手法[13] では，書籍販売サイトにおける商品推薦を目的とし，おすすめ商品を消費者にメールで通知する形で推薦を行っている．カテゴリごとに定義された興味情報を，同一カテゴリ内と異なるカテゴリ内の商品それぞれと比較するため，通常の推薦に加え意外性の高い推薦が可能となるが，入力として消費者からの各コンテンツの評価が必要となる．

3. Web アクセスログ解析への LCS の適用

本章では，我々が以前に提案した，アクセスログから LCS を抽出する方法[7),8)] 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．

3.1 LCS

リスト x の部分列とリスト y の部分列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☆．

2 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 2 つのリストの類似性が高いことを表す．

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共通する特徴を抽出することができると考える．

3.2 LCS を用いた Web アクセスログ解析

3.2.1 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出し，ユーザのセッションを抽出する必要がある．

サイト内におけるユーザごとの移動情報を得るため，IP アドレスや 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．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，Cookie を用いることが好ましいが，これ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，複数のユーザを含む可能性があるため正確性は下がるものの，IP アドレスの情報等で代用する．

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．

アクセスログの中には，画像ファイルへのアクセス等の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．そのため，セッション抽出時に，前処理[14] によりそれらの情報を取り除く必要がある．

3.2.2 LCS の抽出と頻度集計

Web アクセスログから得られた各セッションを，URL を各要素に持つシーケンスと見なし，各シーケンスについて総当たり的に LCS を抽出する．各 LCS の出現頻度の集計を行い，高頻度で出現する LCS パターンを発見する．

まず，2 つのシーケンスから LCS を求める際，LCS を求める問題と等価である SED（Shortest Edit Distance）を求めるための効率化された手法[15] を用いる．

この手法では，比較する 2 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，多くの場合で一般的な動的計

☆ LCSS と略記される場合もある．

画法よりも大幅に小さい計算量で LCS の計算が可能になる．

LCS 抽出のための計算においては，すべてのアクセスシーケンスに対して総当たりで求めるため，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[7]，本研究では，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[8]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．

4. LCS を用いた Web ページ推薦

本章では，3 章で説明した手順でアクセスログから抽出される LCS を用いたユーザにページを推薦する手法[9] について説明する．

抽出された LCS とアクティブセッションのマッチングを行って推薦候補ページを選出する．次に，それぞれのページに得点を付加することで，推薦のための優先順位を決定し，その得点が上位のページから順に推薦する．

4.1 推薦候補ページの選出と得点付け

LCS を用いた Web ページ推薦手法では，アクセスログから抽出した LCS のそれぞれと，現在まで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．

抽出された LCS のうち，全セッション中において数え上げられた回数が閾値 $min.Count$ 以上であり，かつ長さが $min.Length$ 以上である LCS の集合を Large LCS 集合と呼び，$LL = \{lcs_1, lcs_2, \cdots lcs_k\}$ で表す．また，LL 内の i 番目の要素が全セッション中で数え上げられた回数を c_i と表す．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，長さ n 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．

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我々は考える．なぜなら，lcs_i と act_n 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とらえることができるからである．

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，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部分シーケンスであり，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 は同じ傾向を示す可能性が高いと考え，高い得点を付加する．

以上を考慮したうえで，優先順位を付けて推薦を行うための手法として，WRAPL-FL（Web page Recommendation by Access Pattern Lcs with Frequency and matched Length based weighting）法を定義する．Large LCS 集合 LL と長さ n のアクティブセッション act_n に対し WRAPL-FL 法では，あるページ p の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るものとする．

$$point(p) = \sum_{lcs_i \in LL} |lcs_i \cap_p act_n| \cdot c_i^{\alpha} \quad (1)$$

ここで $\cap_p$ は，以下を満たす演算子とする．要素 p，シーケンス l，a に対し，$l \cap_p a$ は l と a の LCS であり，かつ l 内のその LCS のすべての要素より後ろに必ず p が現れるシーケンスを表す．また，α は c_i の重みであり，Web サイトの特徴からその影響度合いを考慮して適切に調節する．

4.2 ページ推薦手法 WRAPL-FL 法

以下の手順で推薦ページの決定を行う．

(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，1 番目から共通部分の最後までの要素すべてを除去する．
(3) 残ったページを推薦ページの候補とし，そのそれぞれの $point$ に $|lcs_i \cap_p act_n| \cdot c_i^{\alpha}$ を加える．
(4) LL の中のすべての要素に対して (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．

ここまでで述べた，LCS を用いた推薦手法の概要を**図 1** に示す．たとえば，ページ推薦のステップにおいて，図のように $act_3 = (A, B, C)$ が与えられたとき，$lcs_1 = (A, B, C, D, E)$ 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 $lcs_2 = (A, D, B, C)$ 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．さらに，$lcs_3 = (B, C, E)$ では E となる．したがって，この例で lcs_1〜lcs_3 から推薦されるページの候補は $\{D, E\}$ となる．

ここでは，ページ E が 2 つの LCS で推薦候補となっている．したがって，各 LCS から算出された得

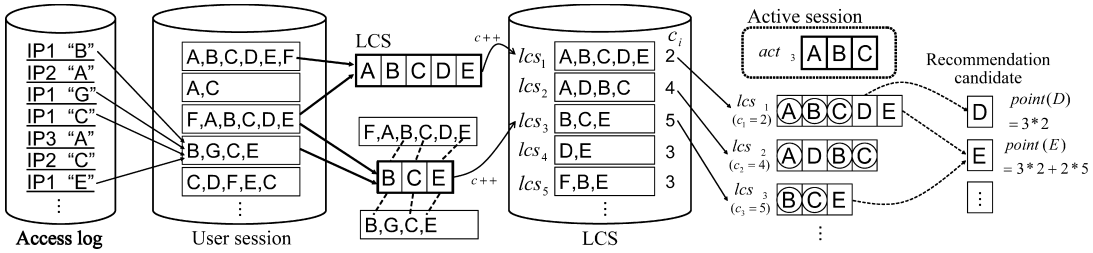


図 1 LCS を用いた推薦
Fig. 1 Web page recommendation using LCS.

点の和がページ E の得点となる．

5. 推薦精度評価のための準備

我々は，LCS を用いた推薦手法 WRAPL-FL 法の有効性を確認するために，実際のアクセスログを使って実験を行い，その効果を測定するとともに，2 章で紹介した 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手法を簡単化した手法を実装し，その推薦精度を比較した[9]．本章では，実験対象として用いたデータの説明と評価指標の定義，さらに比較手法の詳細について説明する．

5.1 実験対象データと評価指標

実験対象として，“The Internet Traffic Archive”（http://ita.ee.lbl.gov/）で配布されているいくつかの Web サイトのアクセスログのうち，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．このデータにはセッション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見なした．ページアクセスの間隔を調査した結果，600 秒程度まではその観測された回数に大きな変動が見られたが，1,200 秒程度を超えると大きな変化は確認できなかったため，1,200 秒以上間隔が空いた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．

全セッションのうち，時期が早い方を学習セットとし，残りのセッションをテストセットと見なしてページ推薦を行い，その評価を行う．

推薦精度の評価にあたり，我々は文献 5) と同様に，テストセット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はユーザが好ましいと思う Web ページであると見なす．そこで，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．

$$precision(Recom) = \frac{|Recom \cap eval|}{|Recom|} \quad (2)$$

$$coverage(Recom) = \frac{|Recom \cap eval|}{|eval|} \quad (3)$$

ここで，$Recom$，$eval$ はそれぞれ対象アクティブセッションから導かれた推薦ページの組，対象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．

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 act_n と見なして Large LCS 集合 LL 中の全要素とマッチングを行う．そこから推薦ページのランク付けを行って上位のページを推薦し，そのセッションの残りのページ $eval$ と比較することで precision と coverage を求めた．

また，学習セットとテストセットの比率を 25 : 75，50 : 50，75 : 25，87.5 : 12.5 としたそれぞれの場合で WRAPL-FL 法を評価した結果を**図 2** に示す．アクティブセッション長を 2 としたときの結果を表している．図 2 より，50 : 50，75 : 25 と 87.5 : 12.5 の場合では同等の結果が得られた一方，25 : 75 では大幅な低下が確認できる．アクティブセッション長が 3，4 の場合にも同様の結果となった．計算時間短縮のためには，学習セットは小さい方が適切であると考えるが，はじめの 50%を学習セットとした場合には，データによっては結果が悪化する可能性があると考える．そこで，

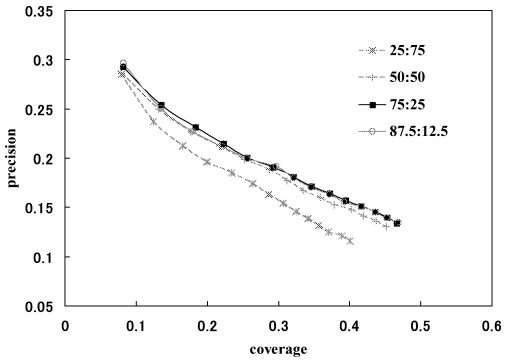


図 2 学習セットとテストセットの比率を変えた場合の比較
Fig. 2 Comparisons of coverage-precision varying ratio of training set size.

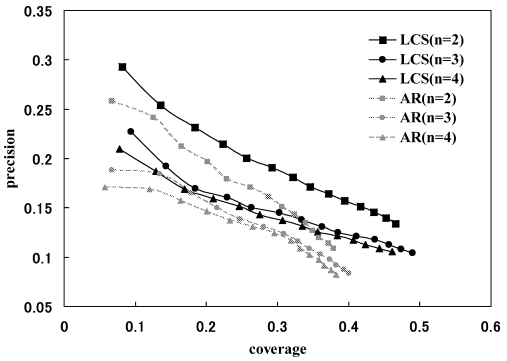


図 3 LCS を用いた手法と比較手法による推薦の評価
Fig. 3 Evaluation of methods with LCS and association rules.

以降では，学習セットとテストセットの比率を 75 : 25 として実験を行う．

5.2 比較手法

我々は，先述の理由で相関ルールを用いた手法とシーケンシャルパターンを用いた手法が比較対象として適切であると考えるが，2 章で述べたように，アクセスログの情報のみから Web ページ推薦を行い，またその方法が詳細に記述されているのは文献 5) のみであった．したがって，今回，これを比較対象として選択する．この手法では，長さ n のアクティブセッション act_n 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 個のページをすべ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とき，{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仮定に基づいている．また，アクティブセッションの長さである n の値を，推薦ページが見つかるまで下げていくという方法で coverage を上げている．

この手法では，推薦ページの順位付けを行っていないため，confidence 値で順位付けを行うこととする．

6. LCS を用いたページ推薦手法に対する解析

我々は本稿の事前検討として，前節の設定で実験を行ったが，比較手法の相関ルールを用いた Web ページ推薦手法の文献内で提案されているすべての改善手法を実装していなかったため，今回機能を追加して改めて比較を行った（6.1 節）．その結果，WRAPL-FL 法の優位性は実証できたが，アクティブセッション長が短い場合で長い場合に比べて良いという，予想に反する結果が得られた．そこで，本稿ではその原因について解析を行うため，以下の 3 点に着目した．

- 今回対象とした Web サイトでは，セッション長が短いものが大部分を占めていた．
- アクティブセッションと LCS とのマッチングにおいて，一致要素数が 1 の LCS も判断材料として用いている．
- 推薦順位の決定において，LCS の出現回数とアクティブセッションとの一致要素数のみを考慮している．

これらが推薦精度に与える影響を調べるために，6.2 節〜6.4 節では実験の条件を変更して詳細な解析を行い，LCS を用いた推薦手法の特徴について調査する．さらに，6.5 節ではそれらから得た知見を利用し，推薦精度の改善を目指す．

6.1 LCS を用いた手法と比較手法による推薦精度の比較

実験に用いたパラメータは，比較手法では一定の最小サポート値 0.008 を用い，要素数が 1 から 5 の頻出アイテムセットを生成した．LCS を用いた手法では，$min.Count$ を 150，$min.Length$ を 3 として LL を作成し，推薦ページの得点付けの式 (1) は $\alpha = \frac{1}{2}$ の場合を用いた．

それぞれについて推薦精度を測定した結果を**図 3** に示す．図中の LCS，AR はそれぞれ，LCS を用いた手法，比較手法を用いた推薦に対応しており，n はアクティブセッションの長さを表す．また，図における各点は，推薦する上位ページの数 $|Recom|$ 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 $|Recom|$ が大きい場合を表している．

図 3 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，比較手法に比べ，LCS を用いた手法で良い結果が得られていることが確認できる．

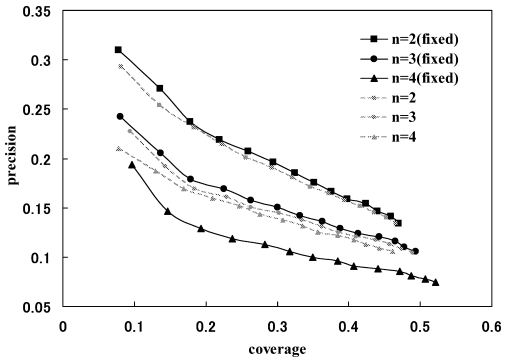


図 4 テストセッション長を 6 に固定した場合の比較
Fig. 4 Case of restricting the length of test sessions to 6.

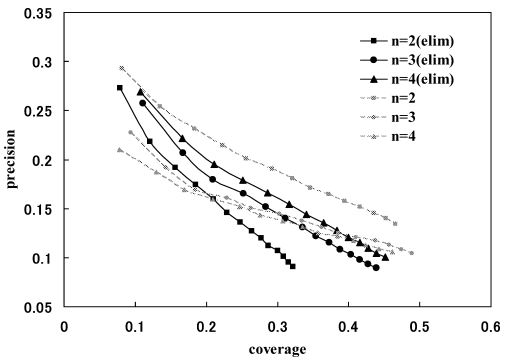


図 5 一致要素数が 1 のものを除いた場合の比較
Fig. 5 Elimination of test sessions whose matched length is 1.

しかし，一般にアクティブセッション長 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 は改善されるはずであるが，図 3 から，n が大きいものに比べ小さいもので良い結果が得られたことが分かる．

6.2 テストセッション長を固定した場合の比較

今回対象としたサイトでは，セッション長ごとのセッション数に大きな偏りがあった．セッション長が 3〜9 のテストセッションの数は順に，3189，1834，1139，717，500，354，217 という分布になっており，短いアクセスでセッションを終えるユーザが多く，長いセッションの数が相対的に少なかった．

長さ n のアクティブセッションから推薦を行う際には，テストセッション長が $n+1$ 以上である必要があることから，調査の対象となるテストセッションに差が生じた結果，等しい条件で比較が行えなかった可能性がある．そこで，ここではまず，テストセッション長を固定して実験を行うことで，可能な限り近い条件で推薦精度の比較を行うことを考える．

5.1 節と同様のデータに対し，テストセッション長を 6 に固定した場合の結果を**図 4** に示す．凡例中の n は，アクティブセッション長を表している．上の 3 つ（fixed）が固定長 6 のテストセッションを用いた場合，下の 3 つがテストセッション長を固定しない場合の結果である．

図から，$n = 2$，3 の場合には大きな差は見られないが，$n = 4$ の場合で精度が低下しているため，テストセッションの差は精度に影響を与えていることが分かる．しかし，近い条件下で精度を比較した場合にも，n が小さいアクティブセッションからの推薦でより良い結果が得られるという傾向は変わらなかった．また，テストセッション長を 5，7 に固定した場合でも，同様の傾向が得られた．

6.3 一致要素数が 1 の LCS を除いた場合の比較

WRAPL-FL 法では，一致要素数が 1 の場合（つまり $|lcs_i \cap_p act_n| = 1$ の場合）にもそれに続く推薦候補を選出し，得点付けを行っている．

しかし，一致要素数が 1 の場合，アクティブセッションと過去のパターンとのマッチ度が低いため，推薦精度に悪影響を与えてしまう可能性がある．そこで，一致要素数が 1 の場合には得点は付加しないものとしたうえで，同様に精度を測定する実験を行った．

結果を**図 5** に示す．凡例中の上の 3 つ（elim）と下の 3 つはそれぞれ，一致要素数が 1 つの LCS を除いた場合と除かない場合の結果に対応している．図から，一致要素数が 1 の場合を判断材料から取り除くと，アクティブセッションを長くとった場合で，短い場合に比べて良い結果が得られることが確認できる．

精度の順番が逆転する要因として，$n = 2$ のときに，一致要素数 1 も含めて推薦を行った場合に比べて精度が大きく低下していることがあげられるであろう．これは，act_2 と LCS のマッチングにおいて，act_2 との一致要素数が 1 である LCS が全体の約 25%を占めているのに対し，一致要素数が 2 である LCS は 2%弱にとどまっていることに起因すると考える．一方，$n = 3$，4 の結果ではほぼ全体的に，一致要素数 1 を含める場合に比べ，精度の向上が確認できた．したがって，n が大きい場合には，一般的な傾向にあてはまり，マッチの度合いが高い LCS が精度向上に影響を与えていることが分かる．

6.4 アクティブセッション中のマッチ位置による比較

多くの大規模なサイトではユーザのアクセス行動が多様であるため，アクティブセッションとの一致要素数が多い LCS は非常に少ない．そこで，一致要素数が少ない場合にも対応できるように，一致要素のアク

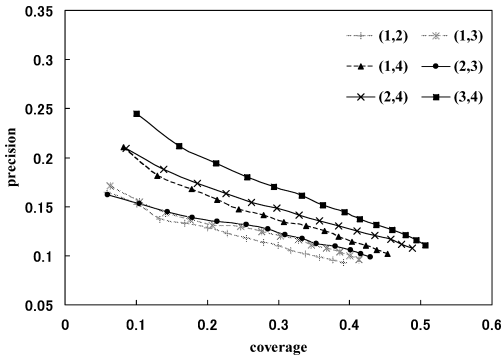


図 6 アクティブセッション中のマッチ位置による比較
Fig. 6 Influence of matched position in each active session.

ティブセッション中における位置に注目し，マッチ位置が推薦精度に与える影響を検討する．

これまでは，テストセッションのはじめの n ページを act_n と見なしてページ推薦を行い，そのセッションの残りの部分 $eval$ と比較することで precision と coverage を求めてきた．しかし今回は，セッション前半のどの部分が後半ページとの関連が深いかを調査するために，テストセッション中のはじめの 4 ページの中から任意の 2 ページのすべての組合せを順序を変えずにとり，それらに対し 5 ページ目以降の部分を評価セット（$eval$）として推薦精度を測ることとする．

結果を**図 6** に示す．凡例中の (a, b) はそれぞれ，はじめの 4 ページ中の a，b 番目のページを順に 2 ページとり，act_2 と見なした場合の結果に対応している．

図より，推薦精度は $(3,4)$ の場合で最も良く，以下は順に $(2,4)$，$(1,4)$，$(2,3)$，$(1,3)$，$(1,2)$ となっていることが確認できる．この結果から明らかに，アクティブセッションの後方のページがユーザの以降のアクセス行動と深いかかわりを持つことが分かる．したがって，推薦を行うための履歴として用いるページの中では，前方のページに比べて後方のページを重視すべきである．

6.5 アクティブセッション中の LCS とのマッチ位置の考慮

6.3，6.4 節で得られた特性を利用して，アクティブセッションと LCS とのマッチングを行う際に，それらのマッチ位置を推薦ページの優先順位付けに反映させる方法を提案する．

6.5.1 マッチ位置を考慮した得点付け

lcs_i と act_n とのマッチングの際に，アクティブセッション中におけるマッチ位置を考慮するために，次のようにマッチ位置重み l_i を定義する．前述のように，アクティブセッションとのマッチ位置が後方にある LCS の方がより重要であるため，l_i は後方ページが重視されるように設定する必要がある．

そこで，act_n と lcs_i を比較し，act_n の m ページ目が lcs_i とマッチした場合，l_i に m を加算する．たとえば，$act_4 = (A, B, C, D)$，$lcs_i = (B, D, E)$ のとき，act_4 中の 2，4 番目のページ B と D が lcs_i と一致するため，$l_i = 2 + 4 = 6$ となる．

このようにして得られた l_i を，WRAPL-FL 法による推薦ページの優先順位付けの式 (1) に掛け合わせることで新たな得点付けの式を以下で定義し，これを用いた推薦手法を WRAPL-FLP（Web page Recommendation by Access Pattern Lcs with Frequency, matched Length and Position based weighting）法と呼ぶ．

$$point(p) = \sum_{lcs_i \in LL} |lcs_i \cap_p act_n| \cdot c_i^{\alpha} \cdot l_i^{\beta} \qquad (4)$$

ただし，β は l_i の重みである．

6.5.2 WRAPL-FLP 法による順位付けの評価

WRAPL-FLP 法による改良の有効性を確認するため，LCS を用いたページ推薦手法において，WRAPL-FL 法，WRAPL-FLP 法を用いた場合の推薦精度の比較を行った．

図 7，**図 8**，**図 9** はそれぞれ，長さ n のアクティブセッションからのページ推薦において，WRAPL-FLP 法を採用した場合の結果を表している．凡例の β は，式 (4) における一致位置に応じた得点 l_i の重みを表す．すなわち，$\beta = 0$ は WRAPL-FL 法に対応している．

グラフより，アクティブセッションにおける LCS とのマッチ位置を考慮することで，結果が改善されることが確認できる．

6.6 考　　察

評価のための指標となる precision や coverage は実験対象とするデータセットに大きく依存するため，他の文献で得られた値と直接比較を行うことはできないが，図 3，7〜9 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，precision と coverage の双方において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確認できる．このことから，相関ルールを用いて共起頻度のみを考慮するよりも，順序情報を用いた方が，よりユーザが好ましいと思うページを正確に推薦できたと考える．

このように，相関ルールを用いた手法に比べ本手法で良い結果が得られる理由として，ユーザのアクセス順序とその出現頻度に特徴的な傾向がある場合を想定する．たとえば**図 10** のように，A，B，C の各ペー

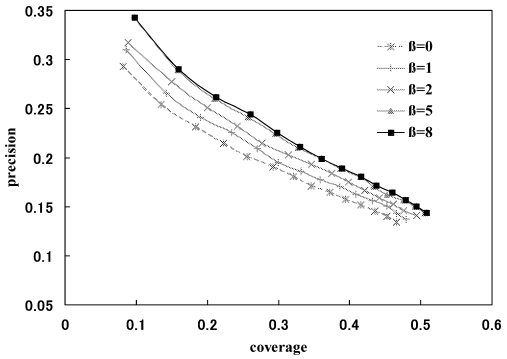


図 7 一致位置の考慮による改良（n=2）
Fig. 7 Improvement by effect of matched position (n=2).

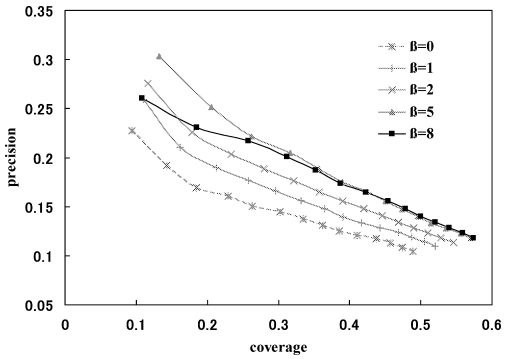


図 8 一致位置の考慮による改良（n=3）
Fig. 8 Improvement by effect of matched position (n=3).

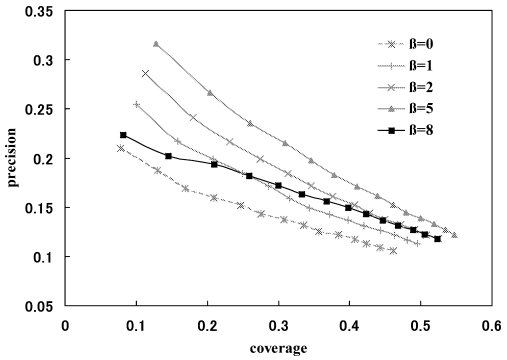
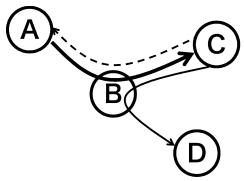


図 9 一致位置の考慮による改良（n=4）
Fig. 9 Improvement by effect of matched position (n=4).

ジに対し，アクセスの順序に右表のような頻度の違いがある場合，アクティブユーザが C → B の順にアクセスすると，C → B に続いてアクセスされる可能性が高いのはページ D であり，WRAPL ではページ D が推薦される．しかし，相関ルールを用いた手法では B，C との共起頻度が高いページ A が推薦されてしまうため，推薦精度に差が生じたと考える．



User session	count
A → B → C	100
C → B → D	50
C → B → A	5

図 10 Web ページ推薦におけるアクセス順序の考慮
Fig. 10 An example of skewd accessing-order frequency.

アクティブセッション中の LCS とのマッチ位置の考慮について，図 8，9 のアクティブセッション長が 3，4 の場合では，$\beta = 8$ の時精度が低下している．これは，後方ページの影響度合いが強くなりすぎることで順序情報等が考慮されなくなり，精度が低下したと考察する．したがって，これらはアクティブセッション長が 2 の場合に比べ，LCS によって表現されるアクセスの順序情報が推薦に良い影響を与えていると考える．しかし，図 7〜9 では全体として，得点付けの式において l_i の重み（$= \beta$）をかなり大きくしても精度の向上が確認できるため，今回対象としたサイトでは，アクティブセッション中で後方に現れるページはその後のアクセスに対して大きな関連性を持つことが分かる．

WRAPL-FL 法を用いた場合に比べ，WRAPL-FLP 法を用いた場合にはその差は狭まるものの，やはりアクティブセッション長を短くとった方が推薦精度が良い．一方で，アクティブセッションとの一致要素数が 1 の LCS を除いた場合にアクティブセッション長が長いもので良い結果が得られた．このことから，今回対象とした Web サイトでは，直前のアクセスページ（実際にリンクが張られているページ）からの推薦が最も精度が良いという特徴があると考える．現在この Web サイトは存在しないためリンク構造等を解析することはできないが，ページ配置が細分化，階層化されており URL 数が非常に多く，また短いセッションが非常に多いことから考えて，目的のページまで迷わずにナビゲートをするユーザが多いためにこのような傾向が現れると推測する．

7. おわりに

本稿では，アクセスログ解析によって抽出した LCS を用い，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推薦する手法である WRAPL について解析を行った．

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 WRAPL-FL 法を用いて推薦するシミュレーションを行った．

その際，条件を限定して実験を行うことにより解析し，手法の問題点や改善点について検討した．さらに，それらから得た知見を活かして推薦ページの優先順位付け方法を改良した WRAPL-FLP 法を提案し，それによって推薦精度が向上することを確認した．

本研究の今後の課題について述べる．まず，本研究の最終的な目的は，Web サイトの利用者に対し，リアルタイムに Web ページを推薦することであるため，今後は，計算量の見積りや他手法との比較，さらに，計算量を削減するための手法についても検討する必要がある．

今回のシミュレーションで採用した評価方法では，アクティブセッションより後にアクセスされたページすべてを正解としたため，サイト内で実際にリンクが張られており，直後にアクセスされやすいページを推薦し，それが正解とされるケースが多かった．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おいては，サイト構造における深さ等を考慮すべきであると考える．したがって，正解ページにも優先順位を付け，それに応じて評価する等，目的に合わせて評価方法も工夫すべきであろう．

また，Web サイトの規模や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，それに適した推薦ページの優先順位付けの方法について再考する必要がある．本手法を他の Web サイトに適用し，サイトの特徴と推薦精度の関連性を調査するとともに，手法の主観的な評価を行うことも今後の課題である．

さらに，LCS を利用したアクセスログ解析の持つ特徴に関する詳細な考察を行うために，似た概念を持つ他のモデルとの比較を行う必要がある．具体的には，不確実性のもとでの予測や意思決定に用いられる確率モデルであるベイジアンネットワーク[16)~18)] を想定する．LCS とベイジアンネットワークには概念としては類似する部分があるものの，実現の手順や方法には大きな差があるため，今後は，双方の相違点について詳細に調査，比較を行っていきたい．

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 の助成により行われた．

参 考 文 献

1) Eirinaki, M. and Vazirgiannis, M.: Web mining for web personalization, *ACM Trans. Internet Techn.*, Vol.3, No.1, pp.1–27 (2003).
2) 大塚真吾，喜連川優：Web アクセスログとその利活用，人工知能学会誌，Vol.21, No.4, pp.410–415 (2006).
3) Kristofic, A. and Bieliková, M.: Improving Adaptation in Web-based educational hypermedia by means of knowledge discovery, *Proc. 16th ACM Conference on Hypertext and Hypermedia*, pp.184–192 (2005).
4) Fu, X., Budzik, J. and Hammond, K.J.: Mining navigation history for recommendation, *Intelligent User Interfaces*, pp.106–112 (2000).
5) Mobasher, B., Dai, H., Luo, T. and Nakagawa, M.: Effective personalization based on association rule discovery from Web usage data, *Proc. 3rd Intl. Workshop on Web information and data management*, pp.9–15 (2001).
6) Mobasher, B., Dai, H., Luo, T. and Nakagawa, M.: Using sequential and non-sequential patterns in predictive Web usage mining tasks, *Proc. IEEE International Conference on Data Mining* (*ICDM'02*), pp.669–672 (2002).
7) 宇根田純治，横田治夫：Web ログの共通シーケンス解析，信学技報 DE2002-2，電子情報通信学会 (2002).
8) 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).
9) 山元理絵，小林 大，小林隆志，横田治夫：Web アクセスログの LCS を用いた Web ページの推薦手法，信学技報 DE2006-40，電子情報通信学会 (2006).
10) Nasraoui, O. and Petenes, C.: An Intelligent Web Recommendation Engine Based on Fuzzy Approximate Reasoning, *Proc. IEEE International Conference on Fuzzy Systems*, pp.1116–1121 (2003).
11) 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，Vol.19, No.3, pp.365–372 (2004).
12) Lieberman, H.: Letizia: An agent that assists Web browsing, *Proc. 15th International Joint Conference on Artificial Intelligence* (*IJCAI '95*) (1995).
13) 寺野隆雄：Web 上の情報推薦システム，情報処理，Vol.44, No.7, pp.696–701 (2003).
14) Banerjee, A. and Ghosh, J.: Concept-based clustering of clickstream data, *Proc. 3rd Intl. Conf. on Information Technology*, pp.145–150 (2000).
15) Wu, S., Manber, U., Myers, G. and Miller, W.: An O(NP) sequence comparison algorithm, *Information Processing Letters*, Vol.35, No.6,

pp.317–323 (1990).
16) Jensen, F.V.: *An Introduction to Bayesian Networks*, New York: Springer-Verlag (1996).
17) 本村陽一：ベイジアンネットワーク，信学技報 NC2003-38，電子情報通信学会 (2003).
18) 本村陽一：ベイジアンネットソフトウェア，人工知能学会論文誌，Vol.17，No.5，pp.559–565 (2002).

(平成 18 年 9 月 15 日受付)
(平成 19 年 2 月 27 日採録)

(担当編集委員　石川 博，有次 正義，片山 薫，
木俵 豊，中島 伸介)

山元　理絵

2005 年東京工業大学工学部情報工学科卒業．同年同大学大学院情報理工学研究科計算工学専攻修士課程入学，現在に至る．データ工学の研究に従事．

小林　　大

2003 年東京工業大学工学部情報工学科卒業．2005 年同大学大学院情報理工学研究科計算工学専攻修士課程修了．現在，同専攻博士後期課程在学中．2006 年より日本学術振興会特別研究員（DC2）．並列ストレージシステムの自律管理に関する研究に従事．日本データベース学会学生会員．

吉原　朋宏

2005 年東京工業大学工学部情報工学科卒業．同年同大学大学院情報理工学研究科計算工学専攻修士課程入学，現在に至る．データ工学の研究に従事．

小林　隆志（正会員）

1997 年東京工業大学工学部情報工学科卒業．1999 年同大学大学院情報理工学研究科計算工学専攻修士課程修了．2004 年同専攻博士課程修了．2002 年同大学学術国際情報センター助手．2007 年より名古屋大学大学院情報科学研究科特任准教授，現在に至る．工学博士．ソフトウェア設計方法論，ソフトウェア再利用技術，複合メディアコンテンツの管理・検索，Web サービス連携等の研究に従事．日本ソフトウェア科学会，電子情報通信学会，日本データベース学会，ACM 各会員．

横田　治夫（正会員）

1980 年東京工業大学工学部電子物理工学科卒業．1982 年同大学大学院理工学研究科情報工学専攻修士課程修了．同年富士通（株）．同年 6 月（財）新世代コンピュータ技術開発機構研究所（ICOT）．1986 年（株）富士通研究所．1992 年北陸先端科学技術大学院大学情報科学研究科助教授．1998 年東京工業大学大学院情報理工学研究科助教授．2001 年東京工業大学学術国際情報センター教授．工学博士．主として分散インデキシング，データ工学向けアーキテクチャ，高機能ストレージシステム，ディペンダブルシステム等に関する研究に従事．日本データベース学会理事．電子情報通信学会，人工知能学会，IEEE，ACM 各会員．