論文 / 著書情報
Article / Book Information

論題(和文)	大規模商用サイトログを用いたWebページ推薦手法 WRAPL の評価と考察
Title(English)	Applying Web Pagfe Recommendation Methods to Access Logs of a Commercial Site
著者(和文)	山元理絵, 吉原朋宏, 小林大, 小林隆志, 横田治夫
Authors(English)	Rie YAMAMOTO, Tomohiro YOSHIHARA, Dai KOBAYASHI, Takashi KOBAYASHI, Haruo YOKOTA
掲載誌(和文)	DEWS2007論文集
Citation(English)	Proceedings of DEWS2007
Vol, no, pages	Vol. , No. , pp. L4-1
発行日 / Pub. date	2007, 3
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2007 Institute of Electronics, Information and Communication Engineers.

大規模商用サイトログを用いた Webページ推薦手法WRAPLの評価と考察

山元　理絵[†]　　吉原　朋宏[†]　　小林　　大[†,††]　　小林　隆志[†††]　　横田　治夫[†††,†]

† 東京工業大学大学院情報理工学研究科計算工学専攻　〒 152–8552 東京都目黒区大岡山 2–12–1
†† 日本学術振興会特別研究員 DC
††† 東京工業大学学術国際情報センター　〒 152–8550 東京都目黒区大岡山 2–12–1
E-mail: †{yamamoto,yoshihara,††daik}@de.cs.titech.ac.jp, †††tkobaya@gsic.titech.ac.jp, †††,†yokota@cs.titech.ac.jp

あらまし　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，Web パーソナライゼーション が注目され，中でも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，またユーザの匿名性を保持したまま利用できるため有用である．我々はこれまでに，Web アクセスログから LCS（Longest Common Subsequences）を抽出して Web ページ推薦に利用する手法である WRAPL を提案してきた. 本稿では，実際の大規模商用サイトのアクセスログを用いて，WRAPL の有効性を評価し考察する．

キーワード　Web，アクセスログ解析，Longest Common Subsequences，Web パーソナライゼーション，Web ページ推薦

Applying Web Pagfe Recommendation Methods to Access Logs of a Commercial Site

Rie YAMAMOTO[†], Tomohiro YOSHIHARA[†], Dai KOBAYASHI[†,††], Takashi KOBAYASHI[†††], and Haruo YOKOTA[†††,†]

† Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology　2–12–1 Ookayama, Meguro-ku, Tokyo 152–8552, Japan
†† Research Fellow (DC), Japan Society for the Promotion of Science
††† Global Scientific Information and Computing Center, Tokyo Institute of Technology 2–12–1 Ookayama, Meguro-ku, Tokyo 152–8550, Japan
E-mail: †{yamamoto,yoshihara,††daik}@de.cs.titech.ac.jp, †††tkobaya@gsic.titech.ac.jp, †††,†yokota@cs.titech.ac.jp

Abstract　Sophisticated websites satisfying users' requirements 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 requirements. We have proposed 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 of a large-scale commercial site.

Key words　Web, Access Log Analysis, Longest Common Subsequences, Web personalization, Web Page Recommendation

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を行う Web パーソナライゼーション [1] が注目され，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．

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．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]．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) の手順に

おいて，Web 利用マイニングを用いる方法では，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といった利点がある．

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．Web パーソナライゼーションのための Web 利用マイニング方法として, [1] では， 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を挙げている．

しかし，相関ルールやクラスタリングを用いた既存の研究 [3]〜[7] では，ページアクセスの順序情報を考慮していないため，ページ参照の順序に特徴的な傾向がある場合など，すでにアクセスしたページを推薦したり，ユーザにとってもはや不要となったページを推薦することで，推薦精度を低下させてしまう可能性がある．

我々は，順序情報を考慮しないという上記の問題を解決するため，アクセスログ中のシーケンスの LCS（Longest Common Subsequences）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きた [8], [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，Mobasher らの相関ルールを用いる手法 [5], [6] と比較して推薦精度が向上した実験結果が得られている．

しかしながら，我々のこれまでの報告では，NASA の Web サイトの 1995 年 8 月という古いアクセスログに対して実験を行っていたため，以下のような問題があった．

- 動的な Web ページの構築方法の普及により Web サイト構成の傾向が異なる．
- 検索エンジンの普及によりアクセス傾向が異なる可能性がある．
- 対象ページ群が既に存在しないため実際に推薦されるページがユーザにとって有用かどうかを評価できない．

そこで本稿では，現存する大規模商用サイトの実際のアクセスログに対し， 我々の提案する LCS を利用した Web ページ推薦手法 WRAPL （Web page Recommendation by Access Pattern Lcs）を適用することで，その有用性と推薦される結果について評価し考察する．また，実際の Web サイトが持つ特徴を調査し，アクセス傾向や推薦されるページの特徴などを把握することで，Web サイトの持つ特徴を考慮した推薦対象を検討する．

以降ではまず，次節で関連研究について述べ，3. 節において，我々の提案手法である WRAPL に関して説明する．4. 節で，大規模商用サイトの実際のアクセスログに対し WRAPL を適用する実験で用いるデータの説明と，その結果に対する考察を行った後，5. 節で本稿のまとめと今後の課題に関して述べる．

2. 関 連 研 究

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. [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. [5] では，アクティブユーザの現在までのアクセスページに対し，段階的に相関ルールと比較することによって，共起頻度の高いページを効率的に推薦できることが示されているが，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．

[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．

クラスタリングを用いた Web ページ推薦手法 [7] では，遺伝的アルゴリズムを用いることにより，少ないパラメ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挙げられており，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．

3. WRAPL: アクセスログから抽出した LCS を利用した Web ページ推薦

本節では，我々がこれまでに提案してきた Web アクセスログから抽出した LCS を用いてユーザに Web ページを推薦する手法である WRAPL [8], [9] について説明する．WRAPL では，推薦のための準備として，ある一定期間中のアクセスログを学習のためのデータとして用い，3. 1 節で述べる手順で LCS を抽出する．推薦においては，抽出された LCS を用いて，3. 2 節のようにして推薦ページを決定する．

3. 1 Web アクセスログからの LCS 抽出

3. 1. 1 Web アクセスログと 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 と略記する(注1)．

これをアクセスログから抽出した， ユーザのアクセスシーケンス群に適用することで，寄り道等の余分な情報を取り除いた，

(注1)：LCSS と表現される場合もある．

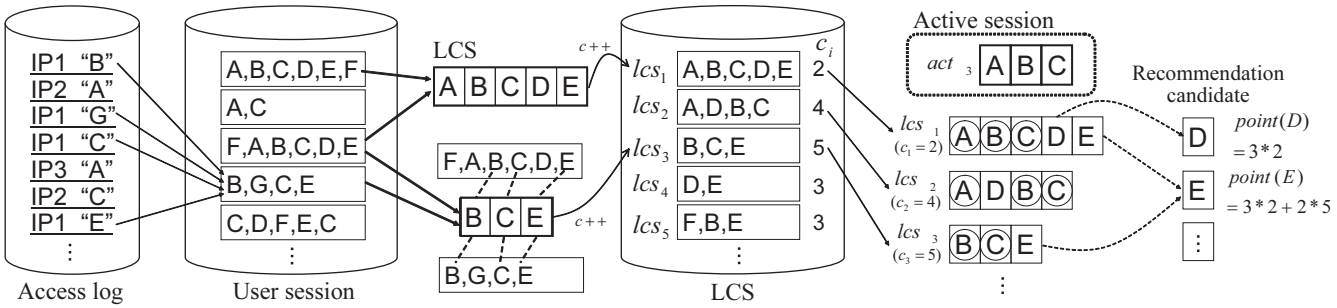

図 1　Web アクセスログの LCS を利用した Web ページ推薦

共通の傾向を発見することが可能となる．

3.1.2　ユーザセッションと LCS の抽出

Web 利用マイニングを行うためには，まず蓄積されている未加工のアクセスログを精錬して，サイト内におけるユーザ毎の移動情報である，ユーザセッションを抽出する必要がある．

ユーザセッションは，各セッションでユーザがアクセスした URL のシーケンスであり，各セッションに一意なセッション ID を割り当て，アクセスログから，セッション ID ごとに整理された URL の集合を時系列順に抽出することで作成する．

セッション ID には，Cookie を用いることが一般的ではあるが，Cookie を使用できない環境や複数のサーバをまたがる場合には，クライアントの IP アドレスなどを利用する．

このように取り出されたユーザセッション中の任意の 2 つのセッションから LCS を求める問題は，SED（Shortest Edit Distance）を求める問題に還元することができる．SED は，長さ M，N のシーケンス A，B の各要素を X，Y 軸上に並べたときの，(0,0) から (M,N) までの最短距離として求められる．移動距離が 0 となる部分の要素のみを抜き出したものが LCS であるため，我々は，この問題に関して効率化された手法 [10] を用いる．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，単純に動的計画法を用いて計算する場合に比べ，大幅に小さい計算量での LCS の抽出が可能になる．

これらの処理を，Web アクセスログから得られた全てのセッションの全組み合わせに対して行い，各 LCS の出現頻度の集計を行うことで，高頻度で出現する LCS パターンを発見する．

この計算は，全てのユーザセッションの全組み合わせに対して行うため，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 [11]．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 [12] を用いることで，LCS 抽出にかかわる計算量をさらに抑えている．

3.2　LCS を利用した Web ページ推薦

3.2.1　推薦候補ページの選出と推薦ページの決定

WRAPL 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 の内，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 は，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に共通する傾向を表す．そのため，lcs_i とアクティブセッションを比較して傾向の類似性を調べる際，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 はそのアクセスの特徴を表現しているとみなすことができる．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．

以上を踏まえ，WRAPL では，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，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 $\{p_1, p_2, \cdots, p_k\}$ を推薦ページの候補とし，そのそれぞれの *point* に $f(lcs_i, act_n, p_j)$ を加える．

（4）　LL 中の全ての要素に対して（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．ただし，$f(lcs_i, act_n, p_j)$ はある候補ページ p_j に対する得点計算のための関数とし，具体的な計算式については 3.2.2 節で述べる．

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．例えば，ページ推薦のステップにおいて，図のように $act_3 = (A, B, C)$ が与えられた時，$lcs_1 = (A, B, C, D, E)$ 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 から算出された得点の和がページ E

表 1　検索条件等のパラメタを含む URL の考慮

	内容
考慮	エリア（地方，県，都市），表示様式（件数，ページ），絞込条件（業態，場所，予算，距離），検索タイプ，特集テーマ，路線検索 [路線，駅]，地図検索 [大／小]
除外	表示様式（表示順），絞込み条件（日時，詳細条件），リンク元，文字コード

の得点となる．

3.2.2　推薦候補ページの順位決定手法

推薦候補ページの優先順位を決定する方法として，我々はこれまでに，WRAPL-FL 法 [8] と，その拡張である WRAPL-FLP 法 [9] を提案してきた．ここでは，WRAPL-FLP 法を説明する．

推薦順を決定するための推薦候補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と考える．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アクセスパスの部分シーケンスであり，重視すべきである．また前述のように，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と考え高い得点を付加する．さらに，アクティブセッション中の後方ページは，それに続くアクセスと関連性が高い [9] ため，後方ページが一致する LCS を重視する．

以上の 3 点を考慮した上で，各候補ページの得点計算のための方法として，FLP（Frequency, matched Length and Position 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 FLP 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．

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

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 の全ての要素より後ろに必ず p が現れるシーケンスを表す．l_i はマッチ位置重みを表し，アクティブセッション中の後方ページが lcs_i と一致する場合に大きくなるように設定する．また，α，γ はそれぞれ c_i と l_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.　大規模商用サイトにおけるアクセスログの解析

本節では，大規模商用サイトの実際のアクセスログに対し，WRAPL を適用し，その有効性を評価する実験に関して説明する．また，推薦結果やサイトの持つ特徴に対して考察を行った後，異なるアプローチについて検討する．

4.1　実験対象データ

実験対象として，商品紹介を行う大規模商用サイト “ぐるなび”（http://www.gnavi.co.jp/）の Web サイトにおける，2006 年 10 月 16 日から 10 月 22 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．

おおまかなサイトの構造としては，トップページの中に，特集やコンテンツページ，エリア別ページや，検索ページなどへのリンクがあり，そこから各レストランページへの直接のリンクが張られている．検索結果の表示ページなどに代表される動的ページへのアクセスが多く，その場合，検索条件によって URL は変化するため，ユーザのアクセス行動は非常に多様であった．我々のこれまでの報告では，動的に生成されるページへのアクセスは解析対象の URL から除外したが，今回は，これらのページも対象とした．

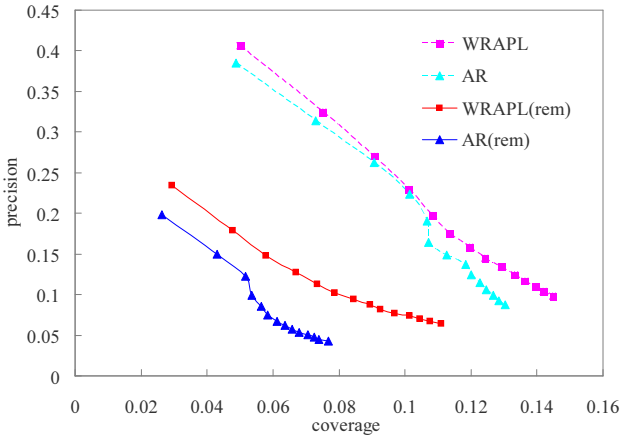

図 2　WRAPL-FLP 法と相関ルールを用いた手法の比較

サイトの構造から，各々のレストランページへのアクセスに比べ，トップページや検索ページ等の上位階層の Web ページへのアクセスが圧倒的に多い．そのため，無数に存在するレストランページの推薦を目的とする場合，条件を限定せずに十分な量のパターンを抽出するためには，膨大な量のアクセスログの解析が必要となる．

そこで今回は，解析対象から各レストランページを除外し，特集ページの推薦を目的として，特集ページへのアクセスを 1 回以上含むセッションのみを対象とし，推薦の評価に利用できない長さが 3 以下のセッションも除外した．

また，固有 URL 数を削減するため，路線検索や地図検索などのページにおけるパラメタや，検索ページ URL 中に含まれる，cgi に対するパラメタの内，不要なものを除外することを考える．全てのパラメタを考慮するのではなく，出現頻度が高く，またユーザの嗜好やアクセスの傾向を反映し得ると予測できるパラメタのみに限定し，それ以外のものを除外することで，URL を再構成した（表 1）．その結果，ログ全体の中に出現する固有 URL 数は 7 日分で 77,007 となった．

4.2　WRAPL の適用とその結果

4.1 節のデータに WRAPL-FLP 法を適用し，相関ルールを用いた Web ページ推薦手法 [5] との比較を行った．

4 日分のアクセスログ（10 月 16 日～19 日，19,974 セッション）を学習セットとしてそこから LCS を抽出し，また，続く 3 日分のアクセスログ（10 月 20 日～22 日，14,951 セッション）をテストセットとして評価を行った(注2)．

(注2)：19 日と 20 日にまたがる 13 セッションは評価対象から除いた．

4 日分のログから，WRAPL では，抽出した LCS の内カウントが 100 以上のものを Large LCS 集合として推薦に利用し，相関ルールを用いた手法では，最小サポート値を 0.005 としてアプリオリアルゴリズムで相関ルールを抽出したものを利用した．

推薦精度の評価に当たり，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 ページであるとみなす．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 を用いた．precision は適合率であり，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 は再現率であり，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．

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

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

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．

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 とみなして，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順位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．

アクティブセッション長を 2 とした場合の結果を図 2 に示す．凡例中の WRAPL，AR はそれぞれ，提案手法，比較手法である相関ルールを用いた手法を用いて推薦を行った場合の結果に対応している．グラフ中の各点は，推薦する上位ページの数 |*Recom*| を 1 から 14 の間で変化させた時の結果に対応しており，右の点ほど |*Recom*| が大きいを場合を表している．ここで，相関ルールを用いた手法では，act_n 中に同じ URL が含まれる場合には推薦ができず，また act_n 中に含まれるページを推薦することもない．したがって，今回の実験では，条件を揃えて比較を行うために，セッションの初めから重複のない n ページを取って act_n とし，WRAPL では act_n 中に含まれるページは推薦しないこととする．

図 2 からは，WRAPL と既存手法の間で優位な差は確認できなかった．そこで，実際に推薦されているページを比較したところ，どちらの場合においても，アクセス頻度が高い，トップページ等のインデックスページ（以降では，これらのページをナビゲーションページと呼ぶ）が推薦ページの上位に入ることが多かった．我々は，ナビゲーションページばかりが推薦されることは，ユーザにとって有益ではないと考えるが，これらのナビゲーションページを基点として各特集ページへ移動し，比較検討を行うユーザ（セッション）が多いことを考慮すると，これらのページは解析対象から除くべきではないと判断する．そこで，LCS の抽出と推薦ページの得点付けはこれまでと同様に行い，特にアクセス頻度の高い「トップページ」,「関東版トップページ」,「関西版トップページ」,「東京版トップページ」,「大阪版トップページ」の 5 つが推薦ページに含まれる場合に，それらを推薦候補から除外し，|*Recom*| + 1 位以下のページを繰り上げて推薦する方法についても推薦精度を計測した．結果を図 2 中の実線（rem）で示す．この結果，トップページやエリア毎のトップページなどの頻出ナビゲーションページを *Recom* から除去することで，推薦の精度が低下することが分かる．

4.3 抽出された LCS の特徴とその利用効果

以下では，抽出された LCS にはどのようなものがあり，またどのような特徴があるかについて述べる．さらに，ユーザにとってより有用な情報の推薦を目的とし，その特徴を利用することで期待できる効果について議論する．

特徴的なアクセスパターンを発見するために，前節の実験で用いた 7 日分のセッションから，1,504,887 本の LCS を抽出した．

得られた LCS を調査したところ，長さが 3〜5 程度のものが大部分であった．また，上位階層への後戻りや，検索ページにおいて固有 URL 数を削減するために除外したパラメタの影響で，同じページを複数回含む LCS も多く見られた．

LCS カウントは，トップページを 3 回繰り返すパターンが最大の 15,297,317 となり，それ以降の分布には大きな偏りがあった．全体的にカウントの小さい LCS が大部分を占めており，LCS カウントが 100 以上のものは，全体の 1.3%程度であった．

アクセスの傾向を見ると，トップページ ↔ 特集／検索ページ ↔ レストランページ間の往復を行うユーザが非常に多く見られた．これは，条件に合うレストランページの閲覧や比較を行うために多くのユーザがこのようなアクセス行動を取ったためであると考える．したがって，トップページや特集／検索ページ等のナビゲーションページのアクセス頻度が高くなる．

WRAPL 等の，Web アクセスログのマイニングを利用した推薦手法では，多くの場合，パターンの出現頻度に応じて推薦ページの順位付けが行われるため，複数のパターンに頻繁に含まれるページ，すなわちナビゲーションページ等のアクセス頻度の高いページが推薦されやすくなる．一方で，アクセスパターンに特徴的な傾向があるものの，アクセス頻度そのものが低い Web ページは推薦されにくい．

これまでの評価では，ユーザに常に何らかのページを推薦することを前提に，各テストセッションに対し，アクティブセッション長に達したとき必ず推薦を行い，それを評価した．この場合，そのセッションの特徴が出現する前にナビゲーションページ等へのアクセスを含むことで適切な推薦が難しかったり，また，セッションが特徴的なアクセスパターンを持つときにも，それに合致するルールの後半部分にナビゲーションページが含まれる場合には，それらのページばかりが推薦されてしまう可能性がある．そのため，ルールの持つ特徴を反映し，適切な Web ページ推薦を行うためには，特徴的な傾向を表す推薦ルールと同様の傾向を示すユーザのみに対し，そのルールを用いて推薦する，絞込み推薦を行うべきであると考える．

そのような方式において，提案手法で利用している LCS が有効であるかを議論するために，抽出した LCS から得られた特徴的なアクセスパターンを基に，アクティブユーザが，各 LCS

表 2　LCS，相関ルールそれぞれのヒット率の比較

ユーザの特徴	推薦されるページ	ヒット率（順序あり）	ヒット率（順序なし）
関東版トップ (→) 女性が喜ぶレストラン特集	デートで行くお店特集	0.0977	0.0797
関東版トップ (→) デートで行くお店特集	女性が喜ぶレストラン特集	0.0762	0.0565
ぐるなび食べ放題トップ (→) 焼肉・ステーキ特集	今月の食べ放題特集	0.2676	0.2112
ぐるなび食べ放題トップ (→) 今月の食べ放題特集	焼肉・ステーキ特集	0.0675	0.0721
ぐるなびシニアトップ (→) ゆっくりと落ち着けるお店特集	今月のおすすめレストラン特集	0.3913	0.3600
ぐるなびシニアトップ (→) 今月のおすすめレストラン特集	ゆっくりと落ち着けるお店特集	0.1333	0.1250
トップ (→) カップル向け個室特集	落ち着いた大人の個室特集	0.1124	0.0967
トップ (→) 落ち着いた大人の個室特集	カップル向け個室特集	0.0857	0.0828
トップ (→) 記念日&誕生日に行きたいお店特集	女性が喜ぶレストラン特集	0.0369	0.0172
トップ (→) 女性が喜ぶレストラン特集	記念日&誕生日に行きたいお店特集	0.0629	0.0597
関東版トップ (→) デートで行くお店特集	記念日&誕生日に行きたいお店特集	0.0714	0.0484
関東版トップ (→) 記念日&誕生日に行きたいお店特集	デートで行くお店特集	0.0725	0.0484

中の初めの数ページへアクセスしたとき，それに続くページをユーザに推薦できるかどうかを検証する．

評価に際し，7 日分から抽出した LCS を調査し，長さ 3 の LCS の中から，特集ページを 2 ページ以上含み，かつ 3 ページ目が特集ページであるものを選択し，特徴的なルールとみなして利用した．

LCS を想定した場合の実験として，先述の 3 日分（10 月 20 日〜22 日）のテストセッションの内，長さが 3 以上のもの（14,371 セッション）の中で，該当する LCS の初めの 2 ページが順番通りにアクセスされているセッションに対象を限定し，その中で初めの 2 ページに引き続き，該当する LCS の最後の特集ページがアクセスされているセッションの割合（ヒット率）を求めた．また，相関ルールを想定して，順序を考慮せずに初めの 2 ページを含むセッションを対象とし，同様の実験を行った．結果を表 2 に示す．なお，表中のユーザの特徴欄の (→) は LCS の場合の順序を示す．相関ルールの場合には，これらの 2 ページのアクセス順は考慮されていない．

表から，大半は 10%前後のヒット率であることが確認できる．また，LCS に対応するアクセスの順序を考慮した場合のヒット率は，相関ルールに対応する順序を考慮しない場合のヒット率よりも，1 件を除いて若干ながら良いという結果が得られている．

さらに,「ぐるなびシニアトップ」から，ぐるなびシニア以下の「今月のおすすめレストラン」と「ゆっくりと落ち着けるお店」へのアクセスや,「ぐるなび食べ放題トップ」から，ぐるなび食べ放題以下の「今月の食べ放題特集」と「焼肉・ステーキ特集」へのアクセスのように，アクセス順によってヒット率が大幅に変わるルールが存在することも確認できた．

このような場合，単なる共起関係では区別をすることはできず，順序の情報が必要となる．本稿での実験では，2 ページのアクセスから推薦を行うため，相関ルールとの差は顕著ではないが，2，3 ページ目の順序が入れ替わることでヒット率に大きな差が生じるルールがあることから考えても，推薦に至るページ数が増えることで，LCS と相関ルールの差は明確になると考える．

4.4 考　　察

4.2 節の実験では，ナビゲーションページを推薦候補から除去した結果，推薦精度が悪くなるという結果を得た．トップページや地域トップページなどのナビゲーションページがシーケンスの途中に含まれることは，例えば，一旦トップページに戻ってから，別のページに移動するなど頻繁に起こることであり，我々はそのような行動は推薦の際の特徴として捉えるべき行動であると考えていた．しかしながら，近年検索サイトの普及により，ユーザはトップページからトップダウンにレストランページを探し始めるだけではなく，レストランページから閲覧をはじめ，各階層のナビゲーションページまで戻って，サイト構造を確認しながら閲覧する場合がある．また，今回対象としたような，多数の商品を紹介する商用サイトの場合，商品を比較検討する段階で，ナビゲーションページや検索ページに頻繁に戻ってくることなどから，ナビゲーションページ含むルールが多く抽出されてしまい，結果として，ナビゲーションページが上位に推薦されてしまった．

4.3 節の実験では，それぞれのルールのヒット率は低いながらも，全てのユーザに対して常に推薦する方法ではなく，特徴的な傾向を表す推薦ルールに合致するユーザのみに，そのルールを用いて推薦する絞込み推薦の効果を示すことができたと考える．本稿の実験では，各ルールを個別に評価したが，実際は複数のルールを同時に適用できるため，推薦の精度は向上するものと考える．そのためにも，実際に絞込み推薦を実現するためのアルゴリズムを検討する必要があると考える．

次に，絞込み推薦がカバーできるユーザの割合について考察する．表 2 において，順序あり，なしの場合の各ルールにおける平均対象セッション数はそれぞれ 221，261 であり，最大のルールでも 406，513 セッションに留まった．この結果について解析を行うために，同様の 3 日分の 14,371 セッションを対象に，個々の特集ページについて，それらを含むセッションの割合を求めた．出現頻度が最も高い 4 つの特集ページと，その割合を表 3 に示す．1〜4 位のページが出現するセッションの数はそれぞれ 876，661，488，402 であり，全セッション数に対して非常に小さい値となっている．3 日分のアクセスログに出

表 3　出現頻度が上位の特集ページを含むセッションの割合

出現順位（内容）	割合
1 位（今月の食べ放題特集）	0.061
2 位（記念日 & 誕生日に行きたいお店特集）	0.046
3 位（少人数でもやっぱり個室特集）	0.034
4 位（お誕生日特典特集）	0.028

現する特集ページが 484 種類と多数であることも考慮し，これらの特集ページに対するアクセスは，大きな偏りがなく分布していると判断する．これによって，絞込み推薦でカバー可能なユーザの割合が小さくなったと考える．前提としている条件が異なるため，単純に比較をすることはできないが，直感的な推薦方法として，最頻出特集ページの推薦を考える場合，ヒット率は表 3 中の “割合” と同程度の値になると予測できるため，表 2 の多くの場合で，より高いヒット率が達成されていることが確認できる．現状では，絞込み推薦でカバーできるユーザの割合は小さいが，今後は，それぞれの特集ページに対する特徴的なルールを考慮し，それらを組み合わせることで，より多くのユーザをカバーできるようになると考える．

5. おわりに

5.1 まとめ

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，Web 利用マイニング手法として，アクセスシーケン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である，LCS（Longest Common Subsequences）に着目し，LCS を用いた Web ページ推薦手法である WRAPL を提案してきた．

本稿では，これまで比較的古く，現存しない Web サイトのログを用いて評価してきた WRAPL を，大規模な商用サイトの新しいアクセスログに適用することで，WRAPL による Web ページ推薦の効果を議論し，また，Web サイトの持つ特徴を調査することで，異なるアプローチである絞込み推薦を考案した．

実験の結果から，現在の Web サイトでは，動的ページや検索サイトの普及から，WRAPL や相関ルールを用いた既存手法，単に適用しただけでは，サイトトップページやコンテンツ毎のトップページなどのナビゲーションページを推薦しやすいことが分かった．そこで，抽出された LCS の中で特徴的ないくつかのルールに対して，対象とするユーザを限定して推薦を行う方法について検討し，その中で，LCS の持つ順序情報は推薦に良い影響を与えることを確認した．

5.2 今後の課題

本研究の今後の課題として，まず絞込み推薦の実現と評価に向け，利用すべきルールの選定条件や推薦を行うための合致度合いの条件，適用のためのアルゴリズムなどを調査する必要があると考える．また，その結果として推薦されたページが，ユーザにとって有用であったかどうかを評価するために，被験者実験を行うことも有意であると考える．本研究の目標は，ユーザにとって有用なページを推薦することであるため，被験者実験を通して，評価指標による客観的な評価のみでは判断が難しい，直感的・主観的な評価を行うことができると考える．

手法の改善に関して，レストランページの推薦や，より高速かつ精度の高い推薦を行うためには，十分な量のログからの LCS の抽出，目的に合った LCS の選択や限定を行うアルゴリズムが必要となる．そのためにも，より高速に効率よく LCS を抽出し，アクティブセッションとのマッチングを行う必要がある．また，本稿では，ページをナビゲーションページ，レストランページ，特集ページの 3 種類に分類し，それぞれを区別したが，より詳細な分類を加え，その分類を，LCS 抽出や推薦の際に利用することで，より高度な推薦が可能になると考える．例えば，レストランページの推薦を目的とし，料理カテゴリなどを用いて分類（抽象化）を行うなどの応用が考えられる．実際のサイトのログを分析し，これらの応用を検討することも今後の課題である．

さらに，今回比較対象とした，Web 利用マイニングの一種である相関ルールを用いた Web ページ推薦手法以外にも，協調フィルタリングを用いた [13] などの推薦手法とも比較を行っていきたい．

謝　辞

本研究で利用したアクセスログの御提供ならびに初期加工に御協力いただきました，株式会社ぐるなびの 福島常浩氏， 古木信司氏，岡本拓明氏に深く感謝致します．また， 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．

文　献

[1] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology(TOIT)*, 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] Andrej Kristofic and Mária Bieliková. Improving adaptation in Web-based educational hypermedia by means of knowledge discovery. In *Proceedings of the 16th ACM Conference on Hyertext and Hypermedia*, pp. 184–192, 2005.

[4] Xiaobin Fu, Jay Budzik, and Kristian J. Hammond. Mining navigation history for recommendation. In *Intelligent User Interfaces*, pp. 106–112, 2000.

[5] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from Web usage data. In *Proceedings of the 3rd International Workshop on Web information and data management*, pp. 9–15, 2001.

[6] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Using sequential and non-sequential patterns in predictive Web usage mining tasks. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'02)*, pp. 669–672, 2002.

[7] Olfa Nasraoui and Chris Petenes. An intelligent web recommendation engine based on fuzzy approximate reasoning. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 1116–1121, 2003.

[8] 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.

[9] 山元理絵, 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. In *Proceedings of the IPSJ DBWeb2006*, pp. 43–50, 2006.

[10] Sun Wu, Udi Manber, Gene Myers, and Webb Miller. An O(NP) sequence comparison algorithm. *Information Processing Letters*, Vol. 35, No. 6, pp. 317–323, 1990.

[11] 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.

[12] 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.

[13] Greg Linden, Brent Smith, and Jeremy York. Industry report: Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Distributed Systems Online*, Vol. 4, No. 1, 2003.