論文 / 著書情報
Article / Book Information

論題(和文)	Web ページ推薦における推薦順位決定のための得点付け手法の比較
Title(English)	Comparison of the Weighting Methods for Web Page Recommendation
著者(和文)	山元理絵, 小林大, 吉原朋宏, 小林隆志, 横田治夫
Authors(English)	Rie YAMAMOTO, Dai KOBAYASHI, Tomohiro YOSHIHARA, Takashi KOBAYASHI, Haruo YOKOTA
出典(和文)	DBSJ Letters, Vol. 5, No. 4, pp. 5-8
Citation(English)	DBSJ Letters, Vol. 5, No. 4, pp. 5-8
発行日 / Pub. date	2007, 3
権利情報 / Copyright	本著作物の著作権は日本データベース学会に帰属します。 Copyright (c) 2007 Database Society of Japan, DBSJ.

Webページ推薦における推薦順位決定のための得点付け手法の比較

Comparison of the Weighting Methods for Web Page Recommendation

山元 理絵♡ 小林 大♡ 吉原 朋宏♡
小林 隆志♢ 横田 治夫♢

Rie YAMAMOTO Dai KOBAYASHI
Tomohiro YOSHIHARA
Takashi KOBAYASHI Haruo YOKOTA

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いという利点があるため，我々は，これまで，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 を提案してきた．本稿では，WRAPL における推薦ページの優先順位付けのための得点付け方法の改良について検討する．実際の Web アクセスログを用いた実験を通して 3 種の手法について比較を行い，考察する．

The sophisticated website satisfying various requirements becomes much more important to propagate information via websites. Web page recommendation methods using web access logs are useful because of its unnecessity of the modification in the client. We have proposed WRAPL as a technique of extracting LCS (Longest Common Subsequences) from web access logs and using them to recommend web pages. In this paper, we compare three weighting methods for Web page recommendation using actual data to improve recommendation accuracy.

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，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 ページ推薦が盛んに研究されている．Web パーソナライゼーションのための情報収集の方法としては，閲覧した情報に対する各々のユーザの興味の有無を何らかの方法で収集し分析する方法や，Web サイトのアクセスログを分析する方法などがある．

前者の例としては，ユーザによる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する TextExtractor[2] がある．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，文や行の単位で興味情報として抽出できるが，クライアント側にプログラムを埋め込まなければならないという問題点がある．

一方，アクセスログを利用する方法は，クライアント側の変更が不要であり，様々な利用方法が研究されている．当初は，利用頻度に基づくリンクの接続性の評価 [3] や，バックトラックポイントの発見 [4] 等のアクセス解析の研究が中心であったが，最近では，ユーザビリティの向上を目的とし，ユーザ行動の予測手法 [5] や Web ページ推薦に関する手法 [6, 7] が提案されている．

ユーザ行動の予測手法 [5] では，ログ中のユーザのアクセスパスの統計を取ってユーザ行動をモデル化し，全てのページに対し，現在のアクセスパスに引き続いてアクセスされるための条件付確率を算出することで，続くアクセスページを予測する．予測モデルの適用例の 1 つとして Web ページ推薦が挙げられているが，このような手法を用いて推薦を行うと，推薦されるのは直後のページのみに限られてしまう．さらに，サイト内のナビゲーションが不適切な時やサイト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，ユーザは多種多様なパスを取り得るため，ページの的確な推薦が難しい．

一方，相関ルールを用いて Web ページ推薦を行う手法 [6, 7] では，アクセスログに apriori アルゴリズムを適用し，1 セッション中で頻繁に共起するページの組の集合を作成してページを予測する．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，前述した，直後のページの推薦に限られるという問題や，後戻りや遠回りを含む場合の問題を解消することができる．しかし，この手法では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には，すでにアクセスしたページを推薦したり，ユーザにとって不要となったページを推薦したりすることで，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）を用いることにより，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，Mobasher らの相関ルールを用いる手法 [6, 7] と比較して推薦精度が向上した実験結果を得ている．

本稿では，我々の提案する Web ページ推薦手法 WRAPL における推薦ページの優先順位決定のための得点付け手法に対して，[8] の FL 法による順位決定に加え，考慮すべき他の要因について検討し，FL 法を拡張した 2 種の推薦順位決定手法を議論する．さらに，それらの手法を用いて実際のデータに WRAPL を適用し，推薦精度の比較と考察を行う．

2. 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，まず，2.1 節で我々がこれまでに提案してきた，Web アクセスログからアクセスシーケンスの LCS を抽出する方法について述べる．次に，2.2 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-FL 法 [8] について説明する．続いて，2.3 節で手法の改良について検討する．

2.1 Web アクセスログからの LCS 抽出

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．

アクセスログから取り出したユーザのアクセスシーケンス群を基に LCS を抽出することで，寄り道等の余分な情報を取り除いた，共通の傾向を発見することができ，その利用によってサイト構成の改善が可能となる [10].

Web アクセスログから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，訪問者が行ったアクセスの URL シーケンスである，ユーザーセッションの情報を抽出する．セッション ID には，Cookie を用いることが一

♡ 学生会員 東京工業大学 大学院 情報理工学研究科 計算工学専攻
{yamamoto, daik, yoshihara}@de.cs.titech.ac.jp

♢ 正会員 東京工業大学 学術国際情報センター
tkobaya@gsic.titech.ac.jp, yokota@cs.titech.ac.jp

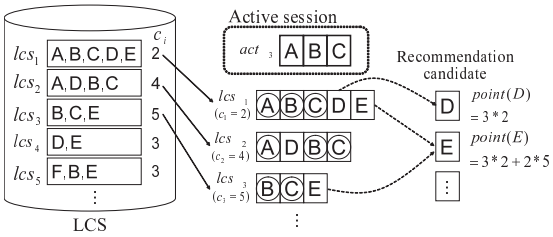

図 1: LCS を利用した Web ページ推薦
Fig.1 Web page recommendation using LCS

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跨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から LCS を求める手法として，我々は，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 [11] 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，多くの場合で動的計画法よりも大幅に小さい計算量で 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セッションの全ての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．

2.2 LCS を利用した Web ページ推薦

我々の提案する 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．

アクティブセッションにマッチした LCS の情報を利用して，推薦候補ページの優先順位を決定し推薦する方法として，我々は，WRAPL-FL（Web page Recommendation by Access Pattern Lcs with Frequency and matched Length based weighting）法 [8] を提案してきた．WRAPL-FL 法では，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．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．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を考慮して適切に調節する．
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．

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．

また，ここではページ E が 2 つの LCS で推薦候補となっているため，各 LCS から算出された得点の和がページ E の得点となる．

2.3 得点付け手法における他の要因の考慮

以下では，2.2 節で説明した WRAPL-FL 法に対し，各候補ページの推薦のための優先順位付けの方法を拡張した，FLD 法と FLP 法について説明する．

2.3.1 FLD 法

本研究では，ページアクセスが進むに従いユーザはトップページ等のインデックスページからコンテンツページ等のリーフページに近づいていき，ユーザが目標とするページは，リーフページである場合が多いと考える．そのため，リーフページを優先的に推薦するための順位付け手法を考える．リーフページへのアクセスに比べ，それらの親ページに当たるインデックスページの方がアクセス頻度が高いため，LCS にはリーフページよりもインデックスページが多く含まれることになる．そこで，得点を付加する際，トップページから遠いページにより高い得点を与えることで，リーフページを優先的に推薦する手法を考える．

あるページ p の深さ d_p は，サイト内の構造を考慮することで，トップページからの階層の深さとして一意に定義するか，もしくはその URL の様式から判断する．

以上を考慮した，各候補ページの得点計算のための方法として，FLD（Frequency, matched Length and Depth based weighting）法を次式で定義する．

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

ただし，β は d_p が $point(p)$ に与える影響度合を表すパラメタである．

2.3.2 FLP 法

我々は，アクセス順序は有意な情報であると考えるため，アクティブセッション中のそれぞれのページは，その位置によって異なる情報を持つと予測する．そこで，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．

重み付けに際し，予備実験を行ったところ，アクティブセッション中の前方のページが LCS と一致する場合に比べ，後方ページが一致する場合に良い結果が得られた．この結果から，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と考え，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．例えば，$act_4 = (A, B, C, D)$，$lcs_i = (B, D, E)$ 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，FL 法による推薦ページの優先順位付けの式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 FLP（Frequency, matched Length and Position based weighting）法と呼ぶ．

$$point(p) = \sum_{lcs_i \in LL} |lcs_i \cap_p act_n| \cdot c_i^\alpha \cdot l_i^\gamma \quad (2)$$

ただし，γ は，l_i の $point(p)$ に対する影響度合を表すパラメタである．

3. 評価実験

本節では，実際のアクセスログに対し，WRAPL-FLD 法と WRAPL-FLP 法を適用し，WRAPL-FL 法による推薦精度との比較を行う．

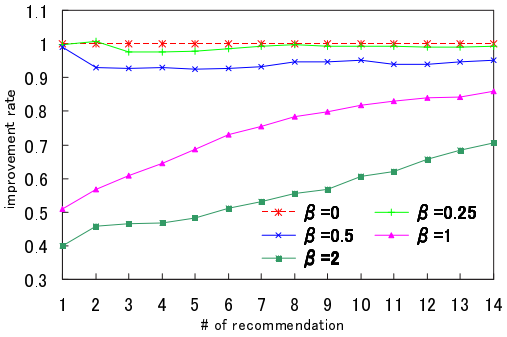

図 2: FLD 法による改善率（precision）
Fig.2 Improvement rate with FLD (precision)

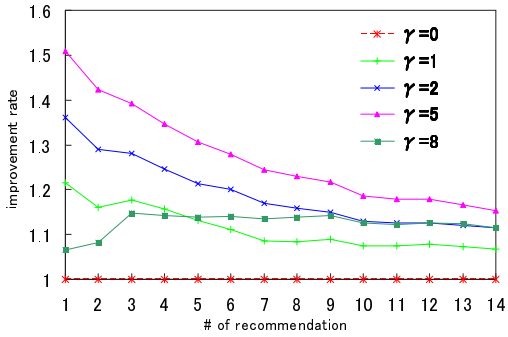

図 4: FLP 法による改善率（precision）
Fig.4 Improvement rate with FLP (precision)

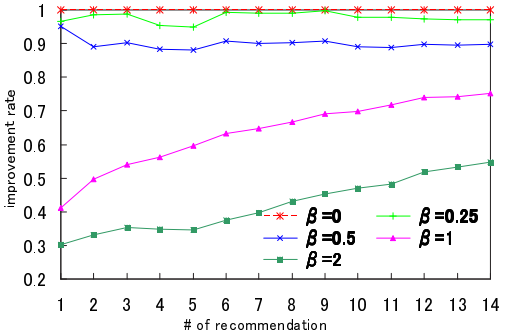

図 3: FLD 法による改善率（coverage）
Fig.3 Improvement rate with FLD (coverage)

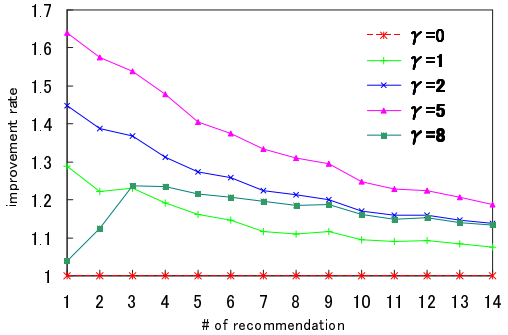

図 5: FLP 法による改善率（coverage）
Fig.5 Improvement rate with FLP (coverage)

3.1 実験に用いたデータ

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 サイトのアクセスログの内，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．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 秒以上の時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．

3.2 実験

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 を抽出し（$min.Count = 150$，$min.Length = 3$ として LL を作成），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アクティブセッション長 2~4 の各場合でページ推薦を行い，その評価を行った．

また，今回対象とした Web サイトは現存しないため，FLD 法の適用に際しては，URL 中の ‘/’ の個数を d_p として実験を行った．

評価のための指標として，precision と coverage を用いた．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，アクティブセッションに引き続いてアクセスされたページの組である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．

まず，FLD 法による効果を調べるために実験を行った．図 2，3 はそれぞれ，長さ 4 のアクティブセッションからのページ推薦において，WRAPL-FLD 法における深さ d_p の影響度合 β の値を変化させた場合の WRAPL-FL 法（$\beta = 0$）に対する改善率を表している．縦軸は，$\beta = 0$ の結果に対する割合を表しており，横軸は推薦する上位ページの数に対応している．グラフより，β の値を大きくするほど結果が悪化していることが確認できる．アクティブセッション長を 2，3 とした場合にも同様の結果が得られた．

次に，FLP 法の効果を調査するために，同様の指標を用いて実験を行った．図 4，5 はそれぞれ，長さ 4 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 法におけるマッチ位置重み l_i の影響度合 γ の値を変化させた場合の WRAPL-FL 法（$\gamma = 0$）に対する改善率を表している．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．アクティブセッション長を 2 とした場合には，γ の値を大きくするほど precision，coverage 共に結果が改善し，$\gamma = 8$ の場合で最も良い結果が得られた．また，3 の場合には図 4，5 と同様の傾向が得られた．

さらに，FL 法，FLD 法，FLP 法を用いて順位決定を行った場合の比較を図 6 に示す．各手法におけるパラメタは，$\alpha = 0.5$，$\beta = 1$，$\gamma = 5$ を使用した．FLP 法で最も高い推薦精度が得られ，続いて FL 法，FLD 法の順となった．

4. 考察

各手法による精度の比較としては，まず，FLD 法を適用した場合，図 2，3 の結果から，FL 法に比べて推薦精度は悪化している．しかし，2.3.1 節で述べたように，FLD 法の適用範囲として，コンテンツ（リーフ）ページの優先的な推薦を想定しており，今回の評価手順では，アクティブセッション以降にアクセスされたページ全てを “正解” としたため，出現頻度の高いインデックスページが評価セットに多く含まれ，それらが推薦されにくくなることで精度が低下したと考える．したがって，目

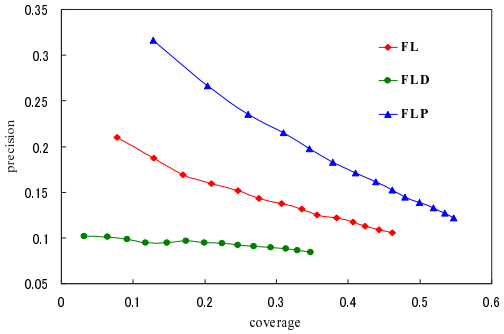

図 6: 各順位決定手法の比較
Fig.6 Comparison of each weighting method

的に合致した異なる評価指標を定義し，再評価を行うべきであると考える．

FLP 法を用いた場合には，FL 法を用いて推薦ページの優先順位付けを行った場合に比べ，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，実際のリンク構造等を解析することはできないが，URL から判断した時，ページ配置が細分化・階層化されており，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．

また，図 4，5 より，FLP 法の適用の結果，$\gamma = 8$ の時精度が低下していることが分かる．これは，後方ページの影響度合が強くなりすぎることで順序情報などが考慮されなくなり，精度が低下したと考察する．しかし，全体を見ると，得点付けの式において l_i の影響度合（$= \gamma$）を大きくしても精度の向上が確認できるため，今回実験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わかる．

5. おわりに

本稿では，我々がこれまでに提案してきた，Web ページ推薦手法 WRAPL における推薦ページの優先順位決定のための得点付け方法に対して，従来の FL 法による順位決定に加え，考慮すべき他の要因について検討し，FL 法を拡張した 2 種の推薦順位決定手法 FLD 法，FLP 法について議論した．また，それらを実際のデータに適用することで，その効果を比較した．

実験の結果，アクティブセッション中の LCS とのマッチ位置を考慮した FLP 法で FL 法に比べて精度が改善し，一方，サイト内の位置を考慮した FLD 法では精度が悪化したため，そのような結果が得られる原因について考察を行った．

今後の課題としては，インデックスページとコンテンツページが明確に分類されたサイトにおけるアクセスログへの WRAPL-FLD 法の適用が挙げられる．コンテンツ（リーフ）ページを優先的に推薦することを目的とし，評価においては，4. 節で言及したように，目的に合ったページのみを正解として評価することで，WRAPL-FLD 法の有用性を詳細に解析する必要があると考える．

さらに，LCS を利用したアクセスログ解析の持つ特徴を明確にするために，他のモデルとの比較を行う必要がある．特に，ベイジアンネットワークを用いた推薦手法は WRAPL と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」及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．

[文献]

[1] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM TOIT*, Vol. 3, No. 1, pp. 1–27, 2003.
[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, Vol. 19, No. 3, pp. 365–372, 2004.
[3] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Creating adaptive web site through usage-based clustering of URLs. In *Proc. of KDEX*, pp. 19–25, 1999.
[4] Ramakrishnan Srikant and Yinghui Yang. Mining web logs to improve website organization. In *Proc. of WWW*, pp. 430–437, 2001.
[5] James E. Pitkow and Peter Pirolli. Mining longest repeating subsequences to predict WWW. In *Proc. of USITS*, 1999.
[6] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective perssonalization based on association rule discovery from Web usage data. In *Proc. of WIDM*, pp. 9–15, 2001.
[7] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Using sequential and non-sequential patterns in predictive Web usage mining tasks. In *Proc. of ICDM*, pp. 669–672, 2002.
[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 *Proc. of IPSJ DBWeb2006*, pp. 43–50, 2006.
[10] 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.
[11] Sun Wu, Udi Manber, Gene Myers, and Webb Miller. An O(NP) sequence comparison algorithm. *Inf. Process. Lett.*, Vol. 35, No. 6, pp. 317–323, 1990.

山元 理絵　Rie YAMAMOTO
平 17 東工大・工・情報工卒．同大大学院・情報理工・計算工・修士課程在学中．日本データベース学会学生会員．

小林 大　Dai KOBAYASHI
平 17 東工大大学院・情報理工・計算工・修士課程了．同大大学院・情報理工・計算工・博士課程在学中．日本データベース学会学生会員．日本学術振興会特別研究員 DC．

吉原 朋宏　Tomohiro YOSHIHARA
平 17 東工大・工・情報工卒．同大大学院・情報理工・計算工・修士課程在学中．日本データベース学会学生会員．

小林 隆志　Takashi KOBAYASHI
平 9 東工大・工・情報工学卒．平 11 同大大学院・情報理工・計算工学・修士課程了．平 16 同大大学院・同専攻・博士課程了．平 14 同大学術国際情報センター・助手．工博．日本データベース学会，日本ソフトウェア科学会，情報処理学会，ACM 各会員．

横田 治夫　Haruo YOKOTA
昭 55 東工大・工・電物卒．昭 57 同大大学院・情報・修士課程了．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．昭 61 (株) 富士通研究所．平 4 北陸先端大・情報・助教授．平 10 東工大・情報理工・助教授．平 13 東工大・学術国際情報センター・教授．工博．日本データベース学会理事，電子情報通信学会，情報処理学会，人工知能学会，IEEE，ACM 各会員．