論文 / 著書情報
Article / Book Information

論題(和文)	複数属性に着目したアクセス履歴からのページ推薦手法
Title(English)	Access Log based Web Page Recommendation Using Multiple Attributes of Web Pages
著者(和文)	岡本 拓明, 横田 治夫
Authors(English)	Hiroaki Okamoto, Haruo Yokota
出典(和文)	Web DB フォーラム 2009 論文集, , ,
Citation(English)	Proc. of WebDB Forum 2009, , ,
発行日 / Pub. date	2009, 11
権利情報 / Copyright	

複数属性に着目したアクセス履歴からのページ推薦手法

岡 本 拓 明[†1]　横 田 治 夫[†2]

近年，Web サイトにおける情報量の増大に伴い，サイト中の適切な Web ページへとユーザを導く Web ページ推薦の重要性が高まっている．特に，Web サイトのアクセス履歴からパターンを抽出して推薦する手法は，ユーザの評価情報を直接収集する必要がないため障壁が少ない．しかしながら，これまでの手法ではアクセスしたページ単位でのパターン抽出を対象としているため，Web サイト中のページ数が増加するとパターンの種類も膨大となり，ユーザに提供すべきページに至るパターンの抽出が困難になる問題があった．我々は，各 Web ページそのものではなく，各 Web ページの持つ複数の属性に着目し，それらの組合わせのパターンを抽出することで Web ページの推薦を行う手法を提案する．複数の属性を対象とすることで，多くのユーザに共通する傾向を適切に把握し，それを推薦に利用することができ，新規の Web ページでも対応できる．適切な推薦を行うためには，どのような属性をどのような粒度で使い，どのように組み合わせるかが重要となる．本稿では，実際の商用サイトのアクセス履歴を用いて，着目すべき属性とその粒度およびその組み合わせ方法に関して検討を行う．

Access Log based Web Page Recommendation Using Multiple Attributes of Web Pages

Hiroaki OKAMOTO[†1] and Haruo YOKOTA[†2]

Because the volume of information in a web site increases quite large recently, web-page recommendation systems leading users to appropriate web pages become very important. Recommendation approaches extracting the pattern from access histories of a web site are efficient, since they do not require evaluation information from users directly. However, when the number of pages in a web site becomes huge, it is hard to find high frequent patterns including the accessed pages in the history because the kinds of patterns in it also increase rapidly. To attack the problem, we propose a method of recommending web pages by using multiple attributes of a web page. It does not find access patterns of the web page itself, but extracts pattern combination of multiple attributes the accessed pages have. The combination of attributes provides the abstracted patterns enabling to derive the access tendency of many users, and to be used for appropriate web-page recommendation. To make the proposed approach effective, it becomes important to find what kind of and what granule of attributes should be used. In this paper, we use an actual access history of a web site in a private company to examine appropriate combination of attributes and their granularity.

1. はじめに

近年，Web サイトにおける情報量の増大から，サイトのユーザの選択肢が増え，提供したい情報にユーザがたどりつけない場合も多くなっている．このため，Web サイトに訪れたユーザを適切なページに導くための Web ページ推薦の重要性が高まってきている．

例えば，飲食店情報検索サイト「ぐるなび」[⋆1] においてもサイトに掲載している飲食店数が増大しており，2009 年 4 月現在，詳細情報掲載店舗数は約 6 万 3 千店となっているが，その中でサイトを訪れたユーザが自分の要求にあった飲食店の情報までたどりつくことは必ずしも容易ではない．このため，訪れたユーザに対してそのユーザが望んでいると推測される飲食店の Web ページを候補として推薦することが重要と言える．

Web ページ推薦手法には，大別すると，ユーザの評価情報を直接収集し分析する手法と，Web サイトに残るアクセス履歴を解析する手法がある．前者は，アクセスした Web ページをどのように評価しているかという情報を収集するためにユーザに評価を求め，各ユーザに労力を要求することから導入の障壁が高いとともに，適切な評価が得られるかどうかがそれぞれのユーザに依存し，ばらつくことが考えられる．一方後者は，一般のアクセス履歴中のパターンを解析するために，ユーザには特別な労力を求めず，全てのユーザから同レベルの情報を得ることが可能となるため，前者に比べると導入が容易で評価のばらつきも少ないと言える．

後者のアクセス履歴を解析する手法としては，アクセス履歴の中からユーザのセッションを抽出し，そのセッション中にアクセスされた各 Web ページをアイテムとして，マイニングアルゴリズムを適用し相関ルールを抽出して推薦する手法[2),3)]，Web ページの頻出

†1 株式会社ぐるなび
GOURMET NAVIGATOR INCORPORATED
†2 東京工業大学 学術交際情報センター
Tokyo Institute of Technology Grobal Scientific Information and Computing Center
⋆1 http://www.gnavi.co.jp/

アクセスパターンを抽出して推薦する手法，アクセスパターンの LCS (Longest Common Subsequence) を抽出して推薦する手法[5] 等が提案されている．

相関ルールを抽出する手法は，アクセスの順番を考慮しないため，ユーザのアクセス動向を的確に抽出できないという問題がある．また，Web ページの頻出アクセスパターンを抽出する手法は，全く同一のアクセスパターンでないと推薦できないため，特に Web ページ数が増えて発生パターンの種類が膨大になると，十分な頻度を持ったパターンを抽出することができない．アクセスパターンの LCS を抽出する手法は，アクセスの順番が完全に一致しない場合でも，アクセスした順番の特徴を抽出することができるため，他の手法に比較すると有効という結果が報告されている[7]．

しかし，Web ページの数が膨大になると LCS を抽出する手法であっても，パターンの種類が多くなり，アクセスパターンをそのまま利用する方法と同様に，十分な頻度を持ったパターンを抽出することが困難となる．後述するように，上記の「ぐるなび」の飲食店のアクセス履歴に対して LCS を求めると，実際には十分な長さを持った LCS はほんの僅かしか得られなかった．

そこで，本稿では，Web ページそのもののアクセスパターンの LCS ではなく，各 Web ページが持つ複数の属性に着目し，それらの組合わせのパターンの LCS を抽出することで Web ページの推薦を行う手法を提案する．属性に抽象化することで，複数の Web ページをまとめて解析することができ，多くのユーザに共通する傾向を適切に把握し，それを推薦に利用することができると期待できる．また，属性が分かっていれば新規の Web ページを含むアクティブセッションを対象にすることも，新規の Web ページを推薦することも可能となる．

このように Web ページの複数の属性を用いて LCS を抽出する場合，属性の組み合わせによって，抽出される LCS の数や長さも変わってくる．また，一つの属性に着目しても，どのような粒度で分類するかによって変わってくる．このことから，適切な Web ページ推薦を行うためには，Web ページのどのような属性をどのような粒度で用い，どのように組み合わせるかが重要となる．本稿では，実際に「ぐるなび」の飲食店サイトのアクセス履歴に対して提案手法を適用し，従来の Web ページ単位の LCS の抽出との比較して評価するとともに，属性の選択と粒度の影響を調べる．

以下，2 では，関連研究について紹介し，3 では従来の Web ページ単位の LCS の抽出とその問題点について述べる．次に，4 において，提案手法である複数属性に着目した LCS 抽出について，その処理の流れと，対象とすべき属性の種類，および属性の粒度について検討する．5 では,「ぐるなび」のアクセス履歴を用いた実験の内容と，属性の選択と粒度を変化させた結果について報告する．

2. 関 連 研 究

Web ページのアクセス履歴に相関ルールマイニング手法を適用する Web ページ推薦手法[2] では，アクティブセッション中の Web ページに対して，今までにアクセスした Web ページと共起頻度の高い Web ページを推薦する．しかし，新規のページにアクセスした場合は推薦できないという問題があり，改良した手法 [3] も提案されているが，ユーザがブックマーク情報を提供する必要があるなどコストが大きい．さらに，いずれの手法も Web ページアクセスの順番を考慮していない．本稿で対象としているような順番に Web ページを絞り込んでいくような Web サイトにおいては，順番を考慮する必要がある．

また，書籍を販売するサイトにおける商品推薦を目的として，協調フィルタリングを用いた手法や[1]，ユーザの評価履歴を元にユーザの嗜好性モデルを作成し，それを基に推薦を行う手法[6] も提案されている．しかしながら，どちらの手法についてもユーザの評価を何らかの形で登録しておく必要があるため，やはりコストが大きい．

一方，コンテンツの内容や性質を定量化し，ユーザごとに評価値を予測するモデルも提案されている[4] が，実際の事例に適用するために，どのように定量化するかが課題となっている．

3. Web ページ単位の LCS の抽出とその問題点

店舗情報を提示する「ぐるなび」のような Web サイトにおいては，複数のユーザのアクセス動向を把握して推薦することが重要であり，そのためには各ユーザのアクセスの順番を考慮することが有用となる．このため，順番を考慮しない相関ルールを抽出する手法は適さない．しかし，アクセスパターンの頻度をそのまま利用すると，全く同一のアクセスの順番にしか推薦できないため，推薦できるものが限られる．このため、アクセスのシーケンスの中から LCS を抽出して利用する方法が，相関ルールを使う方法や頻出アクセスパターンをそのまま使う方法より有効である．

シーケンス x の部分シーケンスとシーケンス y の部分シーケンスの中で両方のシーケンスに含まれるものを共通部分シーケンス (Common Subsequence) という．共通部分シーケンスの中で最も長いものを最長共通部分シーケンス (Longest Common Subsequence) と呼び，LCS と略する．例えば，x=A-F-B-D と y=A-B-C-D の LCS は A-B-D である．

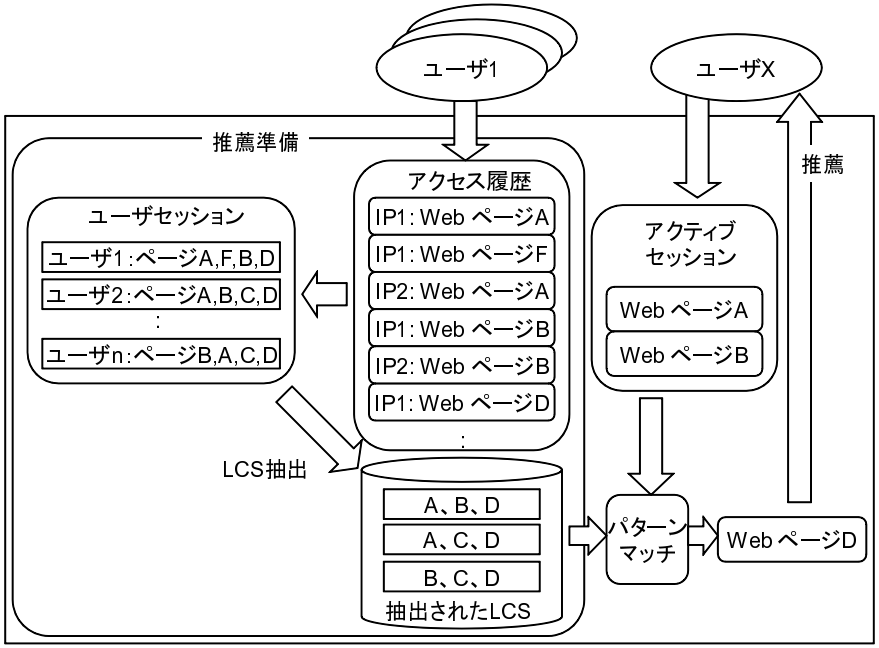

図 1 Web ページアクセスパターンの LCS を用いた Web ページ推薦

アクセス履歴中から抽出したユーザセッションの Web ページのリストから抽出された LCS を記憶しておくことで，途中の横道にそれたアクセス等を除いた多くのユーザが通るパターンを抽出することができる．この記憶しておいた LCS の中から推薦対象のアクティブセッションの Web ページのアクセス順に前半が対応する LCS を探し出して，その後半の Web ページを示すことで，前半に似たアクセスパターンを持つ多くのユーザがその後にアクセスした Web ページを推薦することができる[7]．Web ページアクセスパターンから LCS を抽出して Web ページ推薦を行う処理の流れの概要を，図 1 と対応させて以下に述べる．

Step1: まず，ユーザのアクセス履歴中に含まれる IP アドレス情報と Cookie 情報を基に同一ユーザであると判定された Web ページアクセス履歴を結合することによって，ユーザセッションを抽出する．ここでは，ユーザセッションは Web ページの ID のシーケンスとなる．図 1 の例では，同じ IP アドレス (IP1) を持つ Web ページのアクセス履歴を結合し，ユーザセッション A-F-B-D を抽出している．

表 1 「ぐるなび」の Web ページアクセスパターンから抽出された LCS

LCS 長	種類
3	8
4	3
5	1

Step2: 次に，抽出されたユーザ 1 からユーザ n までの全てのユーザセッションに対して，その任意の 2 セッションに含まれる LCS を算出し，その頻度情報とともに蓄積する．図 1 の例では，ユーザセッションの組 A-F-B-D と A-B-C-D からはその LCS である A-B-D を，別のセッションの組である A-B-C-D と B-A-C-D からその LCS である A-C-D と，B-C-D を得ている．

Step3: 推薦の対象となる現在のアクティブセッションと上で求め蓄積しておいた LCS を比較し，推薦候補を得る．図 1 の例では，ユーザ X のアクティブセッション A-B に対して，蓄積されている LCS の中の A-B-D の前半とパターンマッチし，推薦候補の Web ページ D を得ている．

この手法を用いた文献[7] では，出現頻度の高いアクセスパターンを重視することで，精度の高い推薦を実現している．しかし，推薦対象となる Web ページ数が増大すると，Web ページアクセスパターンの LCS を用いても，適切な Web ページを推薦できなくなる．

実際に「ぐるなび」の飲食店サイトの 2008 年 11 月 1 日のアクセス履歴から 1,000 セッションをサンプリングし，LCS を抽出した結果，表 1 に示す長さと種類の LCS を抽出できたが，表から分かるように十分な長さを持った LCS は少なかった．さらに，この抽出した LCS を用いて，同一の「ぐるなび」の飲食店サイトに対して 2009 年 8 月 21 日からサンプリングした 1,000 アクティブセッションに対して推薦候補の抽出を試みたところ，実際に推薦候補を見つけることができたセッションは 26 セッションにとどまった．これは，適用した「ぐるなび」の飲食店サイトにおいては，推薦対象となる Web ページの飲食店の数が 60,000 店以上存在し，抽出した LCS の数に対して，ユーザのアクセスしたパターンの数のほうがはるかに大きいためと言える．

4. 複数属性に着目した LCS 抽出

Web ページ数が増え，アクセスパターンが膨大になった場合にも十分な LCS を生成するため，以下では，Web ページそのものではなく，Web ページの持つ複数の属性に着目した手法を提案する．

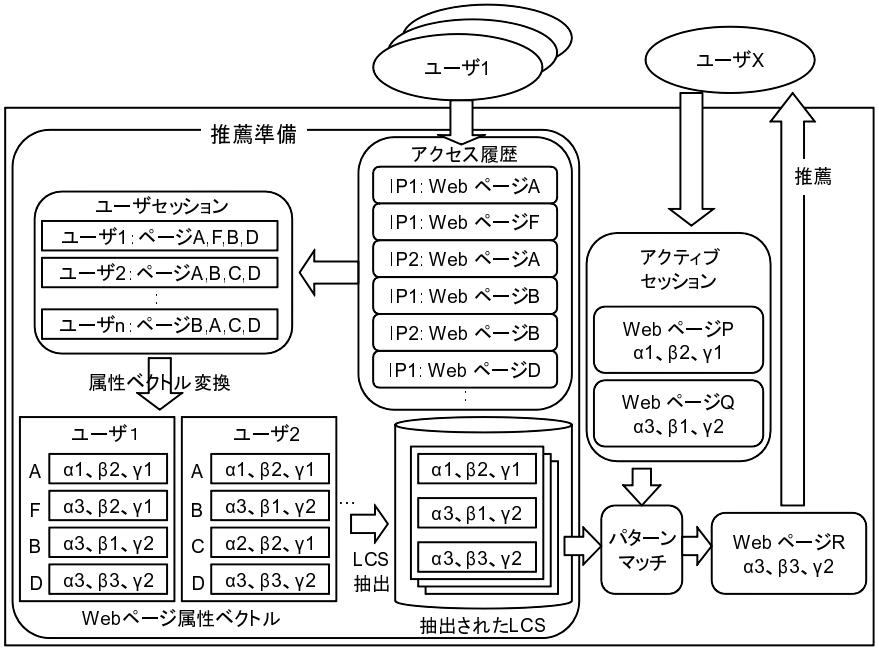

図 2 複数属性に着目した Web ページ推薦

4.1 処理の流れ

提案する Web ページの複数属性に着目した手法の処理の流れの概要を，図 2 と対応させて以下に述べる．

Step1: まず，Web ページ単位の LCS 抽出手法と同様にユーザのアクセス履歴からユーザセッションを抽出する．ここまでは，前述した Web ページ単位の推薦の時と同様である．

Step2: 次に，ユーザセッション中に含まれる Web ページのシーケンスを各 Web ページの持つ属性ベクトルのシーケンスに変換する．図 2 の例では，ページ A が，α 1，β 2，γ 1 という属性を，ページ B がα 3，β 1，γ 2 という属性を，ページ D がα 3，β 3，γ 2 という属性を，ページ F がα 3，β 2，γ 1 という属性を持っていたし，A-F-B-D というユーザセッションは、(α 1，β 2，γ 1)-(α 3，β 2，γ 1)-(α 3，β 1，γ 2)-(α 3，β 3，γ 2) という属性ベクトルのシーケンスに変換される．

Step3: 上で求めた属性ベクトルのシーケンスに対して，属性ベクトルどうしの全要素が等しい場合に同一と判断する LCS を抽出して蓄積しておく．A-F-B-D と A-B-C-D というユーザセッションに対応する属性ベクトルのシーケンスの組からは，(α 1，β 2，γ 1)-(α 3，β 1，γ 2)-(α 3，β 3，γ 2) という属性ベクトルの LCS が抽出される．

Step4: 推薦対象のアクティブセッションに対しても，アクセスした Web ページを属性ベクトルに変換する．図 2 の例では，ページ P が (α 1，β 2，γ 1)，ページ Q が (α 3，β 1，γ 2) という属性ベクトルを持っていたとする．

Step5: アクティブセッションの属性ベクトルのシーケンスと属性ベクトルの LCS の前半を比較し，パターンマッチする属性ベクトルの LCS の後半の属性ベクトルを推薦候補とする．図 2 の例では，(α 3，β 3，γ 2) が推薦候補の属性ベクトルとなる．

Step6: 推薦候補と同じ属性ベクトルを持つ Web ページを推薦する．図 2 の例では，ページ R が (α 3，β 3，γ 2) という属性ベクトルを持っていたため，推薦される．

提案手法では，属性ベクトルを用いることにより，同一の Web ページでなくとも，複数属性が一致していれば同一と判定することで，長い LCS ができる可能性が高くなる．さらに，その LCS を用いた推薦においても，属性ベクトルを用いることでアクセス履歴には含まれなかった Web ページでもあっても推薦が可能となる．図 2 ではアクティブセッションの Web ページ P，Q や，推薦対象の Web ページ R はアクセス履歴に含まれていない場合にも推薦可能である．これに対して，図 1 で示した従来の Web ページの LCS の推薦では，アクティブセッションに現れる Web ページも，推薦対象の Web ページもアクセス履歴に含まれていなければならなかった．

一般に，各 Web ページは複数の属性を持つことが想定できる．1 属性だけを用いると，その属性に偏って推薦を行ってしまうため，ユーザに対して適切なページに誘導できない可能性がある．そこで，提案手法では複数の属性のベクトルに変換する．この複数の属性として，どのような種類の属性で，どのような粒度を用いるかについてが重要となる．以下，まず属性の種類に関して検討し，次にその粒度について検討する．

4.2 属性の種類

Web ページ推薦が有用であると想定される Web サイトにおいて，推薦対象となる各 Web ページが持つと思われる属性の候補を考えてみると以下のようなものを挙げることができる．

- 飲食店の Web ページ：業態，平均予算，エリア，口コミ数，個室あり，…
- 宿泊先の旅館やホテルの Web ページ：ホテル・旅館，宿泊代，エリア，温泉有無，…
- マンションや賃貸物件の Web ページ：賃貸料，広さ，エリア，新築・中古，…

- ニュースや記事などの Web ページ：記事種類，記事タイトル，日時，…
- 本や CD などの商品の Web ページ：ジャンル，著作者，価格，発行年，サイズ，…
- 音楽ダウンロードサイトなどにおける楽曲の Web ページ：ジャンル，歌手，価格，発行年，…

このような属性例の観測から，属性を以下のタイプ別に分類してみる．

カテゴリ　飲食店推薦における業態や本・CD 推薦におけるジャンルが代表的で，主に質的データになる．このカテゴリの分け方は，その推薦の行われている Web サイトの検索の仕様であることが多い.「ぐるなび」では，洋食や和食，居酒屋などの飲食店の業態にあたる．

範囲　飲食店推薦における平均予算やホテル推薦やマンション推薦における宿泊代や賃貸料が代表的な量的データである．ユーザの希望では下限や上限，あるいは両方が決まっているなど，ある一定の範囲を取ることが多い.「ぐるなび」では，飲食店の平均予算にあたる．

距離　飲食店推薦におけるエリアや，ホテル推薦におけるエリア，マンション推薦におけるエリアになる．単なる位置的情報以外にも「駅から○○分」といった形の形式を取ることもある.「ぐるなび」では，飲食店が立地するエリアにあたる．

評価　各推薦における人気ランキングやアクセス数ランキングなどである．過度に重視すると，特定の Web ページに推薦が偏る可能性が有り，取り扱いに注意すべきである.「ぐるなび」では，飲食店の口コミの数や，アクセス数にあたる．

付加情報　飲食店推薦における「個室有り」や，マンション推薦における「風呂トイレ別」など，持っていることに対して，特定のユーザのみがメリットを感じる情報である.「ぐるなび」では，駐車場の有無や，喫煙席の有無などにあたる．

上記で大別した属性のタイプの内,「評価」と「付加情報」は扱いに考慮が必要なことから，今回の評価では,「カテゴリ」,「範囲」,「距離」という 3 属性を対象とする．

4.3　属性の粒度

Web ページの各属性には分類の粒度がある．例えば「ぐるなび」の飲食店 Web ページにおける「カテゴリ」の属性は飲食店の業態にあたるが，和食、洋食、中華といった大きな分類から，和食の中でも，懐石，割烹，寿司，田舎料理といった細かな分類まである.「距離」であるエリアも，東京の中でも，新宿，渋谷といった広いレベルから，新宿西口・都庁前，新宿三丁目・御苑周辺，渋谷道玄坂・神泉といった少し狭いレベルまである.「範囲」である食事の平均予算も，四捨五入でまるめて，100 円単位から，500 円単位，1000 円単位といった粒度にすることができる．

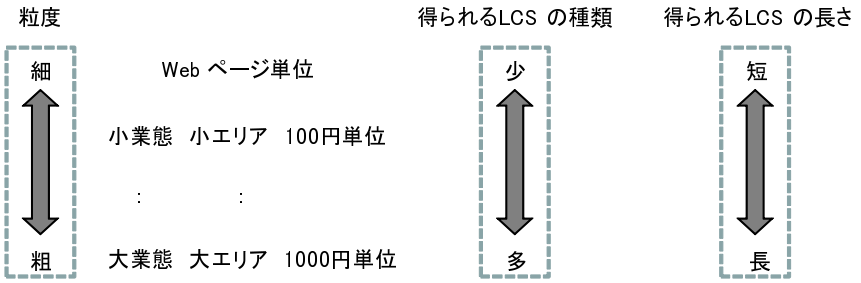

図 3　属性の粒度と抽出される LCS の関係

「ぐるなび」おける業態を「大業態」と「小業態」の 2 種類に，エリアは「大エリア」,「中エリア」,「小エリア」の 3 種類に分類する．2009 年 8 月時点での飲食店サイトの状態は,「大業態」として 12 分類,「小業態」として 127 分類になっている．同様に「大エリア」は 181 分類,「小エリア」は 704 分類となっている．

属性の粒度を変えた場合に抽出される LCS の関係を考えると，一般には，図 3 に示すように，粒度を粗くするほど得られる LCS の長さは長くなり，LCS の種類も多くなる．なお，この比較の上では，もっとも粒度が小さいのは属性に着目せず Web ページ単位の場合となる．

ここで，前述の実際の「ぐるなび」の飲食店サイトの 2008 年 11 月 1 日のアクセス履歴から 1,000 セッションをサンプリングしたデータに対して，業態，エリア，平均予算について，それぞれの粒度を変化させた場合に得られた LCS の数，種類，平均 LCS 長，最長 LCS 長を表 2 に示す．Web ページ単位で抽出した LCS に関しても比較のために示すが，どの属性を用いたとしても，得られる LCS の種類は増加し，LCS 長も長くなっていることが分かる．また，上で解析したように，粒度が粗いほど LCS の種類が増え，長さが長くなっているいることも分かる．LCS が長くなればなるほど，アクティブセッションと共通の属性を含む可能性が高くなるため，推薦できるアクティブセッションの割合も高くなり，推薦すべき属性を包含する可能性も高くなる．

また，得られる LCS が多くなればなるほど，推薦すべき属性を包含する可能性も高くなるが，属性が粗くなるため，推薦すべきではない Web ページを含む確率も上がる．これらは情報検索の分野における，適合率と再現率の関係と同じと考える．属性を粗くしすぎる

表 2 属性の組合せによる LCS の変化

属性の組合せ	LCS の数	LCS の種類	平均 LCS 長	最長 LCS 長
Web ページ単位	25	12	3.20	5
小業態・小エリア・100 円単位平均予算	45	26	3.36	5
大業態・小エリア・100 円単位平均予算	759	34	3.78	6
小業態・中エリア・100 円単位平均予算	96	35	3.79	6
大業態・中エリア・100 円単位平均予算	941	44	3.76	7
小業態・大エリア・100 円単位平均予算	7436	56	4.40	6
大業態・大エリア・100 円単位平均予算	10261	83	4.28	7
小業態・小エリア・500 円単位平均予算	98	30	3.36	5
大業態・小エリア・500 円単位平均予算	829	50	3.74	6
小業態・小エリア・1000 円単位平均予算	478	59	3.36	5

と，再現率は上昇するが，適合率は下がる可能性がある．つまり，両者はトレードオフの関係にあると言え，最も良い粒度の属性を調整する必要がある．

5. 評 価 実 験

次に，従来の Web ページ単位で LCS を抽出する手法と提案手法である複数属性に変換する手法による Web ページ推薦を,「ぐるなび」の飲食店サイトの実際のアクセス履歴に対して適用し，時期を変化させた別のアクセス履歴をテストセッションとして評価を行う．

5.1 実験対象データ

実験対象のデータとして，前述の 2008 年 11 月 1 日に「ぐるなび」の飲食店サイトへのリクエストに対するアクセス履歴を用いた．このアクセス履歴に含まれる Cookie 情報を用いてアクセス履歴を繋ぎ合せることでユーザセッションを作成した．少ない Web ページにしかアクセスしないユーザセッションでは推薦に利用できないと考え, 作成したユーザセッションの内，セッション中にアクセスしたアイテム数が 3 以上のセッション 40,312 セッションを対象にした．

推薦に対して良い属性の粒度を求めることを主眼とし，実行時間の関係から，上記のセッションの内ランダムに 1,000 セッションを抽出し，そのセッションの総当たりを行いアクセスした Web ページでの LCS の抽出と, セッションを業態，平均予算，エリアに変換して LCS の抽出を行った．

抽出した LCS を用いて，文献[7] で提案されている WRAPL-FL 法を用いて推薦を行った．この手法は，あるアクティブセッションに対して推薦を行う際に，まず，あらかじめ LCS を抽出しておく．次に，抽出した LCS とアクティブセッションに共通する Web ページを抜き出し, LCS からその共通部分の最後まで除去を行う．例えば，アクティブセッションが A-B，LCS が A-C-B-A-D だと仮定すると，LCS から A-C-B-A の部分を除去し，推薦候補となる D を得る．このようにして，除去して残った Web ページに対して出現頻度分の得点を加算し, 全ての LCS との得点加算が終了したときに，得点の一番高かった Web ページを推薦候補とする．

評価のため，2009 年 8 月 21 日にアクセスのあった 1,000 セッションをテストセットとし，そのユーザセッション中の前 2 アクセスをアクティブセッションとし，その後に実際にアクセスした Web ページを正解集合として扱った．

5.2 実験結果と考察

実験結果に対して以下に定義する適合率 (precision)，再現率 (recall)，および F 値 (F-measure) を用いて評価を行う．

$$precision = \frac{|recom \cap eval|}{|recom|} \tag{1}$$

$$recall = \frac{|recom \cap eval|}{|eval|} \tag{2}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

ここで，$recom$，$eval$ は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続いて実際にアクセスされた正解ページの組を表す．適合率 (precision) は，推薦されるページ数に対する正解ページ数の割合，再現率 (recall) は，評価セットのページ数に対する正解ページの割合でである．F 値 (F-measure) は適合率と再現率の調和平均である．

Web ページ単位で LCS を抽出する手法と，提案手法において，業態に関して大小の 2 種類の粒度, エリアに関して大中小の 3 種類の粒度，平均予算を 100 円単位，500 円単位，1000 円単位の粒度に変化させて LCS を抽出したものの適合率，再現率，F 値の結果を表 3 に示す．

前述したように，推薦に対して良い属性の粒度を求めることを主眼として，実験回数を増やすために LCS 作成のためのセッションをランダムサンプリングとし，テストセットのセッションも少ないものを使ったことから，再現率，適合率とも高くはないが，属性の粒度を変えたことによる違いは出ている．

表 3 から分かるように，Web ページ単位で LCS によって推薦を行うと，再現率・適合

表 3 属性の種類と粒度を変化させたときの推薦結果

属性の組合せ	適合率	再現率	F 値
Web ページ単位	0.143	0.033	0.054
小業態・小エリア・100 円単位平均予算	0.215	0.035	0.060
小業態・中エリア・100 円単位平均予算	0.304	0.047	0.081
小業態・大エリア・100 円単位平均予算	0.301	0.052	0.088
大業態・小エリア・100 円単位平均予算	0.305	0.051	0.087
大業態・中エリア・100 円単位平均予算	0.310	0.049	0.085
大業態・大エリア・100 円単位平均予算	0.290	0.054	0.091
小業態・小エリア・500 円単位平均予算	0.323	0.050	0.087
大業態・小エリア・500 円単位平均予算	0.494	0.075	0.130
小業態・小エリア・1000 円単位平均予算	0.588	0.085	0.149

率共に最も低かった．これは，ユーザセッションから求めたアクセスパターンに対して, アクティブセッションのユーザのアクセスパターンのほうが多すぎるためだと思われる．これに対して，業態，エリア，平均予算の属性に変換してから推薦を行った手法については，Web ページに対して推薦を行うよりも再現率も適合率も向上した．これは，複数の属性を用いることで Web ページ単位の推薦では対応できなかったアクセスパターンに関しても吸収できるようになったためと考える．また，時期が異なって新しい Web ページがある場合にも対応できていることも示している．

また，小業態・小エリア・100 円単位平均予算を用いた手法について，Web ページ単位で推薦した場合よりも適合率は上昇したものの，再現率については，ほとんど変化がなかった．これは，サンプリングによって差が出づらかったことと，Web ページに対する粒度の粗さの違いがあまりなかったことを示していると思われる．しかし，さらに粒度を粗くしていくと適合率・再現率ともに上昇した．これは粒度による影響が表れていることを示している．さらに粒度を粗くすると適合率が下がり始めるが，これは前述したように粒度を粗くしすぎたことにより推薦すべきでない Web ページも含まれるようになってしまったためではないかと思われる．ただ，今回の評価の範囲では，F 値としては粒度を上げても上昇している．

なお，平均予算の代わりに，業態，エリアの属性と一緒に座席数の属性を用いた実験も行った (ここでは，詳細な結果は省く)．座席数による LCS を用いた推薦に対する適合率・再現率は，業態，エリア属性の粒度を変化させても，Web ページ単位のものに比較して十分な優位性を示すことはできなかった．これは，座席数は実際の店舗選択に影響をあまり与えないためではないかと考察する．

以上のことから，推薦を行う Web ページの属性の選択，および選択した属性の粒度が推薦に大きく影響することが分かった．今回の実験ではカテゴリ, 範囲, 距離のタイプの属性を考慮したが，対象とする Web ページによっては，ここで検討したような属性が必ずしも存在するとは限らない．この属性の選択や粒度を自動的に調節を行うことができれば, 推薦精度を更に向上させることができるはずであるが, これについては今後の課題とする．

5.3 ま と め

本稿では，Web ページのアクセス履歴に基づく Web ページ推薦において，各 Web ページのそのもののアクセスパターンではなく，各 Web ページの持つ複数の属性のパターンに着目し，それらの LCS (Longest Common Subsequence) を抽出することでより適切な Web ページの推薦を行う手法を提案した．複数の属性を対象とすることで，多くのユーザに共通する傾向を適切に把握し，それを推薦に利用することができる．さらに，新規の Web ページを含むアクティブセッションを対象にしたり，新規の Web ページを推薦することも可能となる．

提案手法において，適切な推薦を行うためには，どのような属性をどのような粒度で使い，どのように組み合わせるかが重要となる．本稿では，どのような属性が考えられるか考察を行い，実際の商用 Web サイトにおけるアクセス履歴を用いて，従来の Web ページ単位の推薦と Web ページに対する複数の属性の組み合わせと粒度へ変化させた提案手法の比較を行った．実験の結果は，従来の Web ページ単位の推薦に対する提案手法の優位性と，粒度調整の有効性を示した．

Web ページそのもののアクセスパターンではなく，Web ページの持つ属性のアクセスパターンを用いるアプローチは，LCS を使った推薦手法だけではなく，相関ルールマイニングを使った推薦手法にも有効である．その評価に関しては今後の課題である．

また，得られた属性ベクトルから推薦する Web ページに対して今回は特に絞り込みを行わなかったが，実際の推薦においては大量の Web ページが推薦候補として表示されるとユーザビリティが下がるため，なんらかの順位付けを行って絞り込むことが必要になる．例えば，複数の属性について属性間の距離を考慮したり，対象とする属性の優先順位を付けたりする方法が考えられる．属性による推薦候補の絞り込みも今後の課題である．

この他，今回は平均予算に関しては，範囲をいくつかに区切って用いたが，ユーザにとっては「予算は 5,000 円以下」というような指定はあっても,「5,000 円でなくてはならない」といった要求は少ないと考える．そこで範囲のパラメータについては，事前にクラスタリングを行い，そのクラスタにしたがって本手法を適用することで，更に効率の良い推薦を行うことも可能である．

さらに，今回用いた属性に加えて，Web ページの持っているテキストデータや画像から抽出されるデータを使って推薦を行うことも考えられる．例えば，テキストデータに含まれる「有機野菜」,「アットホームな雰囲気」といった特長語は，ユーザが店舗を検索する際に重要な要素となっていると考えられる．しかし，これらをどのようにカテゴライズするかを検討する必要があり，これも今後の課題の一つとする．

評価においては，今回はテストセットの 2 アクセス目までをアクティブセッションとし，その後に続いたアクセスを正解として評価を行ったが，実際のユーザは意外な Web ページを推薦することを期待している場合もある．そこで，本手法および関連手法に関して実際のサイトに実装し，ユーザに対して推薦を行い，効果を検証する必要もある．

参 考 文 献

1) Linden, G., B.Smith and J.York: Amazon.com recommendations: Item-to-item collaborative filtering,, *IEEE Internet Comput.*, Vol.4, No.1 (2003).
2) 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).
3) 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).
4) 麻生英樹，小野智弘，本村陽一，黒川茂莉，櫻井彰人：協調フィルタリングと属性ベースフィルタリングの統合について，信学技報 NC2006-54(2006-10) (2006).
5) 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 (2002).
6) 黒川茂莉，小野智弘，本村陽一，麻生英樹，櫻井彰人：映画コンテンツ推薦のためのユーザ嗜好性モデルの実験的評価，信学技報 NC2004-182(2005-03) (2004).
7) 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，情報処理学会論文誌データベース No.SIG11(TOD34)， Vol.48 (2007).