アクセスログ分析における接頭部分列からのプラン認識

A Plan Recognition Method using Prefix Computation in Web Access Log Analysis

小島 諒介
Ryosuke Kojima

東京工業大学大学院情報理工学系研究科
Graduate school of Information Science and Engineering, Tokyo Institute of Technology
kojima@mi.cs.titech.ac.jp

佐藤 泰介
Taisuke Sato

(同上)
sato@mi.cs.titech.ac.jp, <http://sato-www.cs.titech.ac.jp/sato/>

keywords: plan recognition, prefix, PCFG, Web access log

Summary

In this paper we propose a new plan recognition method from observations of incomplete action sequences by regarding them as prefixes in a probabilistic context-free grammar (PCFG). In previous work that uses a PCFG for plan recognition, the PCFG receives a sentence, i.e. an observation of complete action sequences to recognize the plan behind it. However, when we deal with real plan recognition problems such as the Web access log analysis, we often cannot have complete sequences of actions and the traditional PCFG approach is not applicable. To overcome this difficulty, we extend the probability computation of PCFGs to prefix probability computation though it requires an infinite sum of probabilities. We applied the proposed method to infer the intended goals of Web site visitors from the online and partial observations of their actions. Also we compared the performance of plan recognition from observations of initial sequences of visitors' actions with that from full observations.

1. 背景

現在、主要な情報発信の手段として Web サイトが用いられ、特に企業などの Web サイトでは多様な情報が発信されている。それに伴い、同一サイトであっても利用者は調査・ニュース・商品の購入など様々な目的で訪れている。例えば、ショッピングサイトを持つ企業のサイトでは商品の購入を目的としたユーザだけでなく、その企業自体に興味のあるユーザも訪れる。ユーザがどのような目的を持ってサイトを訪れるかを知ることが、Web サイトの運営者にとって Web サイトの改善につながる有用な情報であり、訪れたユーザの目的をリアルタイムで知ることが、目的に応じた広告をページに表示することも可能である。そのため、ユーザの目的を知ることがサイトの分析や広告表示における主要な課題の一つになっている。

しかしながら、ユーザがサイトを訪れた目的は通常観測できないため、観測可能なアクセスログやその他のログを用いて推定する必要がある。観測されたユーザの行動から観測されないユーザの目的やプランを推定する問題はプラン認識の問題として多くの研究がなされ、ヒューマンコンピュータインターフェース [Lesh 99] やソフト

ウェアの理解 [Quilici 98] など幅広い範囲で応用されている。

プラン認識には、ユーザの行動はプランによって決定されるというモデルを置き、そのプランを階層構造で表現する方法がある [Kautz 91]。そのような枠組みの一つとしてプラン認識を構文解析の問題として捉える方法がよく使われる [Vilain 90]。構文解析では規則 (文法) によって文が導出されると考え、与えられた文に対して導出過程を表現した構文木を出力する。ここで文をプラン認識における行動列と置き換えて考えると、行動列の生成過程を示した構文木をプランと考えることができる。この時、文の構成要素である単語は行動列を成す単位行動 (以降、単に行動と呼ぶ) と対応する。このようにプラン認識の問題を構文解析の問題に置き換えて考えることにより、言語の分野で有効であることが知られている手法をプラン認識の問題へと応用できる可能性がある。

このような応用の例として自然言語の分野における確率文脈自由文法 (probabilistic context-free grammar: 以降、PCFG) [Manning 99] がある。プラン認識の多くの問題においてユーザの目的は不確実であるためそれを表現することのできるモデルが求められており、言語における不確実性を扱うモデルを利用できる [Pynadath 98]。

ここで、Web サイトを訪れるユーザの目的推定を上述のプラン認識の枠組みに適用してみよう。すると、観測されるユーザの行動列は Web サイトのアクセスログとして記録されており、そのログデータは目的から生成されていると捉えられる。このとき、確率的な規則を利用できる PCFG はユーザの不確実な目的の推定に有効と考えられる。

しかし、PCFG を単純にプラン認識の問題に適用しても問題の要求に十分に答えられない。PCFG では文全体をパースする必要があるため、行動列全体が観測されている必要がある。しかし、アクセスログからオンラインで目的を推定し、広告表示を行うといった場面では開始から現在までの部分的な行動列(接頭部分列)のみが観測される状況で目的を推定する必要がある。また、ユーザが目的を果たすことなくサイトから離脱した場合、同様にアクセスログは接頭部分列のみとなる。このような場面に於いて PCFG を用いたプラン認識を行うことはできない。計算言語学の分野ではこのような接頭部分列に対して prefix 確率を導入する方法が提案されている [Jelinek 91]。本研究では PCFG のプラン認識を prefix 確率を導入することにより拡張し、接頭部分列を扱うことができる手法を提案する。さらに、提案法を用いた目的推定を実際のアクセスログに対して適用し、実験により評価する。

2. PCFG を利用したプラン認識

構文解析を用いてプラン認識を行う手法はいくつか提案されているが、ここでは提案法と関連の深いプラン認識手法として PCFG を用いる手法を説明し、PCFG を用いてアクセスログからプラン認識を行う方法について説明する。

2.1 PCFG

PCFG [Manning 99] は各導出規則に確率が付与されている文脈自由文法 (context-free grammar: 以降、CFG) である。まず、CFG G を四つ組 $\langle N, \Sigma, N^1, R \rangle$ で定義する。この時、 N は非終端記号の有限集合、 Σ は終端記号の有限集合、 R は文法規則の有限集合である。また、 N の要素を N^i で表し、特に N^1 は開始記号である。 R の要素である規則 r は $N^i \rightarrow \zeta$ という形式で与えられるとする。ただし、 ζ は $\Sigma \cup N$ の要素ひとつ以上からなる列であるとする。規則 r は左辺を右辺で置き換え可能であることを表し、規則を N^1 に非終端記号がなくなるまで適用することにより得られる終端記号の列を文と呼ぶ。規則の適用過程は導出と呼ばれ、ここでは最も左の非終端記号から規則を適用していく最左導出を用いることとする。

次に、CFG G に対する PCFG G_Θ を定義する。ここで、 Θ は PCFG の規則 r に対応するパラメータ θ_r の集合である。ただし、 θ_r は $0 < \theta_r < 1$ 、 $\sum_{\zeta: N^i \rightarrow \zeta} \theta_{N^i \rightarrow \zeta} = 1$ であるとする。PCFG では「適用する規則は他に影響さ

れずに選択される」と仮定されるので、文 w を導出するために規則列 $r = (r_1, r_2, \dots, r_n)$ を用いたとすると r による文 w の出現確率 $P(r|\theta)$ は

$$P(r|\theta) = \prod_{j=1}^n \theta_{r_j}$$

と計算される。さらに、ここで扱う PCFG は導出されるすべての文の出現確率の和が 1 になるという性質を満たす、即ち整合性 (consistency) を仮定する。この性質は有限サイズのコーパスから EM アルゴリズムなどの最尤法によりパラメータ学習する場合には満たされることが保証されている [Chi 98]。

PCFG によってある文 w の導出される確率を計算するためには文 w の導出する規則列の集合を R_w とすると

$$P(w|\theta) = \sum_{r \in R_w} P(r|\theta)$$

として計算できる。また、文 w が与えられた時、 w を導出する尤もらしい規則列 $r_v = \operatorname{argmax}_{r \in R_w} P(r|\theta)$ を得る方法として、Viterbi アルゴリズム [Jelinek 85] が知られている。

以上より、PCFG を用いることによって文が与えられた時、それが導出される確率が計算でき、尤もらしい構文木を得ることも可能である。

2.2 アクセスログからのプラン認識

アクセスログはどのような Web サイトであっても比較的容易に手に入るユーザの行動履歴であるが、多くのノイズが存在する。それらを緩衝するため不確実性を扱うことのできる PCFG の利用が考えられる。ここでは、アクセスログから PCFG を用いてプラン認識を行う具体的な方法について論じる。

まず、アクセスログからユーザの行動を取り出す方法について説明する。ここで考えるアクセスログとは、ユーザのドメイン名、タイムスタンプ、アクセスしたページのパスが記録されているデータである。アクセスしたページのパスはサイト内の階層構造を表現したものであり、例えば/(スラッシュ)を区切り文字として「/top/index.html」のように表される。ここではユーザの行動をこの階層構造を持つサイト内を巡回する行動として表現する。具体的にはユーザの行動として、階層構造を上位の階層に移動する (up)、下位の階層に移動する (down)、同じ階層の別のページへ移動する (sib)、同じページを再度要求する (same)、その他の移動 (move) の 5 種類を考え、ユーザはこれらの行動を目的に応じて選択するものとする。例えば「/top/index.html」から「/top/child/a.html」への移動は down とし、「/top/index.html」から「/top/b.html」への移動は sib とする。ユーザの行動は Web サイトの種類によって異なることが予想されるが、本論文ではサイトの階層構造を利用した汎用的なこれらの行動のみを利用することとする。

表1 プラン認識を行うための規則の例

左辺が同じ規則については | で区切って、右辺のみ示した。ただし、開始記号を S とし、小文字で始まる単語は終端記号を表し、大文字で始まる単語は非終端記号を表すとする。

	規則
1	$S \rightarrow \text{Survey}$
2	$\text{Survey} \rightarrow \text{Search Destination}$
3	$\text{Search} \rightarrow \text{Down Up Search} \quad \quad \text{Down}$
4	$\text{Destination} \rightarrow \text{sib Destination} \quad \quad \text{sib}$
5	$\text{Down} \rightarrow \text{Down down} \quad \quad \text{down}$
6	$\text{Up} \rightarrow \text{Up up} \quad \quad \text{up}$

次に、プラン認識を構文解析の問題とみなして計算するためには、プランから行動を生成する規則(文法)を作る必要がある。規則の例を表1に示す。ここでは大文字から始まる単語は非終端記号を表し、小文字から始まる単語は終端記号を表すとする。1番目の規則はWebサイトを訪れた人の主たる目的を表し、ここでは簡単のため調査(Survey)を目的として訪れた人のみを考える。調査(Survey)を目的として訪れた人は始めにサイト内の探索(Search)をし、目的のページに到着(Destination)の後、そのページと同一階層のページにアクセスするというものが2番目・4番目の規則である。ユーザは探索(Search)をする際にはサイトの階層間を上下に移動しながら探索を行うということを示したのが3番目の規則である。また5番目・6番目の規則は階層間の上移動と下移動の繰り返しを表現したものである。ここでは、1番目の規則のように開始記号を左辺に持つ規則は右辺は非終端記号一つで表現することとし、このような非終端記号をトッププランと呼び、ユーザがWebサイトを訪れた主たる目的を表現するよう規則を作成する。

これらの規則に従ってユーザの行動から尤もらしい構文木を得た結果の例が図1である。この図はユーザが主たる目的(トッププラン)である「調査 (Survey)」からどのような導出過程を経て観測された行動に至ったかを示している。例えば、この図を見ることで目的地への到着 (Destination) まで探索 (Search) を何回行ったかなどの情報を得ることができる。さらに、この情報を利用して探索 (Search) の回数の多いページを改善し、ユーザの目的地への達成をより容易にするといったサイトの改善への応用も考えられる。このようにプランを表現した構文木を出力できることは階層的なプラン認識を利用する特徴であり、アクセスログの分析に於いてはこの特徴をサイトの改善に活用できる。

3. 提案法

PCFG によってパースできる行動列全体を完全行動列と呼ぶこととする．完全行動列が与えられた場合には，前節の手法によりプラン認識を行うことが可能である．しかし現実の問題においては，開始から終了までの完全行

図1 表1の規則を用いた場合の構文木の例

終端記号は観測された行動列に対応し、非終端記号は観測されないユーザの目的に対応する。

動列が与えられることは少なく、完全行動列の開始からの部分列である接頭部分列からプラン認識を行いたいという要求がある．例えば Web サイトでリアルタイムでユーザのプランを推定し、適切なリンクや広告を表示する場合には、オンラインでプラン認識を行う必要があり、この場合、接頭部分列のみが観測される．また、ユーザが本来の目的を達成することなく Web サイトを離脱した場合も、観測されるのは接頭部分列のみである．こういった接頭部分列のみが与えられた場合に、プラン認識に PCFG が利用できないという問題点を解決する手法として prefix 確率を用いる手法を提案する．

3.1 prefix 確率計算

PCFG では規則と文に対して確率が定義されているが、文の部分列に対しては確率が定義されていない。PCFG を利用して接頭部分列からプラン認識を行うには接頭部分列に対して確率を定義する必要がある。そこで Jelinek らによって提案されている prefix 確率を採用する [Jelinek 91]。この節では、PCFG における prefix 確率の定義と計算法について述べる。

開始記号 N^1 から文 $\mathbf{w} = (w_1, \dots, w_k, \dots, w_n)$ が導出される時の $\mathbf{w}_k = (w_1, \dots, w_k)$ を接頭部分列と呼び、prefix 確率を同一接頭部分列を持つ文の出現確率の総和と定義する。しかし接頭部分列を扱う場合には、CFG においてある接頭部分列をもつ文は無限に存在するという困難さが伴う。そのため prefix 確率は一般に無限和で表現される。

例えば，開始記号を S として

(規則 1) $S \rightarrow SS$ (規則 2) $S \rightarrow a$

の 2 つの規則が与えられた時に接頭部分列を a, a とすると, S から a, a の導出は有限であるが S から左端が S となる導出は規則 1 を何回適用しても良いため導出が無限に考えられる.

開始記号 S から接頭部分列 w_k の出現確率

$Pres(w_k)$ を接頭部分列に w_k を持つすべての文の出現確率の和として定義する．前の例の規則 1, 規則 2 に対応する確率を θ_1, θ_2 とすると, 接頭部分列 a, a の出現確率 $Pres(a, a)$ は

$$Pres(a, a) = \theta_1 \theta_2^2 (1 + \theta_1 + \theta_1^2 + \dots)$$

と計算できる．この式の $\theta_1 \theta_2^2$ は S から a, a を導出する確率である．また, 無限和はそれぞれ規則 1 を適用しなかった場合, 1 回適用した場合, 2 回適用した場合, ... に対応する．

これを一般化する． w_k の導出を考えると, 接頭部分列 w_k から構文木を逆に辿って合流する最初の非終端記号 N^x が存在する．ここで導出を N^1 から N^x までと N^x から終端記号までに分離して考える．また, 接頭部分列でない部分を導出する非終端記号を根とする部分構文木の確率の総和は整合性 (consistency) の仮定と各規則の確率が 0 より大きいことを仮定すると 1 となる [Chi 98]．これらを仮定することにより N^x から w_k の導出確率の和は有限和となり, PCFG の場合と同様にその確率の総和 $q(N^x \Rightarrow w_k)$ を計算することができる．

一方, N^1 から N^x までについては有限和の場合と無限和を計算する必要のある場合がある．導出が有限の場合もしくは $N^1 = N^x$ の場合は有限和として計算を行うことが可能であるが, 導出が無限に存在する場合には無限和となる．以降は無限和となる場合のみを考える．

規則 $N^i \rightarrow \zeta$ の右辺の記号列 ζ の左端が N^j の規則の集合を R_{ij} として, 左端の文字のみに注目した確率 $P_L(N^i \rightarrow N^j)$ を以下のように定義する．

$$P_L(N^i \rightarrow N^j) \equiv \sum_{r \in R_{ij}} \theta_r$$

また, N^i から左端が N^j となる導出に関する無限和の量 $Q_L(N^i \Rightarrow N^j)$ を

$$\begin{aligned} Q_L(N^i \Rightarrow N^j) &\equiv P_L(N^i \rightarrow N^j) \\ &+ \sum_{A_1} P_L(N^i \rightarrow A_1) P_L(A_1 \rightarrow N^j) \\ &+ \sum_{A_1, A_2} P_L(N^i \rightarrow A_1) \\ &\quad P_L(A_1 \rightarrow A_2) P_L(A_2 \rightarrow N^j) \\ &+ \dots \end{aligned}$$

と定義する．ただし, $A_1, A_2, \dots$ は非終端記号とする． $Q_L(N^i \Rightarrow N^j)$ は, 左端の非終端記号に関して N^i から N^j を導出する確率の和である．ここで, N^i, N^j は任意の非終端記号なので, i 行 j 列目の要素が $Q_L(N^i \Rightarrow N^j)$ の行列を Q_L とすると, i 行 j 列目の要素が $P_L(N^i \Rightarrow N^j)$ の行列 P_L をもちいて

$$Q_L = P_L + P_L^2 + P_L^3 + \dots$$

と書き表せる．このとき,

$$Q_L P_L + P_L = Q_L$$

が成り立っていることを利用して線形方程式を解くことにより Q_L を求めることができる^{*1}．ここで, Q_L の要素 $Q_L(N^1 \Rightarrow N^x)$ を取り出すと N^1 から N^x までの導出の確率の総和を得られる．したがって, 可能なすべての N^x に関しての和をとった次の式により prefix 確率を計算できる．

$$Pre_{N^1}(w_k) = \sum_{N^x} Q_L(N^1 \Rightarrow N^x) q(N^x \Rightarrow w_k)$$

ただし, 可能な N^x は非終端記号の数以下であり, 有限である．

3.2 接頭部分列からのプラン認識

ユーザの目的に応じた広告をリアルタイムに表示する場合や途中で離脱したユーザの目的を推定したい場合には, 接頭部分列からプランを推定する必要がある．接頭部分列の出現確率を prefix 確率により定義することで, 接頭部分列からプラン認識を行うことが可能である．ここではその具体的な方法について述べる．

接頭部分列からであっても PCFG と同様に Viterbi アルゴリズムにより構文木を推定することで尤もらしいプランを推定することができる．一方, 接頭部分列からユーザの主たる目的であるトッププランを予測する場合には prefix 確率を計算する必要がある．

観測として接頭部分列 w_k が与えられた時のトッププラン A を推定するには, A から接頭部分列 w_k をもつ文が導出される確率の総和 $Pre_A(w_k)$ を用いて以下の式により求める．

$$A = \underset{A}{\operatorname{argmax}} Pre_A(w_k) \theta_{N^1 \rightarrow A}$$

具体的には, $Pre_A(w_k)$ は前節の開始記号 N^1 をトッププラン A と置き換えることにより求め, すべての考え得るトッププラン A において $Pre_A(w_k) \theta_{N^1 \rightarrow A}$ を計算することでトッププランを推定する．

4. 実 験

部分列からトッププランの推定を行う問題において, 提案手法が有効であることを確認するために, 人工データと実データを用いて実験を行った．

*1 この線形方程式については更に小さい線形方程式に分割して解くことができる [Jelinek 91] [Sato 12]．後述の実験では小さい線形方程式に分割した後, 線形代数ライブラリ LAPACK (<http://www.netlib.org/lapack/>) に含まれる LU 分解に基づく線形方程式ソルバを利用した．

表2 クラスタ分析結果

クラスタリングの結果 5 つのクラスタに分類され、それらを人手で確認した。

	目的	特徴
1	幅広い調査	サイトの階層構造の上下に行動
2	ニュースなど時間に関係した調査	サイトの階層構造の上下に行動+同じページへのアクセス
3	特定分野の調査	同じ階層のページへのアクセス
4	特定分野のニュースなどの調査	同じページへのアクセス
5	その他	サイトの階層構造に沿わない行動

4.1 実験データ

アクセスログのデータは The Internet Traffic Archive[ITA 01] から U of S(University of Saskatchewan), ClarkNet, NASA の Web サーバのデータを用いた [Arlitt 96]。アクセスログには画像などの HTML ファイル以外へのリソースやアクセスエラーなどの情報も含まれるが、これらの情報は取り除き、成功した HTML ファイルへのアクセスのみを用いて実験を行った。また、ページへのアクセス間隔が 24 時間以内の同一ドメインからのアクセスを一つのセッションとした。このセッション内のログからユーザの行動列を抽出した。抽出した行動列は 2.2 節で用いたものと同様である。この時抽出した行動列の長さが短い行動列や長い行動列では完全行動列とならず、PCFG のプラン認識が失敗することが多い。このような場合でも、提案法により取り扱い可能な場合もあるが、実験において完全行動列と比較できないため今回は行動列の長さ 20 以上 30 以下のデータのみを用いた。ここで長さを 30 以下に制限した理由は計算時間によるもので、特に学習において非常に計算時間がかかるためである。また 20 以下の行動列では目視で目的を確認することが難しく、PCFG の規則を作成することが困難であったためである。得られたデータは全て目的を達成している完全行動列であると仮定し、実験を行った。これらの前処理を U of S の 7 ヶ月分のデータ (June 1/95 - Dec. 31/95), ClarkNet の 7 日分のデータ (Aug. 28/95 - Sep. 3/95), NASA の 1 ヶ月分のデータ (July 1/95 - July 31/95) に適用した結果、それぞれ 652 本, 4523 本, 2014 本の行動列が得られた。

4.2 汎用 Web アクセス文法の構築

実際には、Web サイトを訪れる人の目的はそのサイトの運営者であれば推測できる。そのため、規則は Web サイトの運営者によってサイトに合わせて作成されるのが理想であるが、今回は著者が手動で規則を作成した。手動で規則を作成するための参考とするため、事前にアクセスログデータに対してクラスタ分析を行ない、5 種類の目的があることを確認した。5 種類の目的と特徴は表 2 に示した。この時クラスタリングには文法が等しくパラメータの異なる PCFG を確率的に選択するモデルである混合 PCFG を利用した。ただし、混合 PCFG において

表3 汎用 Web アクセス文法の一部

汎用 Web アクセス文法のうちトッププラン「Survey」に関連した規則のみ抜粋した。StartSurvey, EndSurvey はすべての行動を導出する。InternalSearch 階層の上下を行う行動を導出する。Destination は表 1 を参照。Survey1, Survey2 は InternalSearch と Destination を交互に繰り返すための非終端記号。

	規則
1	S → Survey
2	Survey → StartSurvey Survey1 EndSurvey
3	Survey1 → Destination Survey2 Destination
4	Survey2 → InternalSearch Survey1 InternalSearch

はメモリの制約上大規模な規則を扱うことができなかったため^{*2}、クラスタリング用簡易文法(規則の数は 30 個、非終端記号数 13)を使用し、4.1 節で用意したデータの内 U of S については全データを利用し、ClarkNet, NASA についてはランダムに行動列を 1000 本選択し利用した。

今回はこの 5 種類の目的を参考に、これらに対応する 5 つのトッププランを持つ文法を作成し、実験に用いた。この時用いた規則の数は 102 個であり、非終端記号の数は 32 個である。以降、この作成した文法を汎用 Web アクセス文法と呼ぶ。汎用 Web アクセス文法のうちトッププラン「Survey」と関連のある文法の一部を表 3 に示す。この文法は表 1 で示したものと似ているが、幾つかの点でさらに工夫をした文法である。まず、実際のアクセスログデータを見ると行動列の序盤と終盤においては不規則な行動を行っていることが多くあったため、行動列の序盤と終盤のための非終端記号 StartSurvey と EndSurvey を用意し、これらはすべての行動とその繰り返しを導出できるようにした。次に行動列の中盤では「目的のページとその周辺の調査 (Destination)」と「目的のページの探索 (InternalSearch)」を繰り返していると考え、非終端記号 Survey1 と Survey2 を作成した。また、InternalSearch では階層の上下を主な行動とし、Destination は表 1 のものと同じである。

*2 混合 PCFG の実装には記号的統計モデル言語および処理系である PRISM[Sato 01, Sato 12] を利用し、実験ではメモリサイズが 72GB の Core i7 Quad 2.67GHz OpenSUSE 11.4 の計算機を使用した。

図2 HMM 人工データ (左図), PCFG 人工データ (右図)

人工データの生成に HMM(左図) と PCFG(右図) を利用, 正解トッププラン作成に汎用 Web アクセス文法の PCFG, 予測において提案法 (prefix) と HMM を使用し, 横軸に接頭部分列の長さ, 縦軸に正答率をとって比較した。

4.3 実験方法

実験の目的は, 完全行動列から予測されたトッププランと提案手法によって接頭部分列から予測したトッププランがどの程度一致するかを調べ, 提案法と他の手法で性能を比較することである。そのために 4.1 節で用意した完全行動列から PCFG の EM アルゴリズムによってパラメータを学習して, トッププランの推定を行い, これを正解トッププランと定義する。ここで, 正解トッププランは PCFG の文法により変化することに注意する。この正解トッププラン付きの行動列データを訓練データとテストデータに分け, 訓練データとして完全文と対応するトッププランを利用し, テストデータとして完全文から長さ k の接頭部分列を作成し利用する。提案法ではパラメータを訓練データから EM アルゴリズムで学習し, テストデータで接頭部分列からトッププランを予測し正答率を調べる。比較のための他の手法についても同様だが, 詳細は後述する。ここで利用した EM アルゴリズムは局所解しか得られないため 100 回再スタートを行い, 尤度が最も高くなるパラメータを採用した。

4.4 実験結果・考察

まず, 実験 1 では人工データを用いてデータや文法の違いによるプラン認識結果の違いを探る。次に, 実験 2 では実データを用いて他の手法と比較する。最後に, 実験 1 と実験 2 を踏まえた考察を行う。

§1 実験 1

実験 1 では 2 種類の人工データを用いて実験を行う。1 つは, 自己ループを持つ 3 状態 Left-to-right 隠れマルコフモデル (以降, 単に HMM) を 5 種類用意し, それぞれから生成した長さ 20 の行動列を混合した 1000 本の人工データである。ここで考えた 5 種類の HMM は表 2 に記された 5 つのトッププランに対応するものであり, それぞれの 3 状態 HMM は状態によって異なる確率で 5 つの行動を選択するモデルである。もう一つは, 5 種類の生成用の規則を用いて PCFG から長さ 20 以上 30 以下の

行動列を生成し, それらの行動列を混合した 1000 本の人工データである。両モデルにおいて生成用のモデルのパラメータは事前のクラスタリング結果を参考に手動で与えた。

このように作成したデータを用い, 汎用 Web アクセス文法により PCFG の学習及びプラン認識を行いトッププランを推定した。このときの推定したトッププランを正解トッププランと定義して, 接頭部分列からトッププランを推定するという問題を考え, 提案法と同じく接頭部分列からクラスを予測できる HMM を用いた推定方法と比較する^{*3}。ただし, HMM から生成されたデータのうち 21 本は PCFG の文法でパースできなかったため取り除いて実験を行う。結果は 5-fold cross-variation を行なって評価する。

評価の際, 訓練データを用いた学習場面では, 正解トッププランごとに HMM を作成し EM アルゴリズムによって学習する。テストデータのトッププランの推定においては長さ k の接頭部分列から各 HMM の尤度を計算し, 尤度を最大化するものをトッププランと推定する。2 種類の人工データに対して, 長さ k を 2~20 で変化させた時の正答率を示したグラフが, 図 2 である。ただし, HMM の状態数を 2~8 で変化させて実験を行い, グラフ上では正答率が最大の結果のみを示している。

この結果より接頭部分列の長さが長い時, 両データ共に HMM よりも提案法の方が正答率が高いことが分かる。また, HMM で生成したデータと PCFG で生成したデータを比べると正答率の上昇の様子が大きく異なることが分かる。これは特にデータが短い部分列の時に顕著であることから, PCFG では規則からデータを生成したため, 短い部分列からでもトッププランが推定可能であったと考えられる。

また, PCFG の文法による結果への影響を明らかにするためより簡単な文法を用いて実験を行った。図 3 は PCFG

*3 prefix 確率計算, PCFG, HMM の実装には記号的統計モデル言語および処理系である PRISM を利用する [Sato 01, Sato 12]。

図3 PCFG 人工データ (正解トッププランを簡易文法で作成)

人工データの生成に PCFG を利用，正解トッププラン作成に簡易文法の PCFG，予測に提案法 (prefix) と HMM を使用し，横軸に接頭部分列の長さ，縦軸に正答率をとって比較した。

表4 生成時のトッププラン予測

モデルから人工的に生成したデータから文法・HMM のパラメータ学習を行い学習したモデルを用いて生成時のトッププランを推定した時の正答率。サンプルの数は 1000 とし，5 回同じ実験を行った。

	HMM	PCFG
正答率 (平均)	55.86%	87.38%
分散	1.04	0.57

から生成したデータに対して，汎用 Web アクセス文法ではなく，簡易文法 (規則数 42 個，非終端記号の数 19 個) を用いて正解トッププランを作成した場合の実験結果である。この結果より文法が単純な場合には HMM と提案法にほとんど差が無い事が分かる。したがって，複雑な文法によって決定されるプランを接頭部分列から推定する場合には提案法が HMM に比べ有利であると言える。

この実験では完全な行動列からの PCFG による推定結果を正解トッププランとしているが，参考としてデータ生成時のトッププランを正解とした時の正答率も示しておく。人工データを生成用 HMM，生成用 PCFG (簡易文法を利用) から作成し，サンプルしたデータからパラメータ学習を行い学習後のモデルを用いて生成時のトッププランを推定した時の正答率を表 4 に示す。なお，サンプルの数は 1000 とし，5 回サンプルしなおして平均と分散も示した。この結果より，PCFG の場合にはデータ生成時の文法と予測時の文法が完全に一致していれば 87% 程の正答率で生成時のトッププランを推定可能だということがわかる。このことから，実データではこのような生成時のトッププランや生成時の文法は未知であるが，良い文法を利用することができる場合には完全な行動列からの PCFG による推定はトッププランの推定として有効であると考えられる。

§2 実験2

実験2では実データに対して実験1同様の実験を行い，現実のアクセスログからのトッププランにおける結果を

表5 各データの汎用 Web アクセス文法におけるトッププランの割合

各データを完全行動列から汎用 Web アクセス文法を用いてトッププランを推定した際の各トッププランの偏りを示す。この表の行の番号 1~5 は表2と対応している。

	U of S	ClarkNet	NASA
1	42.03 %	29.44 %	0.99 %
2	2.47 %	4.60 %	69.71 %
3	6.27 %	3.06 %	22.69 %
4	42.36 %	55.21 %	4.07 %
5	6.85 %	7.66 %	2.53 %

確認する。U of S, ClarkNet, NASA のデータのうちそれぞれ 1 本，1 本，2 本は汎用 Web アクセス文法でパースできなかった。このようなデータは今回はデータの数 (2014 本，4523 本，652 本) に比べ非常に少ないので取り除いて実験を行った。また，この時の各データにおけるトッププランの割合を表5に示す。

実データに対する実験では HMM の他に，ロジスティック回帰とサポートベクトルマシン (SVM) による予測の実験結果も比較として用いる^{*4}。これらの手法は不定長には対応できないが，今回の実験では接頭部分列の長さ k を固定して実験を行なっているので，先頭 k 個のデータのみを切り出して作成した訓練データから識別器を構成した。また，SVM については行動を 5 値の k 次元ベクトルとして入力する方法の他に行動の頻度を用いる方法 (bag-of-words) の 2 種類を用いた。

加えて文法とそのパラメータ学習は提案法と同様に行い，トッププランの推定において prefix 確率ではなく PCFG によるトッププランの推定も比較として用いる。この場合，接頭部分列を完全行動列とみなして PCFG により最尤のトッププランを推定することとなる。ただし，提案法では完全行動列の接頭部分列をであれば必ず最尤のトッププランを推定できるのに対し，接頭部分列を完全行動列とみなす方法では，完全行動列の接頭部分列であってもパースできるとは限らず，トッププランを推定できない場合がある。今回の実験ではそのようなパースできない行動列は行動列の長さが 3 以上の時には U of S, ClarkNet, NASA のデータでそれぞれ 2 本，3 本，1 本であり，長さが 2 の時にはすべて行動列がパースできなかった。そのため，接頭部分列を完全行動列とみなした PCFG による方法ではこれらを取り除いて正答率を計算した。

アクセスログから提案法と他の手法を用いて部分列からトッププランを推定した結果を図4に示す。この実験結果より，人工データの場合と同様に部分列の長さが長い時には提案法が HMM，ロジスティック回帰，SVM より優位な傾向があり，一方，U of S や ClarkNet は短い部分列

*4 ロジスティック回帰，SVM の実装には統計解析用のプログラミング言語および処理系の R に含まれる nnet, kernlab パッケージの実装を利用する。

のとき提案法より HMM, ロジスティック回帰, SVM のほうが優位な傾向であると分かる。また, NASA のデータでは提案法と他の手法の差が大きいのに対し, U of S や ClarkNet のデータではそれらの差が小さいことがことも分かる。また, 接頭部分列を完全行動列とみなした PCFG では行動列の長さが短い時に他の方法より正答率が低く, 行動列が長い時であっても提案法とそれほど変わらないことが言える。

また, この実験では完全な行動列からの PCFG による推定結果を正解トッププランとしているが, 参考として 4.2 節で行った混合 PCFG を用いたクラスタリングによる結果を正答とした時の提案法と HMM の正答率を図 5 に示す。この結果は図 4 の結果と比べて提案法と HMM で差がほとんどないことが分かる。これはクラスタリングにおいてメモリの制約から小規模なクラスタリング用簡易文法を使用したためであると考えられる。この結果は簡易文法を使用した実験 1(図 3)における結果と同様の傾向であり, 単純な文法によって決定されるプランを接頭部分列から推定する場合には提案法と HMM の差が小さいことが確認できる。

§3 考 察

実験 1 と実験 2 の結果と合わせて考えると, すべてのデータにおいて部分列の長さが長い時には提案法が他の手法より優位であった。しかし, 提案法と他の手法の正答率の差はデータごとに異なる結果となった。実験 1 では正解トッププランの作成に異なる文法を用いた実験をしており, 複雑な文法の方が提案法と他の手法の差は顕著であった。実験 2(図 4)では汎用 Web アクセス文法を用いており, PCFG のパラメータのみ異なる。ここでこれらのモデルの違いを数値的に確認するために PCFG のエントロピー [Chi 99] を計算する。PCFG のエントロピーは導出される構文木を t として $-\sum_t p(t) \log p(t)$ で定義される量である。実際にエントロピーを計算すると図 2(右図)の実験では 3.90×10^6 であるのに対し, 図 3 の実験では 2.09×10^6 であった。したがって, 実験 1 では正解トッププランの作成にエントロピーが高いモデルを用いた場合の方が, 提案法と他の手法の差は顕著であったともいえる。また汎用 Web アクセス文法を用いた実験 2 についても同様にエントロピーを計算すると, 提案法とその他の手法の差がそれほど大きくない U of S と ClarkNet ではエントロピーはそれぞれ 5.12×10^4 , 2.77×10^5 であったのに対し, 差が顕著であった NASA では 3.14×10^6 と 10 倍以上大きい値であった。このことから, 正解トッププランの作成時のモデルのエントロピーが実験の結果と関係しており, エントロピーの高いモデルであるほど提案法が優位な傾向があると分かった。そのため, このエントロピーが提案法を利用するかどうかの指標になると考えられる。

ここで, 各手法間のモデル化の違いについて見てみると, ロジスティック回帰や SVM では時系列による関係を

上手くモデル化できていないため, HMM や提案法と比べ特に部分列が長い時には不利であったと考えられる。また, 提案法では完全行動列からのプラン認識において利用した文法と同じ文法を部分列行動列からのプラン認識において再利用することが可能である。HMM は PCFG の文法が特殊な場合と考えることもできるので, HMM の結果は完全行動列からのプラン認識に利用した文法をあえて利用しなかった場合の結果とも言える。このことから正答率という点では, 部分列行動列からのプラン認識の枠組みを使って文法の情報を用いた方が有利であるということが言える。ただし, 速度の点からは提案法は他の手法に比べ遅い。今回の各データに対して長さ 20 の場合において予測にかかる時間の平均は HMM と提案法では 10 倍程度^{*5} の差がある。しかし, その時の平均時間はすべてのデータで 1 秒以内となっていたため広告表示や Web サイトの分析といった今回の用途としては十分な速度である。また, 行動列の長さが短い場合には予測性能が他の手法に劣るため, 他の手法との併用をすることも考えられる。

しかし, ここで上げた提案法以外の手法はトッププランの推定のみ可能なのに対し, 提案法では構文木を出力することによりトッププランから観測された行動列をどのように導出したのか説明できる。これは Web サイトの分析に於いて, サイト内のページがユーザにどのように利用されているかというサイトの改善に有用な情報であるため, トッププランの推定性能が劣るような場面であっても提案手法を利用する価値はある。

5. 関 連 研 究

ここでは, プラン認識において確率や構文解析手法を用いる既存の手法と本研究の関係について論じる。

まず, 本研究と同様に確率文法を用いている手法として制限された PCFG からペジアンネットワークを構築する手法 [Pynadath 98] と PCFG を拡張した PSDG (Probabilistic state-dependent grammars) を用いる手法 [Pynadath 00] がある。しかし, これらの手法では PCFG の文法に左再帰を含まないなどの制限された文法を使っている。一方, CFG の左再帰の除去を用いてこの制限を用いる手法 [Geib 11] も提案されているが, 左再帰の除去を行う過程で与えられた規則を変形する必要があり, 出力される構文木で用いられる規則は入力された規則と一致しない。また, HHMM (Hierarchical HMM) [Fine 98, Avrahami-Zilberbrand 07] や AHMM (Abstract HMM) [Bui 02, Bui 03] を用いる方法も考えられるが, これらも左再帰を含まない PSDG の特殊な場合である。しかし, 実際のプラン認識では左再帰を用いた方がより自然に記述できる場合がある。例えば本実験で用いた階層構造上での行動モ

*5 実験では Core i7 Quad 3.40GHz OpenSUSE 11.4 の計算機を使用した。

図 4 実データ: U of S(上図), ClarkNet(中図), NASA(下図)

正解トッププラン作成に汎用 Web アクセス文法の PCFG, 予測において提案法 (prefix), HMM, ロジスティック回帰 (LR), 観測された行動列を特徴とした SVM, 行動の頻度を特徴とした SVM(SVM-BoW), 接頭部分列を完全行動列とみなした PCFG を利用し, 横軸に接頭部分列の長さ, 縦軸に正答率をとって比較した。

デルでは, 過去に階層を上下している行動 (UpDown) をしていた時, 次にとる行動は up か down であるということ を自然に表現すると「UpDown → UpDown down | UpDown up」となり, 左再帰を含む形となる。本研究ではこれらの文法に対する制限を取り除き, かつ入力された規則と一致する規則で記述された構文木による出力を同時に可能にしている。また, 本手法は PCFG を拡張した文法 [Pynadath 00] においても同様の拡張が可能であると考えられる。

また, プラン認識としてはベイジアンネットワークを利用したものも研究されている [Charniak 93] が, グラフサイズが固定のため今回用いたアクセスログのような不定長の行動列からプラン認識を行うといった場合に上手く対応できないといった問題がある。一方, 固定長の行動列からプラン認識を行う際にはこのような手法は有効であり, Web アプリケーションでの行動 (クエリ) から

図 5 実データ (クラスタリング用簡易文法を使用): U of S(上図), ClarkNet(中図), NASA(下図)

4.2 節で行った混合 PCFG を用いたクラスタリングによる結果を正答とした時の提案法と HMM の正答率を示す。提案法で用いた文法はクラスタリング時の文法と同一のものを使用し, その文法は汎用 Web アクセス文法に比べ簡易的な文法である。横軸は接頭部分列の長さ, 縦軸は正答率である。

ベイジアンネットワークを利用して目的を推定するという研究も為されている [Lau 99]。さらに, プラン認識の Web への応用として他にも Web サーバのログから攻撃者の判別と攻撃の種類の特定制の研究もされている [Geib 01]。

6. ま と め

本研究では PCFG を利用したプラン認識を接頭部分列のみが与えられた場合にも適用出来るよう拡張した。その際に prefix 確率を用いることで, 文法規則に特別な制限を設けず, 規則の書き換えも行うことなく拡張できた。また, アクセスログの実データに対し提案手法を適用し, アクセスログの分析やユーザの目的の推定として利用可能であることを示した。さらに, 実験により部分列から

トッププランを推定する場合において、特に観測される部分列が長い場合やモデルのエントロピーが高い場合に提案法が優れていることを確認した。そのため、部分列の長さやモデルのエントロピーがトッププランの推定に提案法を利用するかどうかの指標となる可能性が示された。

また、本研究では非常に簡単なユーザの行動を考えたが、Web アクセスログの取得方法を工夫することでさらに多くのユーザの行動を取得することが可能であると考えられる。このような拡張や拡張した際に考えられる現実の様々なノイズに対するロバスト性については今後の課題である。

◇ 参 考 文 献 ◇

- [Arlitt 96] Arlitt, M. F. and Williamson, C. L.: Web server workload characterization: The search for invariants, in *ACM SIGMETRICS Performance Evaluation Review*, Vol. 24, pp. 126–137 (1996)
- [Avrahami-Zilberbrand 07] Avrahami-Zilberbrand, D. and Kaminka, G. A.: Incorporating observer biases in keyhole plan recognition (efficiently!), in *Proc. of the 22nd National Conference on Artificial Intelligence (AAAI)*, Vol. 7, pp. 944–949 (2007)
- [Bui 02] Bui, H. H., Venkatesh, S., and West, G.: Policy recognition in the abstract hidden Markov model, *Journal of Artificial Intelligence Research*, Vol. 17, No. 1, pp. 451–499 (2002)
- [Bui 03] Bui, H. H.: A general model for online probabilistic plan recognition, in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 3, pp. 1309–1315 (2003)
- [Charniak 93] Charniak, E. and Goldman, R. P.: A Bayesian model of plan recognition, *Artificial Intelligence*, Vol. 64, No. 1, pp. 53–79 (1993)
- [Chi 98] Chi, Z. and Geman, S.: Estimation of probabilistic context-free grammars, *Computational Linguistics*, Vol. 24, No. 2, pp. 299–305 (1998)
- [Chi 99] Chi, Z.: Statistical properties of probabilistic context-free grammars, *Computational Linguistics*, Vol. 25, No. 1, pp. 131–160 (1999)
- [Fine 98] Fine, S., Singer, Y., and Tishby, N.: The hierarchical hidden Markov model: Analysis and applications, *Machine Learning*, Vol. 32, No. 1, pp. 41–62 (1998)
- [Geib 01] Geib, C. W. and Goldman, R. P.: Plan recognition in intrusion detection systems, in *Proc. of the DARPA Information Survivability Conference and Exposition (DISCEX) 2001*, Vol. 1, pp. 46–55 (2001)
- [Geib 11] Geib, C. W. and Goldman, R. P.: Recognizing plans with loops represented in a lexicalized grammar, in *Proc. of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 958–963 (2011)
- [ITA 01] The Internet Traffic Archive, <http://ita.ee.lbl.gov/> (2001)
- [Jelinek 85] Jelinek, F.: Markov source modeling of text generation, in *The Impact of Processing Techniques on Communications*, pp. 569–591, Springer (1985)
- [Jelinek 91] Jelinek, F. and Lafferty, J. D.: Computation of the probability of initial substring generation by stochastic context-free grammars, *Computational Linguistics*, Vol. 17, No. 3, pp. 315–323 (1991)
- [Kautz 91] Kautz, H. A., Kautz, H. A., Pelavin, R. N., Tenenbergs, J. D., and Kaufmann, M.: A formal theory of plan recognition and its implementation, in *Reasoning About Plans*, pp. 69–125, Morgan Kaufmann (1991)
- [Lau 99] Lau, T. and Horvitz, E.: Patterns of search: analyzing and modeling Web query refinement, *Courses and Lectures - International Centre for Mechanical Sciences*, pp. 119–128 (1999)
- [Lesh 99] Lesh, N., Rich, C., and Sidner, C. L.: Using plan recognition in human-computer collaboration, *Courses and Lectures - International Centre for Mechanical Sciences*, pp. 23–32 (1999)
- [Manning 99] Manning, C. D. and Schütze, H.: *Foundations of statistical natural language processing*, Vol. 999, MIT Press (1999)
- [Pynadath 98] Pynadath, D. V. and Wellman, M. P.: Generalized queries on probabilistic context-free grammars, *Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp. 65–77 (1998)
- [Pynadath 00] Pynadath, D. V. and Wellman, M. P.: Probabilistic state-dependent grammars for plan recognition, in *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 507–514 (2000)
- [Quilici 98] Quilici, A., Yang, Q., and Woods, S.: Applying plan recognition algorithms to program understanding, *Automated Software Engineering*, Vol. 5, No. 3, pp. 347–372 (1998)
- [Sato 01] Sato, T. and Kameya, Y.: Parameter learning of logic programs for symbolic-statistical modeling, *Journal of Artificial Intelligence Research*, Vol. 15, No. 1, pp. 391–454 (2001)
- [Sato 12] Sato, T. and Meyer, P.: Tabling for infinite probability computation., in *The 28th ICLP (Technical Communications)*, pp. 348–358 (2012)
- [Vilain 90] Vilain, M. B.: Getting serious about parsing plans: A grammatical analysis of plan recognition, in *Proc. of the 8th National Conference on Artificial Intelligence (AAAI)*, pp. 190–197 (1990)

〔担当委員：鶴岡 慶雅〕

2013 年 7 月 24 日 受理

著 者 紹 介

小島 諒介

東京工業大学情報工学科卒業，現在，東京工業大学大学院情報理工学系研究科 修士過程在学中，2014 年 3 月同大学院修了予定，2014 年 4 月同大学院博士後期課程進学予定。データマイニングの研究に従事。

佐藤 泰介 (正会員)

1973 年東京工業大学・工・電子物理卒，1975 年同大学院修士課程修了（電気工学専攻），同年通商産業省工業技術院電子技術総合研究所入所。1995 年東京工業大学大学院情報理工学系研究科に移り，現在に至る。工博・人工知能，確率モデリングの研究に従事。日本ソフトウェア科学会会員，情報処理学会会員，AAAI 会員。