アクセスログを用いた
Webサイト訪問者の閲覧行動モデルに関する研究

東京理科大学大学院工学研究科
修士2年
岩渕　隆亮

発表構成

I. 研究背景
II. 研究目的
III. 既存の評価モデル
IV. 提案モデル
V. 事例検証と考察
VI. まとめ，今後の課題
参考文献
Appendix

インターネットの利用実態

- 2006年2月時点において日本のインターネット人口は約7000万人と年々増加している [6]
  - 高速通信設備の普及
- 2002年以降インターネット広告市場が拡大

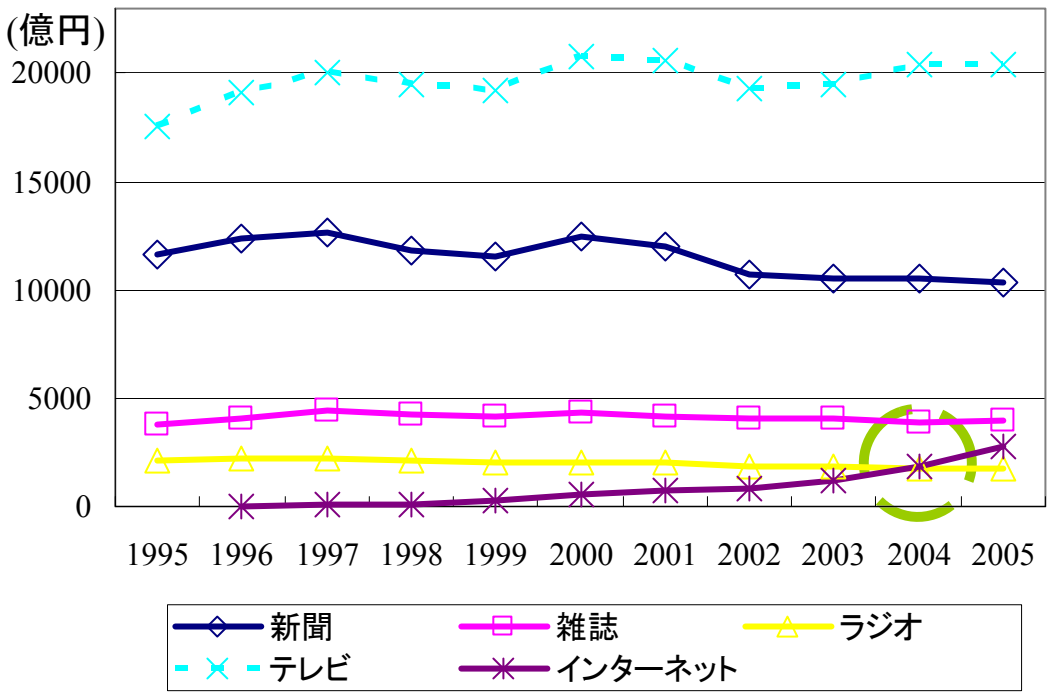
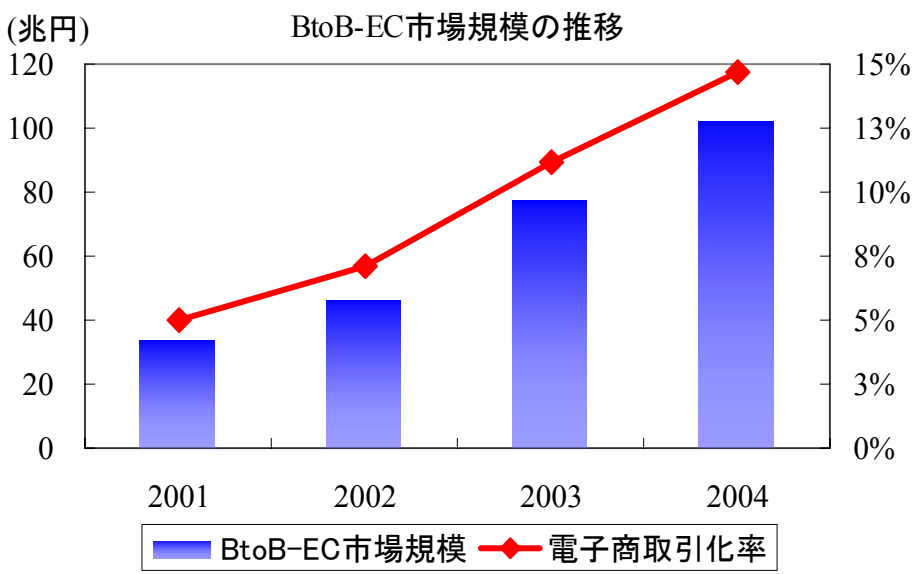


図1.媒体ごとの広告費の推移

- 2004年
  インターネット広告費がラジオ広告費を超える
  （インターネット広告費1814億円，ラジオ広告費1795億円）[7]

EC市場規模の推移

- BtoB-EC(Business to Business Electronic Commerce), BtoC-EC (Business to Consumer Electronic Commerce)の市場規模は年々増加
- 両市場共に電子商取引化率* はこれからも伸び続けると予測されている[5]



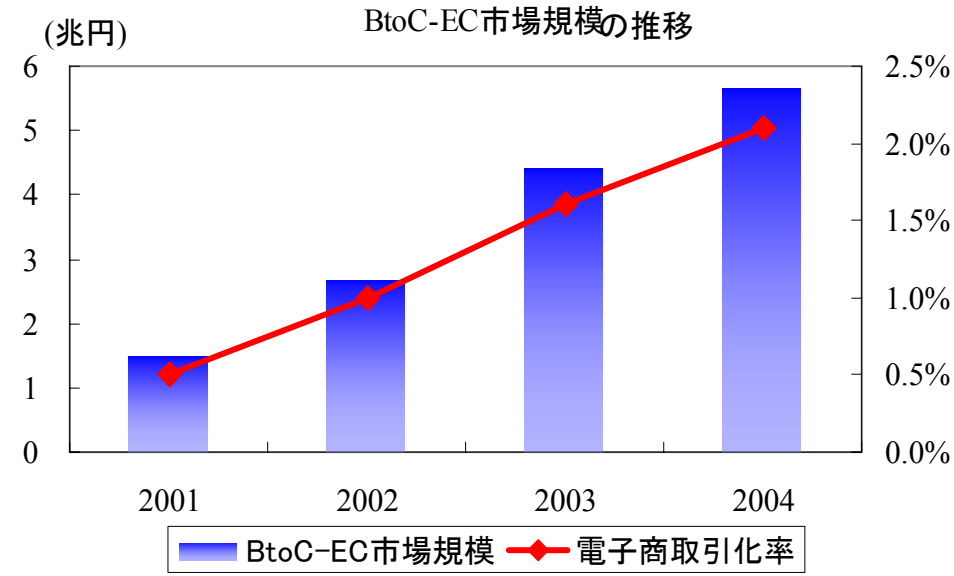


図2.EC市場規模の推移

*電子商取引化率

商品の全販売額・取引額に占めるインターネットを用いた販売額

$$\text{電子商取引化率} = \frac{\text{インターネットを用いた販売額}}{\text{商品ごとの全販売額・取引額}}$$

アクセスログの利用

- インターネットの普及，EC市場の成長によって，ウェブサイトに訪問者を集客するための競争が激化している [7]
- Webサイト運営者が訪問者の閲覧行動を分析する必要性が高まる(経験や勘によるウェブサイト運営では効率が悪い[8])

アクセスログの利用　　**アクセス解析ツールの充実**

➢データに基づいたウェブサイトの評価が可能
➢様々な分析が可能

※ アクセスログ:Webサイト訪問者のアクセス履歴データ
「誰が」「いつ」「どのようにして」「どこから」ウェブサイトに訪れてきたかがわかる

訪問者の購買行動とアクセスログ

- Webサイト訪問者の購買行動とアクセスログの関係

図3.Webサイト内における訪問者の閲覧行動[4]

先行研究

- 近年，アクセスログを用いて，Webサイト内における訪問者の閲覧行動を理解する，ことを目的とした研究が盛んに行なわれている

  ✓ 訪問者の閲覧行動のモデル化に関する研究

  - マルコフモデル(北島らの研究[8],[9](2002年))
    - 訪問者のWebサイト内での閲覧履歴を確率過程とみなし，閲覧行動をモデル化
  - 離散選択モデル(Bucklin[2], Montgomery[3], 里村[4](2005年))
    - Webサイト内での各ページには効用(価値)があるとして，数あるページの中で効用が最大となるページを選択すると仮定したモデル

アクセスログから得られるデータのうち，クリックストリームデータのみを利用した研究がほとんど

研究目的

- ✓ 訪問者の閲覧行動を理解するための情報は
  サーチデータとクリックストリームデータの2つ(図3参照)
- ✓ 先行研究ではクリックストリームデータのみを利用したものがほとんどである

アクセスログを利用した,
ECサイト改善のための分析方法の提案

- 既存の評価モデルとして離散選択モデルを使用
- 離散選択モデルにサーチデータを組み込んだモデルの構築

先行研究1

- マルコフモデルを用いた研究(北村ら[8][9]の研究)

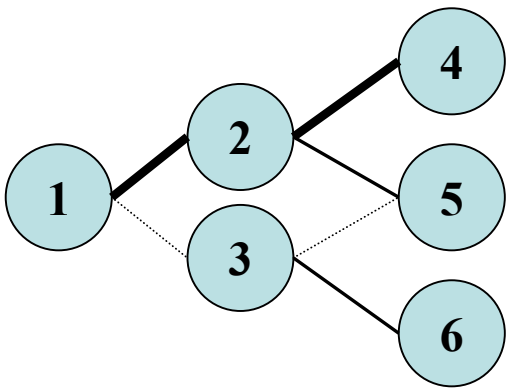


図4.ページ間遷移図

- ✓ 各ページ間での遷移確率(マルコフ過程)を算出し，閲覧行動をモデル化
  - つながりが強いパスはどこか
  - 離脱が多いページはどこかを知ることができる
  - 訪問者別に分析を行なっていない

問題点

どの訪問者特性※が閲覧行動にどう影響しているかが分からない

※訪問者特性：訪問者がWebサイト内を閲覧する際の特性
検索エンジンの利用の有無，過去の訪問回数

先行研究2

- 離散選択モデルを用いた研究(里村[4]の研究)
  - Webサイト内での各ページには効用(価値)があるとして，数あるページの中で効用が最大となるページを選択すると仮定したモデル
  - モデルを用いることにより，各ページ間の遷移確率(回帰式により算出)に影響を及ぼす要因を評価することができる
- 使用データについて
  - ECサイト avance (http://www.avance.jp)のアクセスログデータ
  - データ期間：2004年10月1日～2004年12月31日(3ヶ月)
  - 閲覧ページ数4ページ以上の履歴を分析
  - 分析レベルは各ページの階層

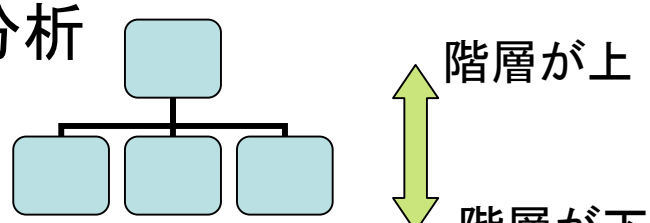


図5.階層の概念図

先行研究2(効用U_{itk}の設定)

- 訪問者の各ページに対する効用を設定

$$U_{itk} = V_{itk} + \varepsilon_{itk}, \varepsilon_{itk} \sim D.E.(\text{二重指数分布}) \quad (1)$$

i : Webサイト訪問者※Appendix$(=1,\ldots,l)$

t : iが次にクリックするページのページ番号$(=1,\ldots,t_i)$

(検索エンジンから入って次のページのページ番号を1とする)

k : 階層(カテゴリ化されたページ群)$(=1,\ldots,K)$

U_{itk} : 訪問者iのページtにおける階層kに対する効用

V_{itk} : 訪問者特性等の説明変数によって確定的に決まる要素

ε_{itk} : V_{itk}以外の様々な要因を含んだランダムな誤差項

先行研究2(選択確率P_{itk}の算出)

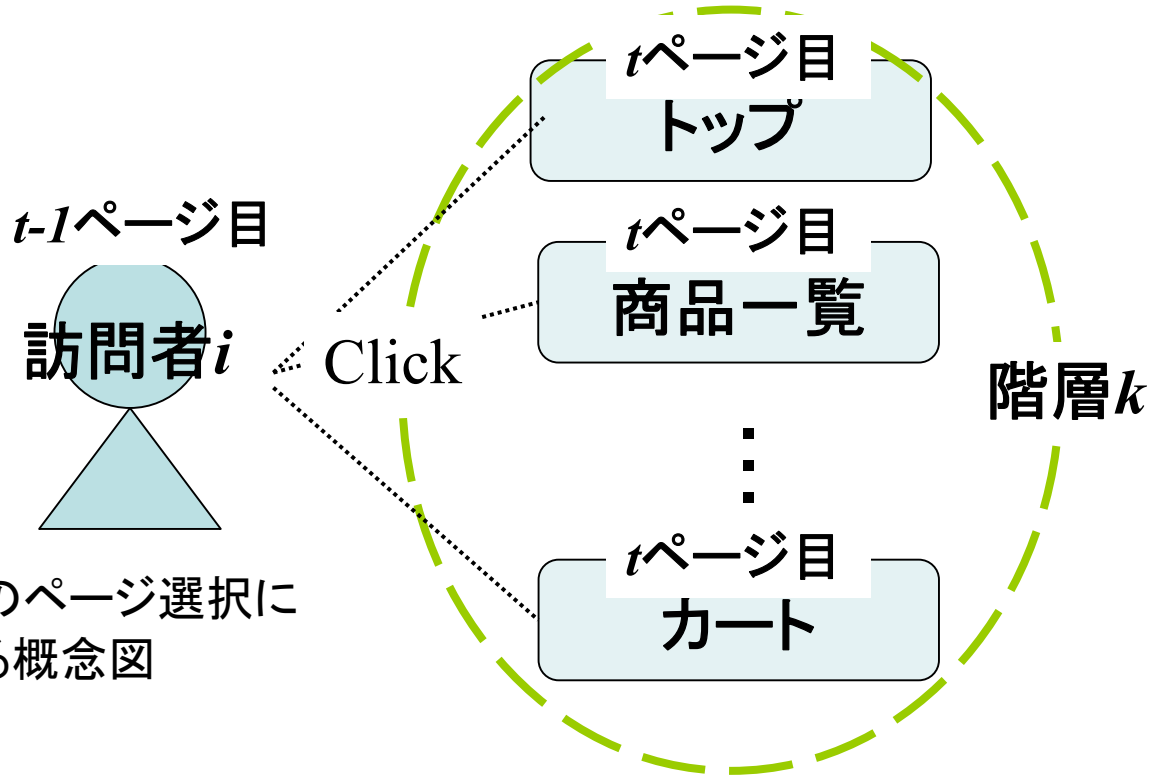


図6.訪問者iのページ選択に関する概念図

- 選択確率の算出(式(1)より導かれる)

$$P_{itk} = \frac{\exp(V_{itk})}{\sum_{k=1}^{K} \exp(V_{itk})} \qquad (2)$$

P_{itk} :個人iが時点tにおいて階層kを選択する確率
V_{itk} :効用の確定的な部分

先行研究2(V_{itk}の設定)

$$V_{itk} = \gamma_{0k} + \gamma_{1k} FV_i + \gamma_{2k} VP_{i,t-1} + \gamma_{3k} SE_i + \eta_1 C^1_{k,i,t-1} + \eta_2 C^2_{k,i,t-1} + \eta_3 C^3_{k,i,t-1} \quad (3)$$

V_{itk} : iがtにおいて選択した階層k

FV_i : iがサイトを2回以上訪問していれば1, それ以外では0

$VP_{i,t-1}$: iが$t-1$までで閲覧したページ数

SE_i : 検索エンジンを利用していれば1, それ以外は0

$C^1_{k,i,t-1}$: iが$t-1$で閲覧した階層がkであれば1, それ以外では0

$C^2_{k,i,t-1}$: iが$t-1$で閲覧した階層がkより上で，かつリンクがあれば1, それ以外では0

$C^3_{k,i,t-1}$: iが$t-1$で閲覧した階層がkより下で，かつリンクがあれば1, それ以外では0

$\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{3k}$: 階層kに関するパラメータ

$\eta_1, \ldots, \eta_3$: リンクに関するパラメータ

既存の評価モデルの問題点と対策

【1】(3)式において，変数*FV,VP,SE*を利用する目的が明確でない

- (3)式の変数の他に候補を挙げ，どのような変数が効くかを調べてみる

【2】時点*t*はページ数の推移を表しているのみで，時間的要素がモデルに組み込まれていない

- 各ページの閲覧時間を何らかの形でモデルに組み込む

【3】使用データはクリックストリームデータのみ

- サーチデータ(検索キーワード)と閲覧行動の関係を明らかにしていく

提案モデル(【1】のみを考慮)

$$V_{itk} = \gamma_{0k} + \gamma_{1k} FV_i + \gamma_{2k} VP_{i,t-1} + \gamma_{3k} SE_i + \gamma_{4k}$$【候補となる変数】

$$+ \eta_1 C^1_{k,i,t-1} + \eta_2 C^2_{k,i,t-1} + \eta_3 C^3_{k,i,t-1} \quad (4)$$

- 候補となる変数

変数として入れたもの

- オンタイム or オフタイム(0/1)
- 各ページごとの閲覧時間
- 広告からWebサイトに入ってきたか否か(0/1)
- Internet Explorer or タブブラウザ(0/1)
- キーワードをカテゴライズ化したもの(0/1)
- 検索フレーズの数

※オンタイム：9時～11時，13時～19時
　オフタイム：12時，19時～8時

分析データ

- アクセスログ：株式会社環のWeb サイト

  (http://www.sibulla.com/)

  - アクセス解析ツール販売のためのWebサイト
  - 閲覧者は，製品情報，製品の導入申し込み，会社概要，サービス等の情報を閲覧することができる
- 期間：2006年1月1日～6月30日(6ヶ月)

表2.アクセスログデータ

項目名	サンプル
年	2005
月	2
日	27
曜日	Sun
時	11
分	44
秒	32
IPアドレス	
ポート	
ホスト名	

項目名	サンプル
ユーザエージェント	Mozilla/4.0 (compatible; MSIE 6.0;
リクエストURL	http://www.nextechcorp.com/
リファラURL	http://search.yahoo.co.jp/bin/search?p=%A5
ユーザID	QiE0ED3T720AAGlH-Ww
セッションID	QiE0ED3T720AAGlH-Ww
UNIX時間	1109472272
ユニークID	QiE0ED3T720AAGlH-Ww
ディスプレイ縦	800
ディスプレイ横	600
色深度	32
訪問回数	1

データの集計結果

表3:データの集計結果

	2006年1月1日～6月30日(6ヶ月間)
訪問者数	18,081
セッション数	22,276
ページビュー数	56,353
1訪問者当たりPV数	3.108
1セッション当たりPV数	2.524

セッション：訪問者がWebサイト内で行う一連の行動をまとめて
1セッションと呼ぶ
同一な訪問者でも次の閲覧までに30分以上間隔があった場合は，
新たなセッションとしてカウント

PV数(ページビュー数)：訪問者がページを閲覧した回数

Webサイトシビラの概要

図7:Webサイトシビラの概観

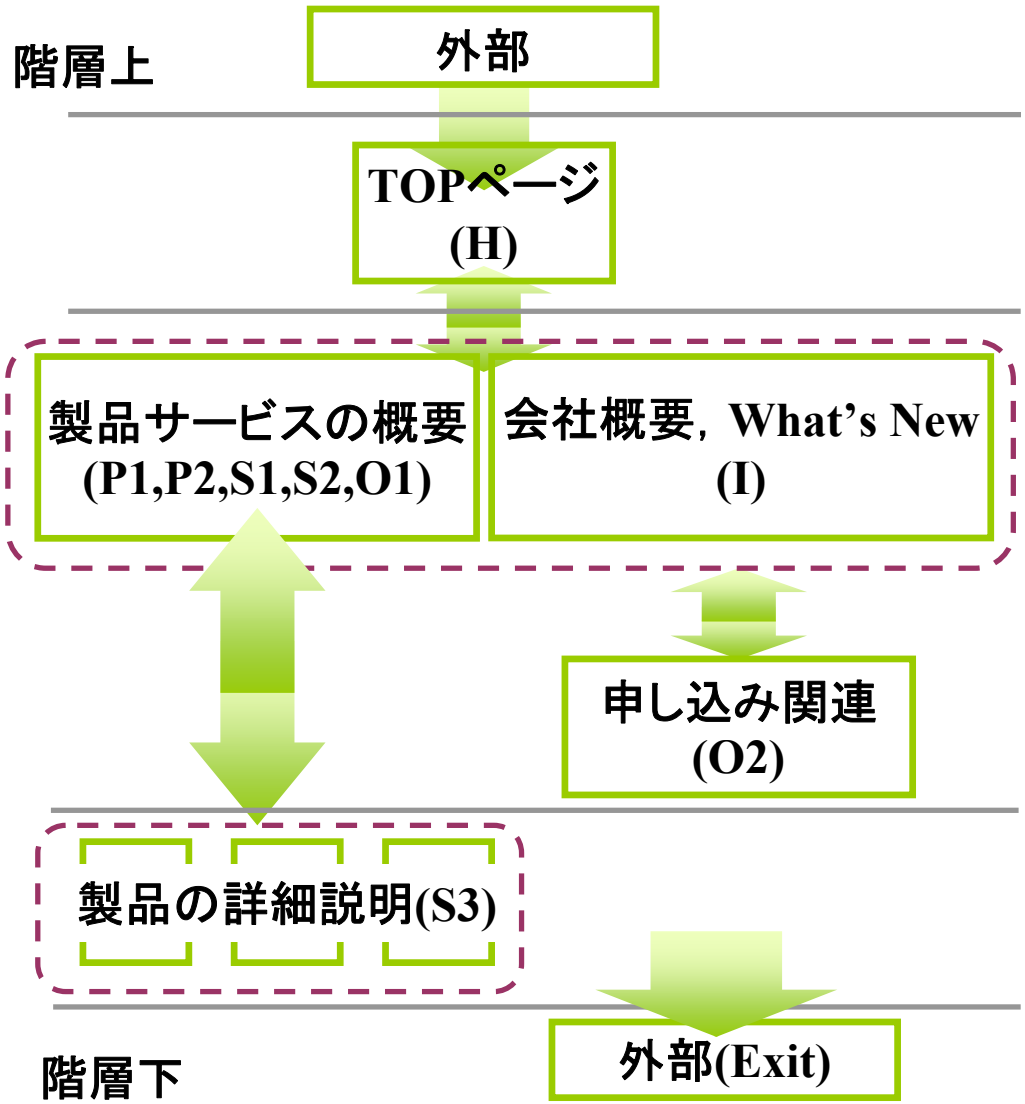


図8:シビラの階層図

階層の設定

表4:階層の設定

対象Webページ	階層名	含まれる情報
トップページ	H(Home)	トップページ
製品の概要	P1(Product1)	製品の特徴
製品の機能	P2(Product2)	機能の紹介
製品の詳細説明	P3(Product3)	製品の詳細な説明
製品申し込みに関する情報	O1(Order1)	料金表，申し込みページ
申し込み完了	O2(Order2)	申し込み確認ページ（製品，サービス含む）
オプションサービス	S1(Service1)	解析レポート，コンサルティング
その他のサービス	S2(Service2)	サポート，メールマガジン，よくあるご質問，提携サービス
その他の情報	I(information)	会社概要，プライバシーポリシー，What'sNew，プレスリリース
外部のページ	Exit	他のWebサイト，検索エンジン，お気に入り，

- ナビゲーションで区画分けされている項目を参考に階層の設定を行なった

※ナビゲーション：ページ間の移動を容易にするために設けられた，Webサイト内のリンク

既存モデルと提案モデル

- シビラのアクセスログを用いて，既存の評価モデルと提案する評価モデル(【1】のみを考慮)の比較を行なう
  - 既存の評価モデル

$$V_{itk} = \gamma_{0k} + \gamma_{1k} FV_i + \gamma_{2k} VP_{i,t-1} + \gamma_{3k} SE_i + \eta_1 C^1_{k,i,t-1} + \eta_2 C^2_{k,i,t-1} + \eta_3 C^3_{k,i,t-1} \quad (3)$$

  - 提案する評価モデル

$$V_{itk} = \gamma_{0k} + \gamma_{1k} FV_i + \gamma_{2k} VP_{i,t-1} + \gamma_{3k} SE_i + \gamma_{4k} ON_i + \gamma_{5k} DU_{i,t-1} + \eta_1 C^1_{k,i,t-1} + \eta_2 C^2_{k,i,t-1} + \eta_3 C^3_{k,i,t-1} \quad (4)$$

ON_i:iがt-1時点で閲覧した時間帯がオンタイムorオフタイム(0/1)
$DU_{i,t-1}$:iがt-1時点で閲覧したページの閲覧時間

両モデルによる推定結果1

表5:既存モデルによる推定結果

	Home	Product1	Product2	Product3	Order1	Order2	Service1	Service2	Infom.	Exit
γ0(Intercept)	1.35(**)	0.34(**)	1.43(**)	0.21(**)	0.39(**)	-0.5(**)	-0.93(**)	-1.11(**)	-0.84(**)	0
γ1(FV)	0.00	-0.02	-0.09(**)	-0.22(**)	-0.04	-0.23(**)	0.05	-0.27(**)	-0.06	0
γ2(VP)	-0.16(**)	-0.05(**)	-0.02(**)	0.01(**)	-0.02(**)	-0.02(**)	0.00	0.02(**)	-0.02(**)	0
γ3(SE)	-0.02	0.01	0.02	-0.01	0.02	0.23(**)	-0.01	0.13(**)	0.00	0

＊＊1%水準有意，＊5%水準有意

表6:提案モデルによる推定結果

	Home	Product1	Product2	Product3	Order1	Order2	Service1	Service2	Infom.	Exit
γ0(Intercept)	1.30(**)	0.28(**)	1.39(**)	0.18(**)	0.35(**)	-0.60(**)	-0.96(**)	-1.14(**)	-0.87(**)	0
γ1(FV)	0.02	-0.01	-0.08(**)	-0.21(**)	-0.03	-0.21(**)	0.06	-0.27(**)	-0.05	0
γ2(VP)	-0.16(**)	-0.05(**)	-0.02(**)	0.01(**)	-0.02(**)	-0.02(**)	0.00	0.02(**)	-0.02(**)	0
γ3(SE)	-0.02	0.01	0.02	-0.01	0.03	0.23(**)	-0.01	0.13(**)	0.00	0
γ4(ON)	-0.10(**)	-0.11(**)	-0.07(**)	-0.05	-0.07(**)	-0.19(**)	-0.08(**)	-0.06	-0.07	0

＊＊1%水準有意，＊5%水準有意

- モデルの当てはまりは提案モデルの方が多少優れている
  - モデル比較の指標は尤度比を使用
- 階層に関するパラメータ($C1, C2, C3$)の値は得られなかった
  - 確率P_{itk}の算出を行なうことができなかった
  - 階層設定(表4参照)の見直しが必要であると考えられる

両モデルによる推定結果2

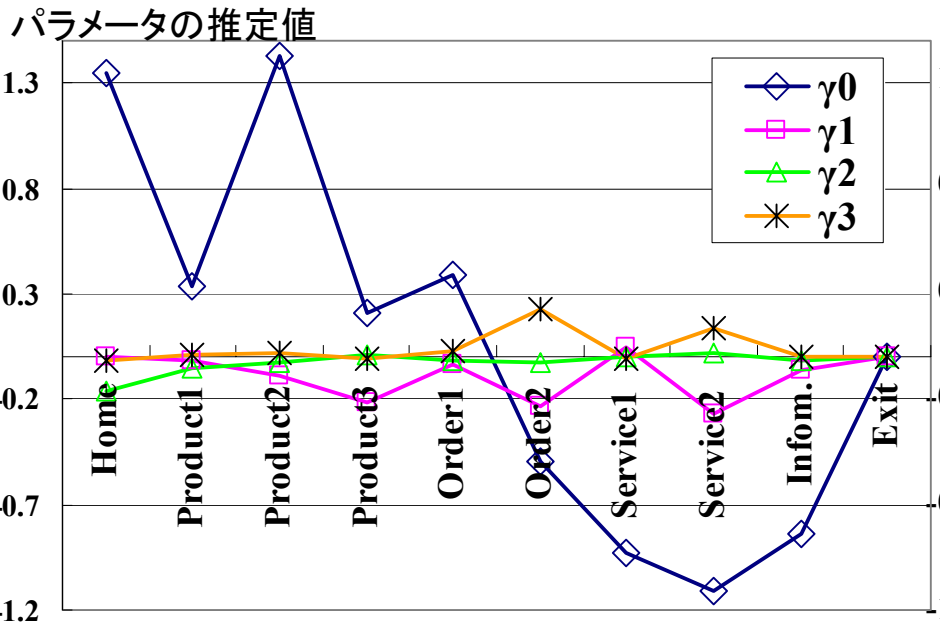


図9:推定結果(既存モデル)

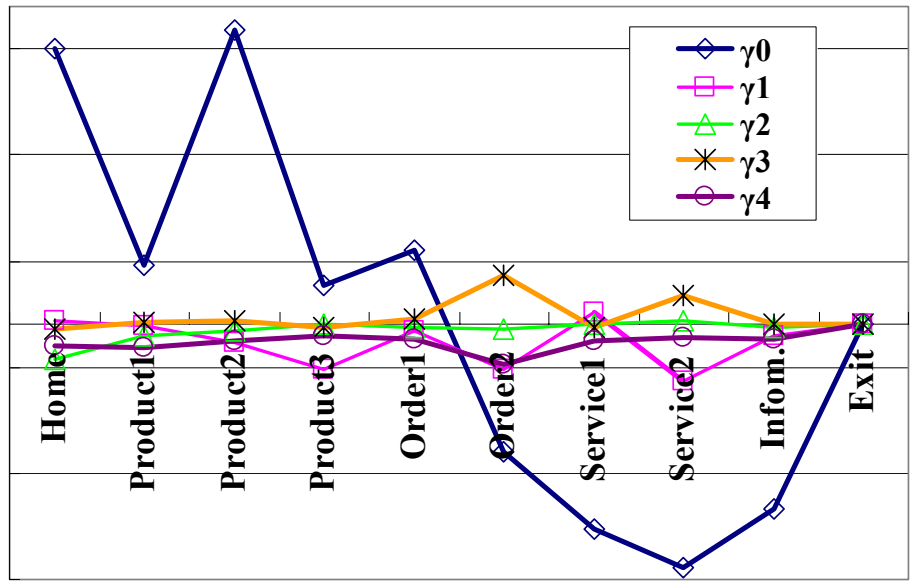


図10:推定結果(提案モデル)

- *Exit*を0(基準値)として相対的なパラメータの値をプロット
- 両モデルにおいてパラメータの値に大きな差はない

結果と考察

図9，図10より

	既存モデル	提案モデル	考察
γ0	階層Service1,2とInformにおいて大きく負の値		その他サービスに関するページは閲覧されにくい
γ1 (FV)	ほぼ全ての階層において負の値		リピーターほどサイト内を閲覧せず離脱する傾向が強い
γ2 (VP)	顕著な特徴は見られない		前のページまでの閲覧ページ数は次のページに進む確率にあまり関係がない
γ3 (SE)	階層Order2とService2において正の値		検索エンジンから来た訪問者ほど，申込に結びつきやすい
γ4 (ON)		全て負の値	オフタイムは閲覧されにくい(BtoB-ECの特徴)

まとめ，今後の課題

まとめ

- ECサイト改善のための手法として，離散選択モデルを用いて分析を行なった
  - 変数に関する問題点を指摘
  - モデルの提案を行なった

今後の課題

- 階層構造の見直し
  - 階層に関するパラメータの推定を可能にする
- 問題点【2】，【3】(P14参照)を考慮したモデルの構築
- サーチデータと閲覧行動の関係を明らかにする
  - テキストマイニング手法に関する文献を読み進める

主要参考文献

[1] 江尻俊章著，“稼ぐホームページ 損なホームページ”，株式会社アスキー (2004)

[2] Radolph E.Bucklin and Catarina Sismeiro(2003),”A Model of Web Site Browsing Behavior Estimated on Clickstream Data”,Journal of Marketing Research,Vol.XL(August),249-267

[3] Alan L.Montgomery,Shibo Li,Kannan Srinvasan, JohnC.Liechty (2004), “Modeling Online Browsing and Path Analysis Using Clickstream Data”, Marketing Science,Vol.23(Fall),579-595

[4] 里村卓也，“ECサイトでのサーチ＆クリックストリームデータのモデル分析”日本マーケティングサイエンス学会第77回研究大会，(2005)

[5] 経済産業省・ECOM・NTT データ経営研究所 共同，“平成**16** 年度電子商取引に関する実態・市場規模調査”，次世代電子商取引推進協議会 (2005)

[6] Internet Watch, “http://internet.watch.impress.co.jp/cda/special/2006/06/19/12386.html”, (最終閲覧日2006年11月7日)

[7] 株式会社 電通 ニュースリリース, “http://www.dentsu.co.jp/marketing/adex/adex2005/_media.html”, (最終閲覧日2006年11月7日)

主要参考文献

[8] 狩谷典之，北島宗雄，高木英明，張勇兵(2002)“Markovモデルを用いたe-コマースサイトのwebデザイン評価” 電子情報通信学会論文誌B，J85-B，10，1809-1812.

[9] 山本哲生，北島宗雄，高木英明，張勇兵(2002) “Markov連鎖を用いたウェッブナビゲーション過程の評価” 情報通信ネットワークの新しい性能評価法に関する総合的研究，189-198.

Appendix

Visual Mining Studioによる分析

- Visual Mining Studio(VMS)を用いた部分
  - 離散選択モデル(3), (4)式を推定するための<u>データセット</u>を作る際にVMSを使用した
- データセットのイメージ
  - アクセスログデータから表Aのようなデータセットを作成した

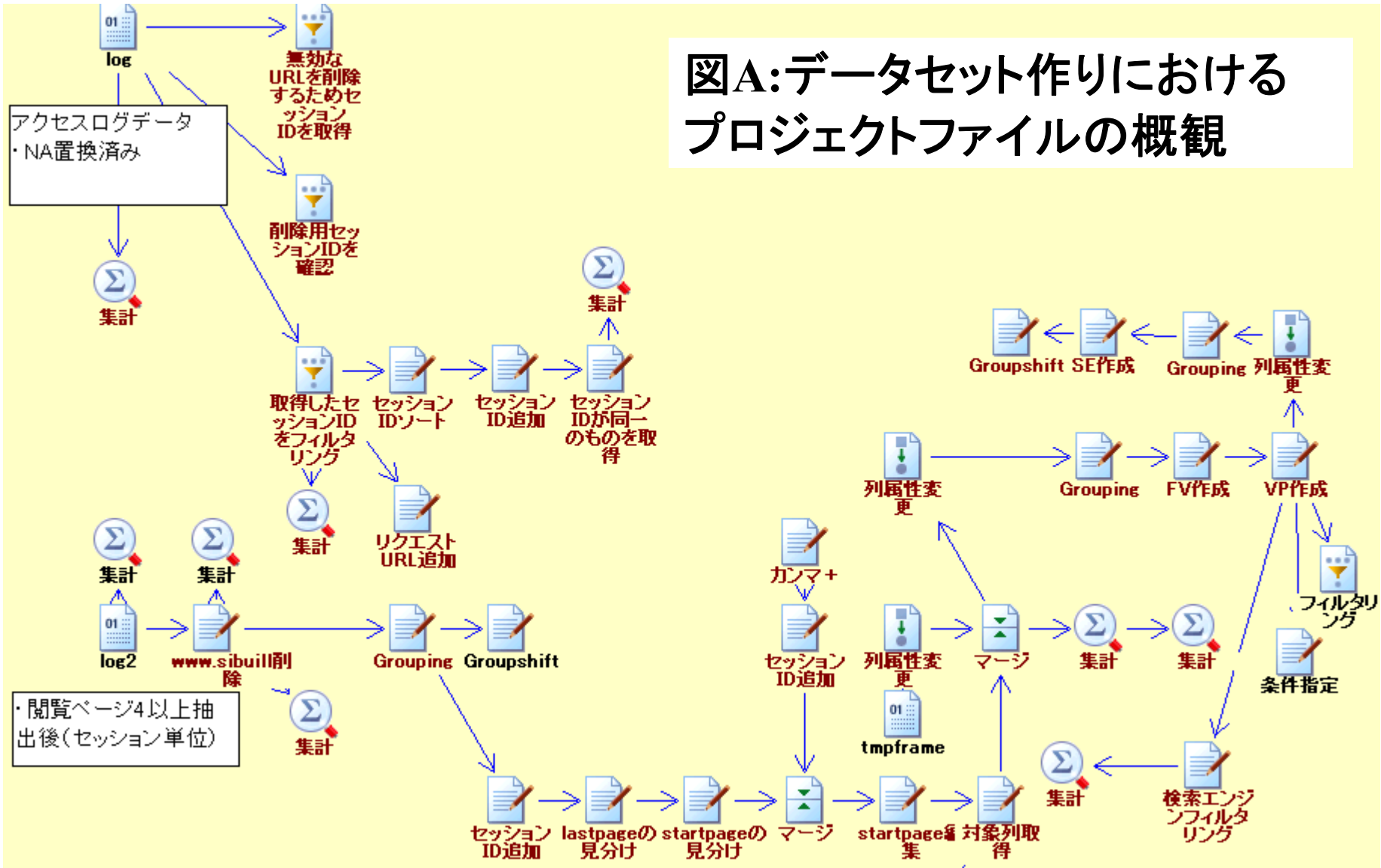
アクセスログデータ

表A:データセットのイメージ

i	*t*	*k*	*FV*	*VP*	*SE*	*ON*	*C1*	*C2*	*C3*
1	1	1	1	1	1	1	0	1	0
1	2	5	1	2	1	1	0	1	0
1	3	3	1	3	1	1	0	0	0
1	4	5	1	4	1	1	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

※各変数の説明はP13,20を参照

Visual Mining Studioによる分析



図A:データセット作りにおける
プロジェクトファイルの概観

Visual Mining Studioによる分析

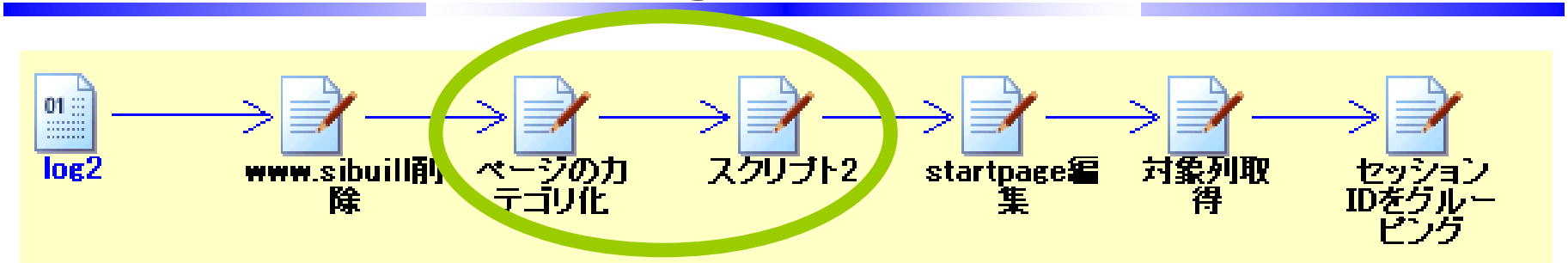
- データセット作りにおいて用いたVMSの機能を紹介する
  - 全てのデータセットを説明するのは困難なので，一部を紹介することとする

1. データの読み込み
2. リクエストURLにおけるhttp://www.sibuilla.com/の文字列を削除
   ① Substringを用いて部分文字列の抽出を行なった

   a(“リクエストURL2”)=substring(logdata(“リクエストURL”),24,50);

   24文字以降の文字列を抽出

Visual Mining Studioによる分析



3. ページのカテゴリ化

- 2.で得られた部分文字列(ex. Path.html)を用いて，各ページの階層分けを行なった(階層の分け方は表4を参照)

t1("Group") = grouping(logdata("リクエストURL2"),

{“top.html”,“company.html”,・・・,”path”},{“H”,”I”,・・・,”P1”});

4. スクリプト2

- 入り口ページ，離脱ページを判別し，それぞれフラグ”1”をたてる
- スクリプト2の内容は次ページに示す

Visual Mining Studioによる分析

～スクリプト1の内容～

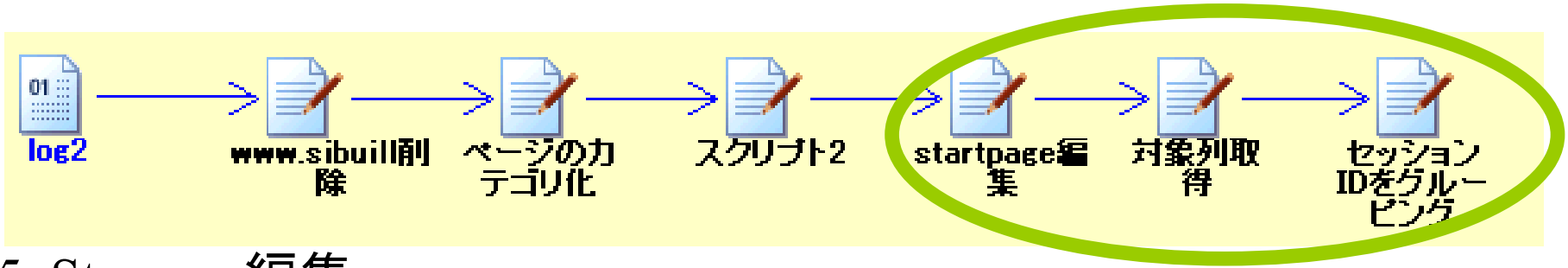
目的：入り口ページ，離脱ページの判別

```
//【行移動】セッション2=セッションIDを全行上へ1移動（-1=1行上へ）
t1("セッション1") = logdata("セッションID");
t2("セッション2") = row_shift(logdata("セッションID"),-1);
result = cbind(logdata,t1,t2);

//【条件抽出】離脱ページの見分け，離脱ページにフラグ"1"をたてる
t1("last") = sel(logdata("セッション1")==logdata("セッション2"),0,1);
result = cbind(logdata,t1);

//【条件抽出】入り口ページの見分け，入り口ページにフラグ"1"をたてる
t2("start") = row_shift(logdata("last"),1);
result = cbind(logdata,t2);
```

Visual Mining Studioによる分析



5. Starpage編集
   - 手順4.においてshift機能を使っている為，1行目のセッションにおけるstartpageのフラグは”0”になっている.そこでstartpageのフラグを”1”に変換した

6. 対象列取得
   - 今後分析に必要な列のみを抽出

result = logdata("リクエストURL","リファラURL","セッションID","UNIX時間","訪問回数","広告情報","検索エンジン","Group","last","startpage");

7. セッションIDをグルーピング(*i*の作成)
   - セッションIDを”1”,”2”のように数値に変換し*i*(=1,2,・・・,*I*)を作成

t1("Group") = grouping(logdata("セッションID"),
{"000A〇389F6","000〇2ED1",・・・,"000〇3YD1"},{"1","2",・・・,"3411"});

Visual Mining Studioによる分析

- *t*の作成
  - セッション単位で閲覧したページ数を1,2,3・・・とする
  - その際，手順4で作成した，入り口ページのフラグを使用する

入り口ページ：1,0,0,・・・,0,1,0,0,1,0,0,0,0,0,0,0,0,1

1セッション　1セッション　1セッション

閲覧ページ数*t*：1,2,3,・・・,7,1,2,3,1,2,3,4,5,6,7,8,9,1

1セッション　1セッション　1セッション

  - 上記のような変換をほどこすためにスクリプトを作成した(次ページ参照)

Visual Mining Studioによる分析

- *t*を作成するためのスクリプト

```
x = as.integer(ファイル名[,対象列])
y = seq(along=x)
y[x==1] = 0
num = which(y==0)
for(i in seq(along=num[-1]) ){
    print(seq(from=num[i]+1,to=num[i+1]))
    if(num[i]+1 != num[i+1]){
      y[seq(from=num[i]+1,to=num[i+1])] =
    y[seq(from=num[i]+1,to=num[i+1])] - num[i]
    }
}
y[seq(from=num[length(num)],to=length(y))] =
    y[seq(from=num[length(num)],to=length(y))] - num[length(num)]
y[y<0] = 0
y = y + 1
tmpframe = data.frame(x,y)
```

付録.リクエストURLの分解手順

- リクエストURLの分解手順（例：http://www.sibulla.com/site/index.html）
- http://www.sibulla.com/を除去
  - 全リクエストURLに共通なので，先頭23文字を除去という文字列関数を用いる
  - リクエストURLの文字列の長さを数える
    tmp1(“フィルタ”) = strlen(log(“リクエストURL”))
  - リクエストURLの24文字目から最後の文字までを取り出す
    tmp2(“リクエストURL2”) = substring(log(“リクエストURL”),24,tmp1(“フィルタ”))
  - 除去後はsite/index.htmlとなる
- site/index.htmlから製品の機能を表すpageという言葉だけを残す
  - /(スラッシュ)を元に，二つの語に分解する文字列関数を用いる
  - 上で取り出した文字列を/を境に二つの文字列に分解する
    tmp3(“リクエストURL3”,“リクエストURL4”) = split.str(tmp2(“リクエストURL2”),“/”,“”)
  - 元データに列を付け加える
    b = cbind(log,tmp3)