携帯電話用ホームページ作成のための数理的基礎*

3M−3

小泉 圭右†
慶應義塾大学大学院 理工学研究科‡

神保 雅一§
慶應義塾大学 理工学部¶

寒河江 雅彦‖
岐阜大学 工学部**

1　はじめに

携帯電話で閲覧可能なホームページは，パソコンを用いて閲覧する一般のホームページと比べて表示画面サイズに制限があるため，メニューページとコンテンツページは分割され，コンテンツページはページ間の関係を表す木構造の端点に置かれることが多い．また，メニューページは内点に対応するが，内点の枝数には表示画面サイズによる制限がある．さらに，任意のコンテンツページ間の推移確率が与えられているとする．本論文の目的は，このような制約のもとで，最適な木構造を見出すことである．

2　携帯電話用ホームページに関する仮定

携帯電話用ホームページに関する仮定として，以下の項目を定めることにする．

- $S = \{s_0, s_1, \cdots, s_n\}$，ただし s_0 はトップページ，$s_1, \cdots, s_n$ はコンテンツページ．
- トップページ s_0 は，組織名などの他にはリンクはなく，唯一のリンク先から root ページに入る．
- 掲示したい各情報 $s_0, \cdots, s_n$ は 1 画面に表示できる量であり，同じホームページ内へのリンクは存在しない．
- 各コンテンツページ $s_0, \cdots, s_n$ は木構造の leaf に置かれており，root ページを含めた各内点は高々 k 個の選択肢が書かれているメニューページである．
- 観測できる量として，サーバのログからユーザ名（あるいはマシーン名），アクセスページ，アクセス時刻が挙げられる．
- ユーザはブラウザ上で現在のページの 1 つ前のページに戻ることができる．
- 任意の木構造に則したメニュー画面を作ることが可能であると仮定する．

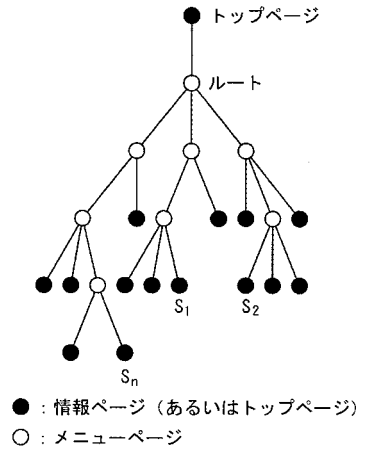

図 1: ホームページの木構造の例

3　ユーザの遷移と目的関数

- $S^{(l)}$ を長さ l の S の要素の並びとする．
- $S^* = \bigcup_{l=1}^{\infty} S^{(l)}$
- 時刻は離散的と仮定し，$p_s (s \in S)$ をある時刻にページ s がアクセスされる定常確率とする．
- $p_{s_j|s_i} (s_i, s_j \in S)$ をある時刻にページ s_i にいるとし，次の時刻にページ s_j に移動する推移確率とする．
- $\boldsymbol{s} \in S^*$ に対して $l(\boldsymbol{s})$ を列 $\boldsymbol{s}$ の長さとし，$\boldsymbol{s} = (s_1, \cdots, s_l)$ に対して，

  $$P(\boldsymbol{s}) = p_{s_1} p_{s_2|s_1} p_{s_3|s_2} \cdots p_{s_{l-1}|s_l} p^{l(\boldsymbol{s})-1}(1-p)$$

  とする．ただし，p はホームページシステムから退去しない確率．
- $\mathcal{T}_{n,k}$ を n 個の leaf を持つ最大次数 k の木の集合とし，$T \in \mathcal{T}_{n+1,k+1}$ を任意に固定して，$n+1$ 個の leaf に S の各ページを対応させる．
- $d(s_i, s_j; T)$ を木 T における s_i と s_j の間のパス長とすると，

  $$d(\boldsymbol{s}; T) = \sum_{i=1}^{l-1} d(s_i, s_{i+1}; T)$$

---

*Mathematical Fundamentals of Homepage Construction for Mobile Phones

†Keisuke Koizumi
‡Graduate School of Science and Technology, Keio University
§Masakazu Jimbo
¶Faculty of Science and Technology, Keio University
‖Masahiko Sagae
**Faculty of Engineering, Gifu University

となる．ただし $\boldsymbol{s} = (s_1, s_2, \cdots, s_l)$．

- $D(T)$ を平均パス長とすると，目的関数は，

$$\begin{aligned} D(T) &= E\left(\frac{d(\boldsymbol{s};T)}{l(\boldsymbol{s})-1}\right) \\ &= \sum_{\boldsymbol{s}} \frac{d(\boldsymbol{s};T)}{l(\boldsymbol{s})-1} P(\boldsymbol{s}) \\ &= \sum_{s_i \in S} \sum_{s_j \in S} d(s_i, s_j; T) p_{s_i s_j} \longrightarrow min \end{aligned}$$

となる．ただし，$p_{s_i s_j} = p_{s_i} p_{s_j|s_i}$．

4 Huffman 符号と最適木に関する定理

情報源記号（本論文では $S = \{s_0, s_1, \cdots, s_n\}$）とその生起確率が与えられた時に，平均符号長が最小になる Huffman 符号 [1] がある．簡単のため，2 元符号の場合における Huffman 符号の作り方を以下に述べる．

(1) まず，M 個の情報源記号 $s_1, \cdots, s_m$ を生起確率の大きい順に並べる．

(2) 生起確率が最小の情報源記号 2 個をまとめ，これを 1 つの情報源記号に置き換え，その合成確率（2 つの確率の和）を新しい情報源記号の生起確率とし，再び確率の大きいものから順に並べ直す．

(3) (2) の手続きを最後の確率 1 の記号が来るまで繰り返す．

Huffman 符号は必ずしも一意には定まらない．例えば，中間段階において生起確率の等しい情報源記号が複数個生じた場合には，その記号の並べ方に任意性が生じ，異なる符合の木が得られるが，いずれも平均符号長は等しく，その平均符号長は最短である．$D(T)$ を最小にする最適木は Huffman 符号を用いて下記のように構成される．

定理 1 ページ s_i の定常生起確率を p_{s_i} とし，$\tilde{p}_{s_i} = p_{s_i} - p_{s_j s_i}$ とおく．$\tilde{p}_{s_0}, \cdots, \tilde{p}_{s_n}$ を大きい順に並べ替えたものを $\tilde{p}^{(0)}, \tilde{p}^{(1)}, \cdots, \tilde{p}^{(n)}$ とする．leaf の数が n，次数が k 以下の全ての木の集合を $\mathcal{T}_{n,k}$ と書く．$\tilde{p}^{(1)}, \cdots, \tilde{p}^{(n)}$ を用いて Huffman k 分木を構成し，その根に $\tilde{p}^{(0)}$ を leaf として付加した木 T^H_{n+1} は，

$$D(T^H_{n+1}) \leq D(T) \qquad \forall T \in \mathcal{T}_{n+1,k+1}$$

を満たす．よって T^H_{n+1} は平均符号長が最短の木である．

5 $\tilde{p}_{s_i}$ の推定

ページ s_i にアクセスしてから次にページ s_j にアクセスする回数を $n_{s_i s_j}$ とすると，$n_{s_i s_j}$ はアクセスログのデータから得ることができる．

	s_0	s_1	s_2	$\cdots$	s_n	計
s_0	○	n_{01}	n_{02}	$\cdots$	n_{0n}	n_{0+}
s_1	n_{10}	○	n_{12}	$\cdots$	n_{1n}	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
s_n	n_{n0}	n_{n1}	n_{n2}	$\cdots$	○	n_{n+}
	n_{+0}	n_{+1}	n_{+2}	$\cdots$	n_{+n}	N

（○はアクセスログからは不明）

よって，$p_{s_i s_j}$ の最尤推定値は，

$$\hat{p}_{s_i s_j} = \frac{n_{s_i s_j}}{N}$$

となる．よって，$\tilde{p}_{s_i}$ の最尤推定値は，

$$\hat{\tilde{p}}_{s_i} = \sum_{s_j \neq s_i} = \frac{n_{s_i +}}{N}$$

となる．ただし，$n_{s_i +} = \sum_{s_j \neq s_i} n_{s_i s_j}$．

6 結論と今後の課題

直観的には $\tilde{p}_{s_i}$ だけでなく，ページ s_i からページ s_j への推移確率により最適な木構造が決まるように思われるが，実際には $\tilde{p}_{s_i}$ だけの情報から最適木構造が決まるため，最適木を求める複雑なアルゴリズムが不要であり，$p_{s_i s_j}$ を推定することなくアクセスログの情報だけで最適木構造が決定できることを明らかにした．また，定理 1 により木構造が決まるが，トップページが対応付けられた leaf が木の一番上に来るように木を書き直せば，ホームページの木構造が見やすくなる．また，今後の課題として以下のことに取り組みたい．

- ページ s_i からすぐにページ s_j に移動したのか，それとも s_i から外部のページ経由で s_j に移動したのかがアクセスログからでは区別がつかない．そのため，2 つのページにアクセスした時間間隔も考慮に入れた $\tilde{p}_{s_i}$ の推定が必要．
- 携帯電話のメモリー内に，直前にアクセスした 1, 2 ページが記憶されるため，この点を考慮した $\tilde{p}_{s_i}$ の推定が必要．
- 今後の課題として，メニューの中に“○○に戻る”などの戻りや，リンクを加えた場合の最適グラフ構造を考え，その際どのようなグラフが最適なのかを調べてみたい．

参考文献

[1] 宮川 洋，原島 博，今井 秀樹．岩波講座 情報科学 - 4 情報と符号の理論．岩波書店，1982 年．