WWWサーバアクセス履歴からのユーザモデルの構築*

4Aa−3

三浦 信幸†　島 健一‡

©NTT ソフトウェア研究所
E-mail: {miura, kshima}@slab.ntt.jp

1 はじめに

従来，WWWサーバの履歴の解析[1]は，各ページへのアクセス回数をカウントする等の，アクセスしたユーザの全体的な動向を分析するために行われてきた．しかし，ユーザが使いやすい，再び使いたいと思うような，より高品質なサーバを構築するためには，個々のユーザの性質を把握し，サーバ構築にフィードバックする必要があると考えられる．本研究では，WWWサーバ履歴から，個々のユーザのアクセス履歴を抽出するために，WWWサーバの拡張および個人履歴抽出ツールの作成を行い，抽出結果から，個々のユーザに対するユーザモデルを構築する方法について考察を行った．ユーザモデルは，WWWサーバ上の各ページの内容および構成・ページ間のリンクを各ユーザごとに動的に変更すること等に応用できると考えている．

2 WWWサーバの拡張

WWWサーバには種々のものがある[1]が，ここでは，NCSAのhttpd[2]をとりあげる．このサーバはデフォルトでは，表1のような，access_log，referer_log，agent_logという3種類のlog fileを生成する．ただし，referer_logは，browserによっては移動元の情報を通知してこない[3]ため，記録されない場合がある．

このようなlog fileでは，サーバにアクセスした人全員のlogが一つのファイルに収められている．また，従来のlog解析ツール[1]のほとんどは，access_logのみを対象としており，その場合，パスワードによるアクセス制限をしていないファイルが参照された際には，access_logの第3フィールドであるuser名は記録されず，host名だけでユーザを特定しなくてはならない．特に，一般的になっているproxy server経由のアクセスの場合，host名はすべてproxy serverのhost名となり，同一proxy server経由の複数のユーザの履歴を識別することが非常に困難である[2]．

このような問題点を解決するには，これらのlog fileのうち，access_logとreferer_logの2つとサーバ上にあるHTML file等のリンク状況とを統合的に解析する必要がある．そのためには，各log fileの各行どうしの対応づけが必要だが，表1の例でわかるように，referer_logには，host名，user名，日付，時間などが記録されていないため，access_logと対応づけることができない．そこで，まず，本研究では，従来の3種類のlog fileとは別に，両者の情報を合わせた表2のようなlog fileを生成するようにWWW Serverを拡張した．source fileへの追加は，57行で，全体の0.3%に相当する．

なお，Ver.1.5以降のNCSA httpdでは，log fileの生成の仕方を選択することができ，本拡張のような出力形式で3つのlog fileを一つのfileに出力するようにもできるが，この場合，従来の解析ツールが正常に動作しなくなる場合がある[4]．今回の拡張では，従来の3つのlog fileは従来通り出力するため，他の解析ツールとの併用に関し，何ら支障がない．

3 個人履歴抽出ツールの作成

拡張サーバのlog fileから，個々のユーザのアクセス履歴を抽出するには，次のように行えば良い．

1. 第1field(host名)ごとに，log fileを分割する．
2. サーバ上のfileのリンク関係を読み込む．各リンクには，方向別に重みづけを行う．通常良くたどられる方向のリンクに大きな重みをつける．
3. host名ごとに分割されたlog fileを個々に，次の優先順位に従って，各ユーザ毎に分割する．
   (a) 第3field(user名)がある場合，そのユーザの履歴とする．
   (b) 第8field(リンク移動元)がある場合，一定時間まで遡って，過去にその fileを参照したユーザがいれば，そのユーザの履歴とする．複数候補がある場合には，時間的にもっとも近い参照履歴を採用する．
   (c) 読み込んだリンク関係から，リンク移動元の候補を探し，それぞれの候補に対し，前段同様に，一定時間遡ってユーザを特定する．複数候補がある場合，時間優先・リンク間の重み優先で特定する．
   (d) 上記のいずれでも，ユーザが特定できない場合，あるユーザによる新たなアクセスとして扱う．
4. 同一ユーザで複数のhostからのアクセスがある場合，それらを一つのfileにマージする．

以上の抽出方法にしたがって，個人履歴抽出ツールをPerlを用いて実装した．

本ツールの出力形式は，NCSA httpdのdefaultのaccess_logと同一であるため，本出力結果に対し，従来のlog解析ツール[1]を適用すれば，個人ごとの各pageへのアクセス回数などを計算することができる．

4 個人履歴抽出ツールの適用例

今回は，本ツールをJapan Telephone Directory[5][3]に適用してみた．適用イメージは図1の上半分のようになる．

適用結果を利用して，電話番号の検索を行ったことのある2165人のユーザのうち，利用回数の多い上位10人について，電話番号の検索方法の嗜好について分析してみたところ，表3のようになった．例えば，Aさんは，電話番号の検索のうち，名称別の検索を行ったのは24%，業種別の検索は69%，地域別の検索は0%，地図上からの検索は7%であった．

全体の傾向では名称別検索と業種別検索が大半との結果だが，Bさんのように業種別検索ばかりのユーザやGさんのように名称別検索の多いユーザ，Hさんや Iさんのように地図上からの検索が多いユーザなど

---

* Composing the User Model by Analyzing the Access Logs of the WWW Server

† Nobuyuki Miura, NTT Software Laboratories

‡ Ken'ichi Shima, NTT Software Laboratories

[1] http://www.yahoo.com/Computers/World_Wide_Web/HTTP/Servers/.

[2] http://hoohoo.ncsa.uiuc.edu/

[3] 例えば，NCSA Mosaic.

[4] 例えば，wwwstat. http://www.ics.uci.edu/WebSoft/wwwstat/

[5] 弊社で発行している英語版電話帳である City Sourceをもとに，インターネット版マルチメディア電話帳として構築したもの．各種技術の研究用として，2年間に限り，実験的に公開する．http://www.pearnet.org/jtd/

表 1: NCSA httpd が生成するアクセス・ログの例

```
・access_log
        gungun.slab.ntt.jp - - [19/Dec/1995:11:21:53 +0900] "GET /~miura/home.html HTTP/1.0" 200 1281
        (host 名もしくは IP address, RFC931 で規定される個人情報, user 名, 日付, 時間, HTTP request の内容,
                                                server の status code, 転送された file の byte 数 )
・referer_log
        http://gungun.slab.ntt.jp/index.html —> /~miura/home.html        (リンク移動元 —> リンク移動先)
・agent_log
        NCSA Mosaic for the X Window System/2.4 (L10N-2.4.0) libwww/2.12 modified
        Mozilla/1.1N (X11; I; SunOS 4.1.4 sun4m)
        (access した人が使っている Browser 名)
```

表 2: 拡張 WWW Server の log file の例

```
gungun.slab.ntt.jp - - [19/Dec/1995:11:21:53 +0900] "GET /~miura/home.html HTTP/1.0" 200 1281
                                                "http://gungun.slab.ntt.jp/index.html"
```

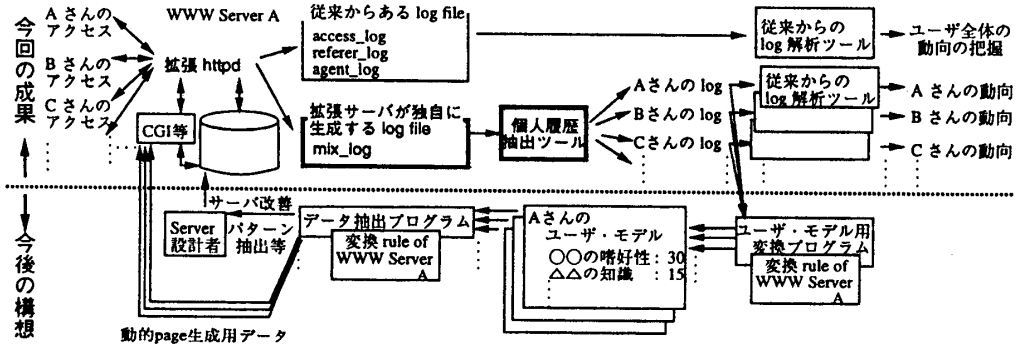

図 1: 個人履歴抽出ツールの適用とユーザモデル構築・利用

表 3: 検索方法の嗜好の分析結果 (割合 (%))

検索方法	名称別	業種別	地域別	地図上
A さん	24	69	0	7
B さん	7	88	0	5
C さん	91	8	1	0
D さん	61	38	1	0
E さん	48	50	2	0
F さん	20	31	18	31
G さん	88	2	10	0
H さん	8	23	0	69
I さん	21	28	9	42
J さん	7	46	20	27
全体 (2165 人分)	46	34	13	7

も存在することが分かる．このように，個人履歴抽出ツールの適用により，従来の log 解析ツールでは分からない，個々のユーザの性質を把握できる．

5 ユーザ・モデル構築とサーバへのフィードバック法に関する考察

前節の適用例で得られた，各ユーザの検索方法の嗜好性はユーザ・モデルの中のある一つのパラメータであると考えられる．もし，ユーザ・モデルをこのようなパラメータの集合と考える時，抽出された個人履歴をユーザモデル用の入力に変換する必要があり，この変換プログラムには，個々の WWW Server 依存の rule が必要になると考えられる．また，どのようなパラメータがユーザ・モデルに必要かという点に関しては，いわゆるロボットの収集データを基に，現存する WWW サーバを検討する方法が考えられる．

また，得られたユーザ・モデルの利用法については，現在，次の 2 つを考えている．

- ユーザの嗜好に合わせて，ユーザ毎に動的に page を生成する．
  例えば，サッカーに関する page を良く参照しているユーザには，業種別検索の業種一覧の箇条書きの中でサッカーに関するものを page の始めの方に持ってくる．
- ユーザの共通の操作パターンを抽出し，各 page 内，または page 間の構成 (User Interface) を改善する．
  例えば，業種別検索のあと，検索方法の選択に戻り，地域別の検索を行うという操作パターンが多くのユーザに頻繁に見受けられる場合，業種別検索結果から地域別検索に直接移れるようなリンクを追加する．

このように，ユーザ・モデルの利用も，個々の WWW Server に依存した rule が必要になると考えられる．

以上をまとめると，図 1の下半分のようになる．

6 おわりに

今後の課題として，以下のようなものがある．

- 個人履歴抽出問題をグラフ理論の問題に帰着させて，より厳密なアルゴリズムで解く．特に，proxy 等の cache による log の欠落を補完できるような解法を検討する．
- 5節で述べた内容を具体的に検討・適用・評価する．
- CGI が独自に生成する log を合わせて解析する．
- application 操作履歴からの UI 改善の研究 (例えば [4]) を検討し，各 page の UI 改善を考える．

謝辞

日頃，様々な面で御指導頂いている，伊藤正樹グループリーダに感謝致します．また，Japan Telephone Dirrctory の構築や日頃の議論でお世話になっている NTT 情報研の高橋克巳研究主任に感謝致します．

参考文献

[1] http://www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/HTTP/Servers/Log_Analysis_Tools/.

[2] 新井克也, 坂本啓, 中畝弘, 桑名栄二. "WWW によるグラフィカル情報サービスの構築 – WWW のユーザ行動モデル–". マルチメディア通信と分散処理ワークショップ, pp. 173–180. 電子情報通信学会, Oct. 1995.

[3] 島健一, 高橋克巳, 三浦信幸. インターネット版マルチメディア電話帳の構築. In *Japan World Wide Web Conference '95*. 日本インターネット協会, Nov. 1995. http://www.pearnet.org/jtd/paper/jtd-overview.html.

[4] 岡田英彦, 旭敏之, 井関治. "使いやすさ評価ツール"GUI テスタ"における共通操作パタン抽出方式の提案と評価". 情処研報, Vol. 95, No. 104, pp. 37–42, Nov. 1995.