Ｗｅｂサイトにおけるユーザーのふるまいに関する分析手法

4Ｓ－8

岸 晃司、坂本 泰久、坂本 啓
ＮＴＴソフトウェア研究所
日本電信電話株式会社

1.はじめに

最近ホームページを作成して情報を公開する個人、企業が増えている。ホームページがどの位の頻度でアクセスされているかを知るにはサーバの残すアクセスログを調べればよい。そのアクセスログに何らかの処理をすれば様々な情報を引き出すことができるのではないかと考えられる。そしてその結果からホームページの作成、改良に関する何らかの知見が得られる可能性がある。今回はログファイルの分析方法と、それをあるエンターテインメントサイトに適用した結果について述べる。

2. 目的

サーバに残されるアクセスログは基本的にはサーバに対するリクエストを時間順に羅列したものである。このログから比較的簡単に得られる統計として、クライアントのドメイン別のアクセス数、時間帯ごとのアクセス数、リクエストされたファイルごとのアクセス数などが挙げられる。実際、その種の統計値を計算するログ解析ソフトは複数存在する。しかしアクセスログからこれら以上の情報を得ることができるのではないかと思われる。例えば、どのような道筋でリンクがたどられているのか、またどのくらいの時間をかけてページが見られているのかなど、制作側にとって興味深いと思われる情報である。アクセスログからそのような情報を引き出すことがこの研究の第一の目的である。さらにそれをリンク構造の改良、あるいは効果的なバナー広告の挿入位置などに関する何らかのヒント、知見を得ることにつなげることが、この研究の第二の目的である。

3. ビジットについて

今回の分析では“ビジット”という概念を用いている。ビジットとは同一の IP アドレスからの連続したアクセスをまとめたものである。アクセス間が 10 分以上あいた場合は別のビジットとした。（Internet Profile 社のサービス I/PRO の“ビジット”では、10 分ではなく 30 分である。）一般的に、一つのビジットはある一ユーザーの連続したアクセスとみなしてよい。アクセスログをビジットの形に変換することにより、そのユーザーがどのような履歴でどのくらいの時間をかけてページを参照したのかが分かる。以後、ビジットの形をしたデータを元に分析を行う。

An Analysis of User behavior on a Web-site
Kouji Kishi, Yasuhisa Sakamoto, and Akira Sakamoto
{kouji, sakamoto, akira}@slab.ntt.co.jp
NTT Software Laboratories

4. 分析

4.1 アクセスログをビジット形式に変換

まず、サーバのアクセスログをビジット形式に変換する。

4.2 ページに関する冗長性の排除

次に、ビジットのデータから画像ファイルとフレームの構成ファイルに対するリクエストのアクセスを排除することによって、ページに関する冗長性をなくす。いわゆる“ヒット”ではなく、“ページ”を扱うということである。この操作の後のビジットのデータに含まれるアクセスの数を“ビジットの深さ”と呼ぶ。

4.3 ページにラベルを付与

次に、ページをその内容によって分類し、その内容に対応するラベルをそのページに付与する。これは、その web サイトを構成しているページ数がある程度大きい場合に必要な操作である。以降、分析に使うのはページそのものの名称ではなく、そのページに付与されたラベルである。つまり、内容の類似しているページは同じ種類のページとして扱う、ということである。

4.4 詳細な分析

考えられる分析の視点として以下のようなものが挙げられる。

4.4.1 全体の分析

(1) ビジットの時間 ： ビジットの最初のアクセスから最後のアクセスまでの時間。ユーザーがホームページに滞在していた時間である。

(2) ビジットの深さ ： ビジットで参照された総ページ数。ユーザーがホームページで参照したページの総数である。

(3) ラベルの種類 ： ビジットで参照されたラベルの種類。この値が大きいということは、いろいろな内容のページが参照されたということである。

(4) max ： ビジットで最も多く参照されたラベルを選び、そ

の参照回数を max と呼ぶ。max がビジットの深さに等しいということは、一種類のラベルのページしか参照されなかったということを表している。また、max が小さいということは、比較的いろいろなラベルのページが参照されたということを表している。

以上のそれぞれの視点において、① 平均値 ② ヒストグラム を求める。ヒストグラムにおいて窪みが観測された場合、その指標においてビジットの集合にクラスタリングが見える、ということである。

4.4.2 それぞれのラベルごとの分析

(1) 参照時間 ：ページが参照された時間。すなわち、そのページに対するアクセス時刻から次に参照されたページに対するアクセス時刻までの時間。

(2) 連続度 ： ラベルのページが連続して参照される度合い。ある計算式によって求められる。ページ内容、リンク構造などに依存すると考えられる。

(3) 参照回数 ： ビジットにおけるそのラベルのページの参照回数。

(4) 参照率 ： ビジットの総参照ページ数に対する(3)の割合。

(5) 参照位置 ： ビジットにおいて何ページ目に参照されたかを表す。

以上のそれぞれの視点において、① ラベルごとの平均値 ② ラベルごとのヒストグラム ③ 異ラベル間での平均値に関する相関係数を求める。

4.4.3 ページ（ラベル）の履歴に関する分析

ページ（ラベル）の履歴に関して、どのようなパターンが多く見られるか、の検証。

5.あるサイトでの結果

上に挙げた分析手法をあるエンターテインメントサイトのアクセスログに適用した結果について述べる。

5.1 結果

全体の分析に関しては、“max”に関してクラスタリングが見られた。(詳細は5.2を参照のこと)

ラベルごとの分析に関しては、ラベルの特徴がよく分かるおもしろい結果を得ることができた。特に、あるラベルの“ビジットにおける参照回数”についてクラスタリングが見られた。このことは、ある内容のページに関しては、そのページが多く参照されるビジットとほとんど参照されないビジットに分けられるということを表している。

5.2 興味深い結果が出た分析の例（max に関する分析）

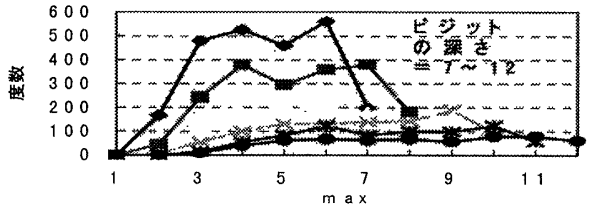

図1 max のヒストグラム

図1はmax のヒストグラムである。ビジットの深さごとに折れ線を描いた。これを見ると、山が二つあることが分かる。左の山は比較的いろいろなラベルのページを参照したビジット、右の山は比較的限定されたラベルのページを参照したビジットを表している。この視点において、クラスタリングが見られる、ということである。

6. 課題

6.1 サーバに残るアクセスログを使用する場合の問題点

サーバに残るアクセスログを元にユーザー動向を分析する場合の問題点として、ブラウザのキャッシュ機能やキャッシュサーバの存在によってユーザーが実際に参照したページの記録がアクセスログに残らないということがある。

また、ログに記録されるクライアントの IP アドレスと実際のユーザーとが一対一に対応しないという問題もある。

今回の分析に関しては、後者の問題は影響を持たない。

6.2 分析結果の応用について

今回は、分析結果をリンク構造の改良やバナー広告の挿入位置などに関する知見にフィードバックさせるにはいたらなかった。そのような知見を得るために、今後さらに詳しい分析を行う必要であると考える。

7. まとめ

比較的簡単に得ることのできるサーバのアクセスログを用いてユーザーの動向を様々な角度から分析する手段を考案し、その方法をあるwebサイトに適用したところ興味深い結果を得ることができた。特にある視点（“max”と“ビジットにおける参照回数”）においてはビジットの集合にクラスタリングを見ることができた。

8. 参考文献

[1] 坂本,岸,“ユーザアクションにもとづく Web サーバアクセス履歴の分析”,情報処理学会 Interaction97(1997)