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

坂本 泰久(sakamoto@slab.ntt.jp) 岸 晃司(kouji@slab.ntt.jp)
NTTソフトウェア研究所 東京都武蔵野市緑町 3-9-11 TEL 0422-59-2398 (fax 2489)

概要

本論文は、インタラクティブサービスとしてのWebのマーケティング的側面に注目し、アクセス履歴から有効なマーケティング情報を得るための分析手法に関するものである。

はじめに

Webの普及にともなって、提供される情報の量や種類が膨大になり、また品質のばらつきも大きくなっている。これは情報氾濫に対する選択限界という利用者側の問題として扱われることが多いが、情報を伝えたい人に伝えることができないという提供者側の問題でもある。つまりは情報を伝えたい人と情報を得たい人の間の適切な情報のマッチングに対する要求が増大している。この要求に対し、Webのようなインタラクティブサービスの持つ本質的な双方向性を前提として、適切な流通支援の仕組みをシステムに組み込んでいこうというアプローチがあり、インタラクティブマーケティングと呼ばれている。

インタラクティブマーケティングの基本的な考え方は、利用者のニーズを何らかの形で顕在化させ、それにもとづいて利用者の情報選択を支援する一方、提供者側は蓄積されたニーズを分析し解釈して利用動向を把握し、提供するサービスに反映するということである。顕在化されたユーザニーズは通常プロファイルと呼ばれる。情報フィルタリングにおけるモデル[1]などをもとに整理したインタラクティブマーケティングにおける情報の流れを図1に示す。

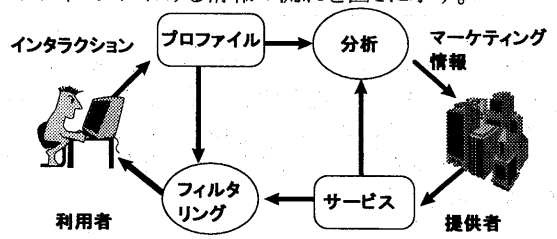

図1 インタラクティブマーケティングにおける情報の流れ

支援システムの実現にあたっては、フィルタリングとマーケティングの両方に活用できるプロファイルの定義とシステム上への獲得方法が主要な課題となる。インタラクティブマーケティングの観点から見ると、ユーザのインタラクションの結果であるWebのアクセス履歴は、最も手軽に得ることができるユーザプロファイルとみなすことができる。

アクセス履歴分析の目的

Webサイトのアクセス履歴分析の目的は、大きく内部的なものと対外的なものの2つがある。

内部目的とは、アクセス量がどのように推移しているか、どの内容が人気があるか、どのような時間帯に利用されているかなど、ユーザの利用傾向をWebサイト運用者が知り、サービスの向上に役立てることである。

対外目的とは、何人のユーザが利用しているか、広告ページの参照回数はどのくらいかなど、コンテンツ提供者や広告主向けに提示するためのメディア効果指標を算出することである。

いずれの場合にも、Webアクセス評価の基準となる測定単位を明確にする必要がある。

Webアクセスの基本測定単位

基本単位としては、TVやラジオの場合と同様に、参照時間、つまり利用者が情報を見ていた時間が、Webサイト種類に依存しない公平な尺度であると考えられる。

しかし参照時間の測定は技術的に困難であることから、インタラクティブサービスにおける利用者の能動的なアクションすなわちシステムへの要求動作を基準とするのが一般的である。Webにおける要求動作としては、ハイパーリンクのクリックのほか、URLの直接入力やブックマーク使用なども含まれる。基本単位としてユーザの一回の要求動作で得られる情報をページ、同一ユーザからの連続したページ表示要求をビジットと定義する(表1参照)。名称などの違いはあるが、これらは米国を中心とした多数の組織において標準的な単位として認知されている。 例えば Novak と Hoffman は、現在使用あるいは推奨されている指標を総合的に調査し、ページとビジットを基本構成要素とする包括的なWebアクセスの指標を提案している[2]。

名称	定義
ページ	ユーザの一回の要求動作により表示される情報。
ビジット	同一ユーザからの時間的に連続した一連のページ表示。ある一定期間以上の間隔があいた場合に別ビジットとする。

表1 Web分析の基本測定単位

サーバログからの基本単位の抽出手法

Webサーバで得られたアクセス履歴(以後サーバログ)からページとビジットを抽出する手法は次の通りである(図2参照)。一行が一リクエストを表すテキストファイル形式のサーバログを対象とする。

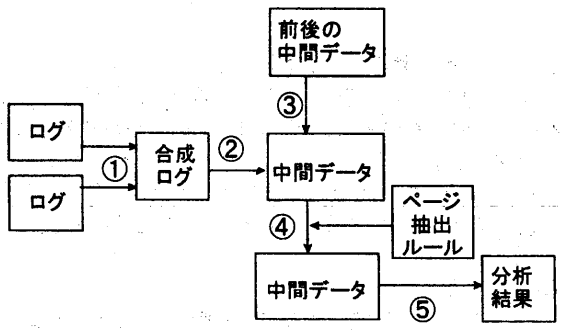

図2 基本単位抽出のためのサーバログ分析の流れ

①Webサイトが複数サーバで構成されている場合は、それぞれのサーバログを合成し、リクエストを時間順に並べ替える。
②同一IPアドレスからの時間的に連続したリクエストを抽出し、ビジットとしてまとめた中間データに変換する。
③前後の期間の中間データと比較して、期間をまたがるビ

ジットの結合を行う。
④サイト構造にもとづいたページ抽出ルールを参照しながら、全リクエストのうちページだけを取り出す。イメージやフレームなどの構成要素が削除される。
⑤中間データをもとに各種の統計分析を行う。

ツールの互換性の観点から、ビジットを表現する中間ファイルのデータ形式(以後SDF[1]形式と呼ぶ)の定義および共通化が重要となる。現在の仕様は次の通りであるが、逐次改良していく予定である。

SDF ≡ <時刻>, <クライアントIP>, <ページ数>, <開始時刻>, <終了時刻>, { <URL?パラメータ>, <相対秒数> }+
{ }+ は1個以上の繰り返しを示す

大規模Webサーバでの適用評価

大規模なアクセスを有するWebサイトにおいて上記の手法を適用し、統計分析を実際に行った。その結果、以下のような効果が検証された。

(1) 効果指標の算出に関して(対外目的)

ユーザ認証を行わないサーバから推定できる値の中では、ビジット数[2]は、ページ数やリクエスト数に比べ、サイト間で比較する効果指標として適している。図3は、日ごとのビジット数、平均ビジット深さ(ビジットあたりのページ数)、平均ページ深さ(ページあたりのリクエスト数)の3ヶ月間の推移を示している。深さは、アクセス量にかかわらず安定しており、ビジット深さ(約4)はサイト滞留度、ページ深さ(約6)は画面の複雑度にあたるサイト特性を表している。したがってこれらの特性が排除されたビジット数はサイト依存性が低い正規化された値だといえる。

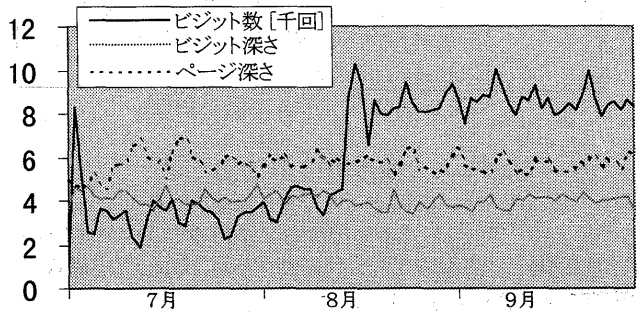

図3 ビジット数の推移

(2) ユーザ行動の把握に関して(内部目的)

ビジット内におけるページ参照履歴、いわゆるアクセスパターンの分析を行うことにより、ユーザ挙動の統計的把握が可能となる。例えば、図4はビジットのどの位置で外部サイトへジャンプしたかついての頻度を示しており、ビジットの後ろほどジャンプ率が高くなっていることがわかる。

(3) その他

ログデータが圧縮できる。SDF形式は冗長な情報の繰り返しがないため、情報量を落とさずにバイト数を少なくできる。実験では元のログの約5分の1になった。

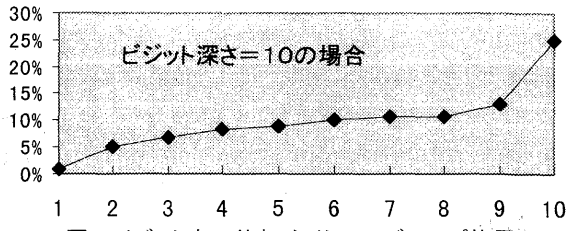

図4 ビジット中の外部サイトへのジャンプ位置

今後の課題

(1) ビジットの時間間隔の定義

ビジットを切り出す判定に用いる時間間隔を今回は10分と設定したが、妥当性に関しては検証できていない。サイト依存性が強いと予想されるので、複数サイトでの比較実験が必要である。

(2) ログの不完全性による誤差の補正

サーバログはユーザ行動と完全に対応しているわけではない。下記のような問題に対処するためには、クライアント側あるいは中間サーバ側で得られるアクセスログにより補正するなどの対処が必要となる。

- キャッシュ機構によるログの減少
  プロキシーサーバが典型的な例である。
- 自動巡回機構によるログの増大
  検索サイトやオートパイロットソフトなど。
- クライアント側との時刻の不一致
  参照時間算出にあたっては、ネットワーク遅延や、ユーザによるブラウジングの中断などの考慮が必要。

(3) ユーザの特定

ユーザ認証を行わないサイトの場合、識別単位はマシンのIPアドレスであり、ビジットにおいてもユーザと厳密には1対1には対応していない。例えばプロキシーサーバの場合、複数ユーザが1つのIPアドレスに対応している。

(4) アクセスパターン分析支援ツール

前節(2)で述べたようにビジット分析により、最終的にはセグメンテーションやターゲット広告などに応用できるユーザモデルが得られる可能性がある。しかし実際には、サイト依存性が高く、多くの試行錯誤が必要となる。多様な列データ(SDF形式)を、さまざまの観点から視覚的に分析できる環境が求められる。

まとめ

Webサーバアクセス履歴から、ユーザアクションにもとづく基本測定単位であるビジットおよびページを抽出する手法を示した。この手法を実際のWebサイト分析に適用し、マーケティング分析上の有効性、すなわち(1)ビジット数などの統計値がWebサイトの効果指標として適していること、(2)ビジット内の参照パターン分析によりユーザ行動の把握が可能なこと、を検証した。

参考文献

[1] 森田、速水,「情報フィルタリングシステム」、情報処理 Vol. 37, No. 8 1996/8 (1996)
[2] Novak and Hoffman, "New Metrics for New Media: Toward the Development of Web Measurement Standards", Project 2000 White Paper (1996) http://www2000.ogsm.vanderbilt.edu/

1 Session Data Formatの略。
2 Novakは、この値をSite Exposureと定義している[2]。