LO-004

IRT による情報提供型 Web サイトのアクセス分析
Web Access Analysis with IRT for an Information-Supplying Type of Website

橋間 智博 †
Tomohiro Hashima

1. はじめに

今日，Web サイトは経済活動における役割がより大きくなってきている．また，インターネット上の情報量は爆発的に増加している．そうした中で，利用者の利便性を高め，さらに情報を発信する力を高めるために，受動的な情報発信ではなく，Web サイトから訪問者へ積極的に働きかける，能動的な情報発信が求められている．

本研究では，項目反応理論を用いて Web サイトアクセスログ解析を行うことで，Web サイト内における各ページの特性を求め，訪問者を目的別にクラスタリングすることを目的として研究を行った．Web サイトを用途別に，情報提供型，コミュニティ型，物販型と大きく 3 つに分ける．情報提供型は Web サイトから訪問者に情報を提供することのみに目的をおいたサイトであり，従来からある紙媒体によるパンフレットに近い Web サイトである．本研究では情報提供型の Web サイトについてアクセス分析を行った．

2. 項目反応理論

本章では，項目反応理論 [1] について述べる．

2.1 概要

主に教育心理学の分野で古典的テスト理論を代替する形で発展した潜在変数分析のひとつであり，学力テストや性格検査などによく用いられる．

2.2 項目特性曲線

各項目は，その困難度と測定する特性値 θ との相関関係の強さなどでそれぞれ異なる特徴を持ち，各項目の正答確率を特性値 θ の関数で表すことができる．2 パラメータ・ロジスティック・モデルにおいて，a:項目識別力,b:項目困難度とすると

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{2.1}$$

と表すことができる．

2.3 項目情報関数

テスト全体の測定精度に対する各項目の貢献度を表し，2 パラメータ・ロジスティック・モデルにおける項目情報関数は式 (2.2) で示される．

$$I_i(\theta) = 1.7^2 a_i^2 P_i\{1 - P_i(\theta)\} \tag{2.2}$$

2.4 テスト情報関数

テストの精度を適切に表現するために，テスト情報量を求める必要がある．また特定のレベルに限定しないテスト情報関数は式 (2.3) で示される．

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta) \tag{2.3}$$

†兵庫県立大学経済経営研究所

3. 解析対象 Web サイトの概要

本章では，本研究で使用したアクセスデータの収集元となった Web サイトの概要を述べる．

3.1 概要

対象とした Web サイトは兵庫県立大学経済学部の Web サイト (www.econ.u-hyogo.ac.jp) である．本サイトは 2005 年 4 月から運用されている．Web サイトの内容は，入試情報を中心とした，受験生へ向けた内容と，大学本部サイトよりも学部独自の情報元として，学部生向け情報提供の為のサイトとなっている．アクセスの半分は兵庫県立大学の本部 Web サイト (www.u-hyogo.ac.jp) からとなっている．約 85%の訪問者がトップページから訪問している．また，検索エンジン経由での訪問は約 8%となっている．

3.2 Web サイトの構造

トップページ上部には大学本部 Web サイト，英語版トップページ，サイトマップへのリンクがある．また，目的別メニュー群と訪問者別メニュー群があり，それぞれのメニューページへリンクされている．そして，それぞれのメニューページから各コンテンツへとリンクされている．例えば,「受験生の方へ」メニューには,「オープンキャンパス」「入試日程」などの内容へのリンクがある．図 1 にサイト構造図を示す．

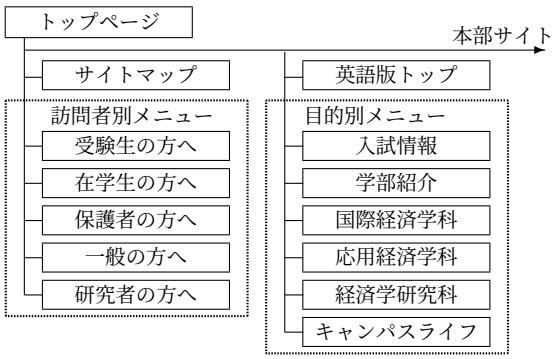

図 1: サイト構造

3.3 アクセスログの収集

表 1 にアクセスログ収集時の条件を示す．

表 1: アクセスログの条件

期間	2005 年 11 月 1 日から 2006 年 3 月 31 日
アクセス数	12665 件
ページ数	69 ページ

アクセスログは文献 [2] で筆者が作成した，アクセスログ解析ツールを用いて集計を行った．なお本サイトは CMS(Content Management System) である XOOPS を用いて構築されている．各コンテンツへのリクエストはクエリを含め集計する必要があった．そこでアクセスログ解析ツールに必要な機能を追加し，集計を行った．

4. 項目反応理論を用いたアクセス分析

本章では，Web サイト内におけるアクセス分析への項目反応理論の適用方法と分析結果を述べる．

4.1 項目反応理論のアクセス分析への適用

分析に使用したアクセスデータは 3. で述べたサイトのアクセスログデータを用い，4 ページ以上を閲覧した訪問者のみを対象とした．また，サイト管理者のアクセスと，検索エンジンのクローラによるアクセスも対象から削除した．対象となる訪問者数は 3653 件である．

Web ページ（以下，ページと呼ぶ）の閲覧の有無を項目とし，閲覧回数が一定以上のページを項目として採用した．訪問者 i がページ j の閲覧時を 1，非閲覧時を 0 とし，反応パタンベクトル U_{ij} を生成した．U_{ij} を用いて項目母数と被験者母数を推定し，訪問者の特性値 θ を推定した．

4.2 項目母数の推定

38 ページ (項目) においてベイズ推定法を用いて項目母数の推定を行った．推定された結果から極端な値をもつ項目を取り除いた．具体的には，項目困難度の絶対値が 4.0 を超えるものを取り除いた．表 2 に主な項目の項目識別力と項目困難度を示す．

表 2: 推定された項目母数

項目	識別力	困難度	ページタイトル
u_3	1.4285	1.4837	入試情報 - 募集定員
u_7	1.0315	1.8485	入試情報 - 資料請求
u_{20}	1.4352	1.2062	入試情報 - 入試日程
u_{21}	1.4429	1.1705	入試情報 - 選抜方法
u_{25}	1.0767	1.1017	入試情報 - 入試結果
u_{26}	-0.9980	-1.1089	大学院 - 教員紹介
u_{27}	-1.6475	-0.8150	応用経済学科 - 教員紹介
u_{28}	-1.2675	-0.7417	国際経済学科 - 教員紹介
u_{29}	1.1865	0.8926	入試情報 - 出願状況
u_{31}	0.7661	0.7804	入試情報 - オープンキャンパス
u_{33}	-0.5481	-0.9503	大学院 - 概要
u_{34}	-0.5999	-0.6189	応用経済学科 - 理念と特色
u_{35}	-0.4293	-0.3629	国際経済学科 - 理念と特色
u_{37}	1.1990	-0.0347	訪問者別メニュー - 受験生

トップページ ($j = 38$) は取り除いた項目の一つである．分析対象の訪問者 3653 件中 3296 件がトップページから閲覧し，また，訪問中にトップページを閲覧した訪問者は 3470 件である．約 95%の訪問者がトップページを閲覧しており，トップページの困難度は -7.92 と非常に小さな値となった．したがって，トップページ等の，ほとんどの訪問者が閲覧するようなページは，項目反応理論によるアクセス分析に用いる項目としては，情報量が無いに等しいことから，削除した．

4.3 項目母数の解釈

表 2 の項目の中で入試に直接的に関係するページ（u_3,u_7,u_{20},u_{21},u_{25},u_{29}, u_{31},u_{37}）をグループ 1 とする．グループ 1 以外で，学部や大学院の情報を提供するページ（u_{26},u_{27},u_{28},u_{33},u_{34},u_{35}）をグループ 2 とする．

図 2 に 4.2 で項目母数の推定を行った項目の項目特性曲線を示す．図 2 の中でグループ 1 は単調増加，グループ 2 は単調減少の項目特性曲線になっている．グループ 2 は単調減少の曲線になっており，逆転項目になっていることがわかる．グループ 2 がグループ 1 に対して逆転項目になっているということは，グループ 1 のページを閲覧した訪問者は，グループ 2 のページを閲覧する確率は低く，逆に，グループ 2 を閲覧した訪問者はグループ 1 のページを閲覧する確率が低くなることを示している．

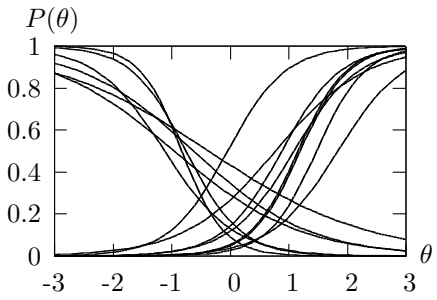

図 2: 項目特性曲線

4.4 被験者母数の推定と解釈

4.2 で推定した項目母数から訪問者の特性値 θ を推定した．特性値 θ の分布がどのようになっているか見るために，ヒストグラムを図 3 に示す．ヒストグラムの平均値は -0.135，分散は 0.774 であった．

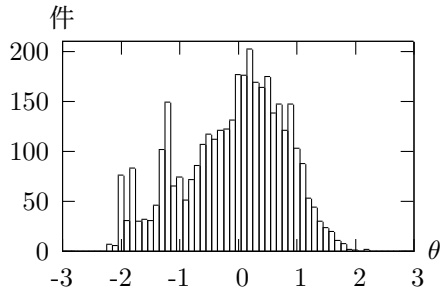

図 3: 特性値 θ の分布

訪問者の特性値 θ の分布は標準正規分布を仮定して推定を行った．しかし，$0 \leq \theta \leq 1, -2 \leq \theta \leq -1$ において偏りが現れている．これは，$0 \leq \theta \leq 1$ の特性値 θ をもつ訪問者と，$-2 \leq \theta \leq -1$ の特性値 θ をもつ訪問者がそれぞれにクラスタを形成していると考えられる．以下，$0 \leq \theta \leq 1$ の特性値 θ をもつ訪問者のクラスタをクラスタ 1，$-2 \leq \theta \leq -1$ の特性値 θ をもつ訪問者のクラスタをクラスタ 2 とする．

2 パラメータ・ロジスティック・モデルの場合，訪問者の特性値 θ は，閲覧したページの識別力をすべて足しあわせることで推定することができる．よって，クラスタ 1 に含まれる訪問者は，項目識別力 a の値が大きいページを多く閲覧したことになり，クラスタ 2 に含まれる訪問者は，項目識別力 a が小さく負になるようなページを多く閲覧したことになる．したがって，クラスタ 1 はグループ 1 のページを多く閲覧した「受験生」が多く含まれると推測され，クラスタ 2 はグループ 2 のページを多く閲覧した「在学生」などの「受験生」以外が多く含まれると推測される．

4.5 IRT によるアクセス分析とアクセスログの検証

4.4 での推測と，実際のアクセスログが一致しているかを調べるために，アクセスログ解析ツールでの集計時に，訪問者の特性値 θ を指定できるように，検索条件に特性値 θ の上下限を指定する条件項目を追加した．

まず，クラスタ 1 の訪問者について集計を行った．条件に一致した訪問者は 1544 件であった．検索エンジンから訪問した訪問者の検索キーワードには，「18 年度出願状況」「出願状況 兵庫県立大学」「兵庫県立大学入試」などの入試に関係する検索キーワードが多く見られ，検索エンジンからの訪問者の約 50%を占めた．図 4 に日付別の訪問者数のグラフを示す．訪問者数のグラフではセンター試験直後の 2006 年 1 月 23 日から数日間，アクセス数が急増しているのが見て取れる，さらに前期日程の合格発表日である，3 月 6 日から数日間もアクセス数がかなり増えている．

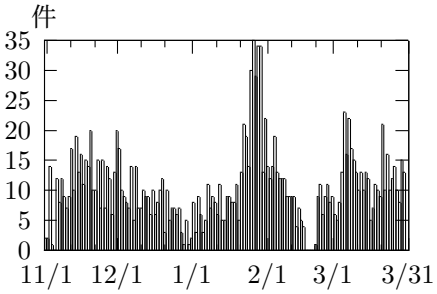

図 4: 日付別の訪問者数

次に，クラスタ 2 の訪問者について集計を行った．条件に一致した訪問者は 643 件であった．検索エンジンから訪問した訪問者の検索キーワードに入試に関係するキーワードはなかった．主に目立ったキーワードとしては，大学名と学部名をあわせたものが約 42%，教員名が約 27%，科目名が約 10%，残りは体育会のクラブ名などであった．また，日付別のアクセス数については，クラスタ 1 の日付別アクセス数と違って，集計期間すべてにおいて平坦なグラフとなっており，センター試験などでアクセス数が変化することはなかった．

以上の結果から，クラスタ 1 付近には受験生と考えられる訪問者が多く含まれ，クラスタ 2 には受験生と考えられるような訪問者は含まれていないということがわかる．よって，4.4 での推測と，実際のアクセスログから読み取れる訪問者の傾向は，ほぼ一致している．

4.6 テスト情報関数による判定精度

図 5 にテスト情報曲線を示す．このグラフは，どの程度の特性値を持つ訪問者に対して，判別の精度が最もよくなるかを示している．

テスト情報曲線を見ると，訪問者の特性値 θ が -0.6 付近と 1.2 付近で測定の精度がピークを迎えていることがわかる．$\theta = -0.6$ のときの標準誤差が 0.404，$\theta = 1.2$ のときの標準誤差が 0.337 であった．$-0.6 \leq \theta \leq 1.0$ において測定精度が高く，特にクラスタ 1 とクラスタ 2 を判別するために重要である $-1.0 < \theta < 0.0$ の範囲で測定精度が高いことは，非常に良好な結果と言える．

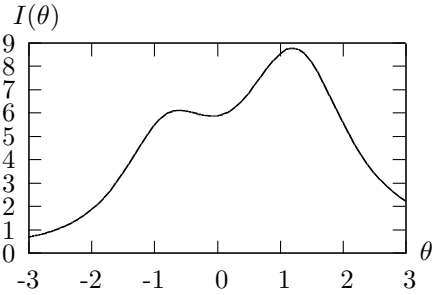

図 5: テスト情報曲線

5. まとめと今後の課題

本研究では，項目反応理論を用いて，アクセスログから Web サイト内における各ページの特性を求め，訪問者の訪問目的ごとにクラスタリングを行うことを目的とし，研究を行った．本研究の結果から，入試関連のページと学部在学生向けのページが逆転項目になるなど，Web サイト内における各ページの特性を項目特性曲線として得ることができた．訪問者を訪問目的ごとにクラスタリングすることについては，どちらのクラスタにも属さない訪問者が残ってしまったが，大きくは 2 つのクラスタに分けることができた．しかも，対象の Web サイトにおいて，最も重要な訪問者である「受験生」を，かなり精度良く判別できた意味は大きいと思われる．

今後の課題としては，以下の点があげられる．

2 値データ以外での分析

本研究では，各ページについて訪問者の閲覧の有無だけの，2 値データモデルとして扱った．しかし，ページ数が多く，メニューが多階層になっている場合などは，同じページを繰り返し閲覧することになる．こういった場合，多値項目反応モデルを用いることで，行動の分岐点となるページの特性を分析することが容易になるのではないかと考える．

リアルタイムでの分析と動的なナビゲーション

訪問者の Web サイトに求める目的をリアルタイムに判別することができれば，訪問者を目的のページへと誘導することが可能になり，訪問者が目的のページを閲覧できずに逃げてしまうといった事が減少すると考える．リアルタイムでの分析と動的なナビゲーションによって，訪問者の利便性向上につなげることができると考える．

サイトの欠陥発見のための応用

ナビゲーションに欠陥があると，サイト設計者が意図していない特性をナビゲーションページが持ってしまう．項目特性曲線を用いることで，欠陥を発見する手法への応用が考えられる．

謝辞 本研究を進めるにあたり，ご指導頂いた兵庫県立大学経済学部秋吉一郎教授に心より感謝いたします．

参考文献

[1] 豊田秀樹 著: ”項目反応理論 [入門編・理論編]” 朝倉書店

[2] 橋間智博 著: ”Web サイト訪問者の分類を行うアクセス解析ツール” FIT2004, pp. 413-416 O-022