情報処理学会論文誌　コンシューマ・デバイス & システム　Vol.3　No.3　85–93 (July 2013)

研究論文

Webアクセスログに基づくユーザの革新性推定手法の提案

市川 裕介[1,a)]　岸本 康成[2,b)]　小林 透[1,†1,c)]

受付日 2012年12月14日, 採録日 2013年4月26日

概要：本稿では，ユーザの Web サイトへのアクセスログから，革新的ユーザか否かを推定する手法を提案する．革新的ユーザとは，新製品に興味を持ち自身の考えを持って商品を選択することができるユーザを指し，一般的にイノベータやアーリーアダプタと呼ばれるものである．推定にあたっては，革新的なユーザと保守的なユーザは EC サイトにおける Web ページの参照行動，具体的には各種 Web ページへの参照頻度が異なるという仮説に基づき，ユーザの購買商品の平均採用時期を正解データに，EC サイトの Web ページへのアクセス回数データを素性にして，Support Vector Machine を適用して革新的なユーザを推定する手法を提案する．本手法を，7,000 名規模のパネルから収集した Web アクセスログを用いて評価した結果，正解データの妥当性，およびアンケート回答を用いた方法を上回る精度での推定ができることを確認した．

キーワード：サイコグラフィック属性，イノベータ，アクセスログ，EC，Web，SVM

A Proposal of Extracting Innovative Users with Web Access Log of an E-commerce Site

Yusuke Ichikawa[1,a)]　Yasunari Kishimoto[2,b)]　Toru Kobayashi[1,†1,c)]

Received: December 14, 2012, Accepted: April 26, 2013

Abstract: We study a way of extracting user groups having a specific psychographics. Especially we define innovative users as that segment interested in buying new items. Identifying such users will allow more firms to develop and market new products more efficiently. This paper assumes that innovative users are interested in specific information found on particular web pages on an e-commerce website. Once those web pages have been identified, new users can be categorized as innovative based on the pages they access. We propose a method of analyzing web access logs to classify users as innovative or not and report its effectiveness.

Keywords: psychographics, innovator, access log, e-commerce, Web, SVM

1. はじめに

ブロードバンドネットワークの普及により，商品の情報収集や購入を目的としてインターネットを利用するユーザが増えている．平成 22 年度に総務省が行った調査によると，日本のインターネットの人口普及率は 78.2%に及び，インターネット利用者の 44.3%が商品・サービスの購入・取引を目的としてインターネットを利用すると回答している [1]．また，小売市場が縮小している中で，消費者向け電子商取引（BtoC EC）市場は継続的上昇傾向を示し，平成

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[2] 日本電信電話株式会社 NTT ソフトウェアイノベーションセンタ
NTT Software Innovation Center, NTT Corporation, Musashino, Tokyo 180–8585, Japan
[†1] 現在，長崎大学大学院工学研究科
Presently with Graduate School of Engineering, Nagasaki University
[a)] ichikawa.yusuke@lab.ntt.co.jp
[b)] kishimoto.yasunari@lab.ntt.co.jp
[c)] toru@cis.nagasaki-u.ac.jp

© 2013 Information Processing Society of Japan

22 年の経済産業省の調査では，前年比 16.3%増の 7.8 兆円となった．EC の浸透を示す指標である EC 化率も前年比 0.4 ポイント増の 2.5%になっている [2]．このように，今後インターネット上での消費活動の増加にともない，インターネット上の各サーバに蓄積された Web アクセスログの量も増加すると考えられる．Web アクセスログには多くのユーザの行動が記録されていることから，今後多くのユーザの動向を分析し，マーケティング等に活用したいと考える事業者等によって Web アクセスログが有効利用されることが期待される [3], [4]．

Kotler らによると商品の販促活動において，売り手は個々の買い手に向けて別々のマーケティングプログラムを設計することが理想であり，個々の買い手をとらえる方法として消費者を各種属性によって細分化する方法を紹介している [5]．従来，ユーザを細分化するための属性として，性別や年齢といったデモグラフィック属性がさかんに利用されてきたが，ユーザの価値観の多様化により性別や年齢だけではユーザの実態をとらえることが難しくなったといわれている．そこで，ユーザの活動や関心事，意見あるいはパーソナリティや態度といったサイコグラフィック属性に注目が集まっており，サイコグラフィック属性を用いてユーザを細分化し，実態をとらえようとする傾向にある [5], [6]．

サイコグラフィック属性の代表的な事例として，Rogers が提唱したイノベータ理論がある [7]．イノベータ理論ではユーザはイノベーションに対する採用時期の早さにより，イノベータ，アーリーアダプタ等の 5 つのグループに分類し，各グループにおけるライフスタイルやパーソナリティの違いがあることを示している．

我々は，サイコグラフィック属性の 1 つとして，Rogers が提唱したイノベータ理論を参考に，新製品に興味を持ち自身の考えを持って商品を選択することができるユーザ（イノベータとアーリーアダプタに該当するユーザ）を革新的ユーザと定義し，商品購入に向けた情報収集や実際に商品購入を行う EC サイトのアクセスログを用いて革新的ユーザを抽出することを目指している．たとえばユーザに商品のクチコミ情報等を提供する情報収集サイトにおいて，新商品に興味を持つ革新的ユーザの商品に対する評価情報は，新商品のその後の売り上げを予測したり，革新的ユーザにいち早く新商品情報を提供してブログ等で商品に対する感想を書いてもらうことにより，革新的ユーザを販売促進活動の 1 つとして活用することもできる．このように，革新的ユーザの抽出は，より多くの商品を販売したい事業者にとって，マーケティングに利用できる重要な意味を持つ．

本稿では，先述の商品情報収集サイトや EC サイトを適用対象サイトとして，サイトへのアクセス傾向の違いに着目した革新的ユーザの推定・抽出手法を提案する．以下，2 章で従来の革新的ユーザ抽出手法とその課題について説明し，3 章で提案する推定手法について説明する．4，5 章で実ログデータを用いた提案手法の評価実験とその結果，考察について述べ，6 章でまとめについて述べる．

2. 従来の推定手法と課題

従来，イノベータ理論のイノベータやアーリーアダプタといった属性を持つユーザを抽出する方法として，アンケートデータを分析する方法が用いられてきた．たとえば，斉藤は心理スケールを構築するために 700 項目の変数を用意し，アンケートから日本人の価値観の類型化することにより，イノベータ理論の分類に適応するグループの抽出を実現している [8]．

しかし，ユーザにとってアンケートへの回答負担は大きく，すべてのユーザがアンケートに回答するわけではない．そのため，アンケートに回答した特定のユーザのみしか対象にできない推定手法では，大規模に特定の属性に該当するユーザを抽出することに適さないという課題がある．

一方で，商品レコメンデーションに活用することを目的として，ユーザ間やアイテム間でのコンテンツ採用の伝搬率や順番を考慮して，ユーザごとに先にコンテンツを採用したユーザを革新的ユーザとして抽出する手法が提案されている．Song らは過去の購買履歴から作成される information flow network を用いて将来の購買を予測する手法を提案している [9]．Kawamae らは時間が経過すると指数関数的に他へ及ぼす影響力が減少することを仮定しモデル化した personal innovator degree を提案している [10]．

これらの手法は，アンケートのような明示的入力を必要としない．しかし，特定の商品に対するユーザの購買順序を用いているため，多くのユーザが同一の商品を購入する分野（たとえば家電製品等）においては有効だが，各商品の販売数が少なくユーザごとに購入商品が異なるような場合（たとえば，他人と被ることが嫌われる服飾のような分野や，商品種類数が多く興味分野が広範囲に分散しがちな書籍・音楽等の分野）においては利用できない．

本研究では，以上で述べた従来手法の課題である，(1) アンケートのようなユーザの明示的な入力に頼らず，(2) 多くのユーザに共通する特定の購入商品がなくても推定可能な方式の構築を目的とする．

3. 革新的ユーザ推定手法の提案

3.1 提案概要

Rogers は，ユーザのイノベーション採用時期の違いによって，パーソナリティやライフスタイルに差があり，異なる行動の特徴を持つことを示している [7]．このことは，ユーザのサイコグラフィック属性によって，インターネット上での行動特性も異なる可能性があることを示唆しており，それにともない Web ページへのアクセス傾向も異な

© 2013 Information Processing Society of Japan

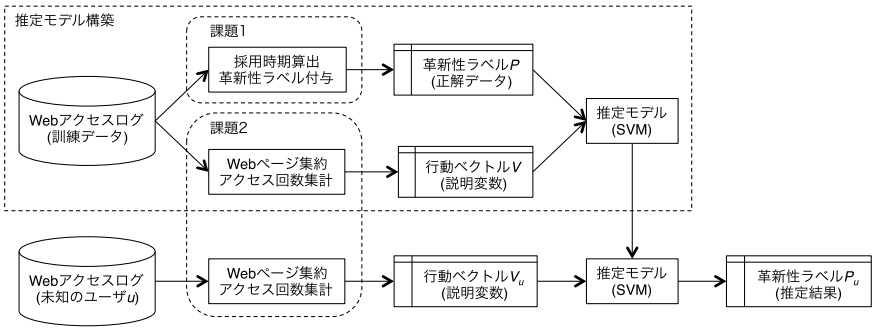

図 1　革新性推定概要
Fig. 1　Outline of extracting innovative users.

る可能性がある．筆者らは以前の研究において，ユーザの Web アクセスログをサイト単位で集計すると，アクセス回数の分布がべき乗則を示し，そのべき指数の大きさの違いにより，Web 利用の目的がはっきりしているユーザと多様な目的で Web を利用するユーザを分類できる可能性があることを示した [11]．同様に，あるユーザが革新的ユーザか否かについても，Web アクセスログから得られる Web ページへのアクセス傾向の違いを用いて推定できると考えられる．

たとえば，新製品に興味を持つユーザは Web サイトにおいて新商品を紹介する Web ページを頻繁に閲覧するだろう．また，自身の考えで商品を選択するユーザは商品の選択に必要な情報を収集するために，商品の詳細情報が掲載されている Web ページにアクセスすると予想される．逆に保守的なユーザはランキングや価格情報を気にして，それらの情報が得られるページへのアクセスが多いかもしれない．このように，ユーザが革新的ユーザか否かによって興味を持つ情報が異なるため，アクセス先の Web ページもまた異なることが予想される．そこで，Web ページが含んでいる情報の種類（以降，「ページ種別」と呼ぶ）によって Web ページを複数の Web ページ群に分類し，各種 Web ページ群に対するユーザのアクセス傾向の違いに着目することで革新的ユーザの抽出が可能であると仮設を立てた．

ユーザの各種 Web ページ群へのアクセス傾向を定量化する手段として，行動ベクトルというベクトル量を導入する．ユーザ i の行動ベクトル V_i は下記のように定義される．

$$V_i = (v_{i1}, v_{i2}, v_{i3}, \cdots, v_{ij}, \cdots, v_{ip}) \tag{1}$$

式 (1) の行動ベクトル V_i の次元数 p は Web ページ群の数であり，p 種類の Web ページ群に対応してベクトルの要素が存在することを示している．行動ベクトルの j 番目の要素 v_{ij} は，ユーザ i の j 番目のページ種別の Web ページ群 x_j へのアクセス回数 a_{ij} に正規化等の処理を施した数値である．ここで，Web ページ群 x_j へのアクセス回数 a_{ij} とは Web ページ群 x_j に含まれる各 Web ページへのアクセス回数の総和で定義される数値である．

前述した仮説より，行動ベクトル V_i からユーザ i の革新性ラベル P_i を導出する推定モデルが存在すると考える．ここで革新性ラベル P_i とはユーザ i が革新的ユーザか否かを表す 2 値のラベルを表している．推定モデルを構築することにより，革新性ラベルが未知のユーザ u に対して，Web アクセスログから算出可能な行動ベクトル V_u を用いて革新性ラベル P_u を推定することが可能となる．推定モデルの構築には機械学習の 1 つである教師つき学習を用いる．教師つき学習は行動ベクトル V_i と革新性ラベル P_i の組 (V_i, P_i) を複数組用意し，これを訓練データとして用いることにより行う．なお，教師つき学習には Web サイトに含まれる Web ページの数に比例して行動ベクトル V_i が高次元になる可能性が高いことを考慮し，高次元の線形識別問題に有効な Support Vector Machine（SVM）[12] を用いる．以上述べた革新性ユーザ推定のフローを**図 1** に示す．

ここで課題となるのが，課題 1：教師データとなる革新性ラベル P_i をアンケートに頼らない方法でいかにして付与するか，課題 2：Web ページをいかにして Web ページが含んでいる内容の種類によって各種 Web ページ群 x_j へ集約して行動ベクトル V_i を生成するかの 2 点である．以下，それぞれ革新性ラベルの付与方法，および行動ベクトルの生成方法について詳細を述べる．

3.2　課題 1：革新性ラベルの付与方法

Rogers の定義ではユーザのイノベーション採用時期の違いによって，ユーザが革新的か否かの分類を行っている．そこで，あるユーザが商品を購入する際に，その購入商品が発売開始されてからどれだけ経過しているか，平均経過時間が短いか長いかによって革新性分類する方法を提案する．

これは，革新的なユーザはイノベーションであるか否かに限らず，発売直後の新製品を購入する傾向が強く，そうでないユーザは発売されてからある程度期間が経って，その信頼性が口コミ等によって知られている商品を買う傾向があるという仮説に従っている．

© 2013 Information Processing Society of Japan

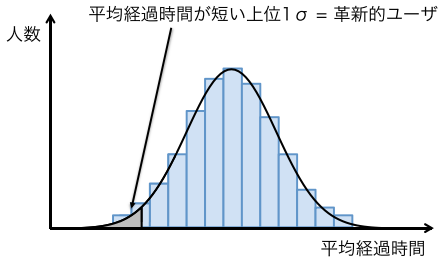

図 2　ユーザ毎購入商品の発売日からの平均経過時間の分布
Fig. 2　Distribution of the average time from the release date of purchase items.

また，ユーザの革新性は，特定の商品群ごとに異なる可能性があることを考慮して，サイト別に分けて行うことを提案する．たとえば，電子機器について新商品を積極的に購入していても，服飾については無難なものを選ぶといったように，同一ユーザでも商品群ごとに異なる革新性を持つと考えるのが自然なためである．

革新性ラベルの付与は以下の 2 ステップで行う．

Step1　購買履歴等に基づき，ユーザが購入した商品購入日と商品発売日の差（経過時間）を求め，ユーザごとにその差の平均値を算出する．平均経過時間の分布は**図 2** に示すように，Rogers が定義する人数分布と一致すると考えられる．

Step2　Step1 の平均経過時間の分布から，平均経過時間が全体の -1σ より短いユーザを抽出し，革新的ユーザとしてラベル付与を行う（図 2 の灰色網掛け部分に該当するユーザが革新的ユーザとなる）．

本手法でのラベル付与は，各ユーザの商品購買が複数観測されている必要があり，特に商品購買サイクルが長いサイトにおいては長期的な観測が必要となる．あくまで教師データ作成用の手段として利用可能な方法である．

3.3　課題 2：行動ベクトルの生成方法

提案手法は革新性が同じであれば，アクセスする Web ページ群，すなわち行動ベクトルが類似していることを前提としている．すでに述べたように，行動ベクトルの成分は Web ページに対応しているのではなく Web ページの種別に対応している．このようにする理由は，Web ページの粒度では細かすぎて，行動ベクトルがスパースになってしまい，革新的ユーザを推定するモデルがうまく構築されない可能性が高いと考えられるからである．Web サイトに含まれる Web ページの数は，サイトの規模によって異なるものの数百から数千に及ぶため，村田 [13] にも指摘されているように，同一の Web ページを閲覧するユーザが複数存在する可能性は低い．そのため，同じ傾向を持ったアクセス行動であっても同じ Web ページにアクセスしなかったために行動ベクトルが異なり，ユーザ同士を関連付けることが困難になると予想される．

たとえば，3.1 節で述べたように，商品詳細ページへ頻繁にアクセスするかどうかは，革新的ユーザを推定するのに重要な情報と考えられるが，通常，商品詳細ページは商品ごとに異なる URL で存在することが多いので，Web ページの粒度で行動ベクトルを生成した場合，異なる 2 人のユーザが商品詳細ページにアクセスしても商品が異なることにより行動ベクトルの異なる成分としてカウントされてしまう．そこで，本研究では行動ベクトルの生成時に商品詳細，検索結果等の同じ種類の情報を含む Web ページは同種の Web ページとして，複数の Web ページへのアクセス回数を集約して行動ベクトルを生成する．行動ベクトル V_i は，下記の 2 ステップで生成する．

Step1　商品詳細ページや検索結果ページといった Web ページに含まれる内容の種類別に，Web ページを複数のグループに集約する．たとえば，ある Web サイトにおいて商品の詳細情報を示す Web ページが，商品ごとに複数存在する場合には，それらの複数の Web ページを集約して 1 つの商品詳細ページグループとして分類する．

Step2　Step1 で分類したグループ単位でアクセス回数を集約し，必要に応じて正規化等の処理を行い行動ベクトルの要素値である v_{ij} を構築する．たとえば，Web ページ群 x_j に含まれる $WebPage_a$〜$WebPage_c$ の各 Web ページに 3 回，0 回，2 回とアクセスした場合，商品詳細 Web ページ群 x_j に対する行動ベクトルの要素値 v_{ij} は，アクセス回数の和をとって算出した $a_{ij} = 3 + 0 + 2 = 5$ に必要に応じて正規化等の処理を施して算出する．

4.　提案手法の評価

4.1　実験概要

ユーザが革新的ユーザであるか否かの正解データとして，ユーザが商品詳細ページへアクセスした商品が発売日からどれだけ経ったものかの平均値をとった．そして，Rogers の定義に従いその平均値が短い上位 1σ のユーザ（発売して間もない商品の情報ばかり閲覧しているユーザ）を革新的ユーザと定義した[*1]．本検証では，本定義の革新的ユーザを提案手法を用いて推定できるか検証した．また，ベースラインとなる比較手法として，実際のマーケティング活動ではアンケートデータを用いてユーザのサイコグラフィック属性を推定していることをふまえ，アンケートデータを用いて同様に先述の定義の革新的ユーザを抽出できるか検証し，比較した．

[*1] 本来の定義に従うならば「採用 = 購入」として発売日から購入日までの時間差で測定すべきだが，実験で用いたデータは購入が観測できないため「採用 ≃ 商品情報への最初のアクセス」とした．

© 2013 Information Processing Society of Japan

表 1　URL 変換の例
Table 1　The example of URL conversion.

変換ルール	変換前	変換後
ID の集約	http://example.com/item/002D314A1/itemdetail.html http://example.com/item/D31D3S155/itemdetail.html	http://example.com/item/<id>/itemdetail.html
キーワードの集約	http://example.com/search/keyword=%52%8E%82%89%82··· http://example.com/search/keyword=%82%8D%82%89%82···	http://example.com/search/keyword=<kwd>
送信データの削除	http://example.com/feature.html?ref%5F=pe%5F4322%···	http://example.com/feature.html

4.2　実験データ

株式会社ビデオリサーチインタラクティブが収集した Web 視聴データと特性調査データを用いた[*2]．Web 視聴データは，収集対象ユーザが Web ブラウザを用いて閲覧した Web ページの URL，ユーザ ID．閲覧日時，閲覧時間等の情報が記録されたデータである．また，特性調査データは，Web 視聴データの収集対象ユーザに対して 2006 年 11 月にアンケートを実施して取得したデータである．本検証では，Web 視聴データに含まれるデータ項目のうち，Web アクセスログに含まれる可能性の高い “Web ページの URL” と “ユーザ ID”，“閲覧日時” の項を用いて検証を行った．データの収集期間は 2007 年 1 月 1 日から 2007 年 12 月 31 日までの 1 年間，観測対象ユーザ数は月間のアクティブユーザ数が約 7,000 人となるよう調整された，延べユーザ数 11,424 人，総ログ数 191,967,395 レコードとなっている．

今回の検証では，本研究の適用対象と考えている商品購入サイトと情報収集サイトの中から，サイトが提供する機能が一般的で，Web ページの内容と URL が 1 対 1 に対応しており，取扱商品の多様性が高いサイトの中で，アクセスユーザ数が上位 1 位のサイトを各 1 サイトずつ，合計 2 つのサイト（以降，「サイト A」「サイト B」と呼ぶ）のユーザを対象に検証を行った．

4.3　革新性ラベルの付与

革新性ラベルの作成にあたり，各ユーザの商品詳細ページへアクセスしたログから，商品の発売日を抽出し，該当商品への初回アクセス日との差を算出する．発売日は該当商品詳細ページに記載の発売日情報をクローラを用いて取得した．発売日が取得できた数は，サイト A は 49,259/129,832 ページ（40%），サイト B は 107,831/162,099 ページ（67%）であった．算出した差（経過日数）の平均を用い，ユーザの革新性を定義する．革新性の算出の対象となるユーザは，平均値に意味のある最低数として，各サイトで発売情報が取得できた商品を 10 商品以上アクセスしているユーザに限定した．

各サイトごとに経過日数の分布を求め，採用時期平均が全体の -1σ より短いユーザを革新的ユーザとしてラベル付けした（図 2 参照）．なお，サイト A については，後述の Web ページ集約方法 2 において DVD カテゴリのみのルールを作成したことに対応するため，全カテゴリ横断的に算出したものと，DVD カテゴリに限定して算出したものの 2 パターンを用意した．

4.4　行動ベクトルの作成

3.3 節で述べたとおり，Web ページを内容の種別ごとに集約する必要がある．本実験では以下に示す 2 つの方法で Web ページを集約してアクセス回数をカウントしたものを行動ベクトルとした．

4.4.1　集約方法 1：引数変換ルールによる Web ページの集約

多くの EC サイトでは，商品 ID，検索キーワード，ユーザ ID 等が URL に含まれる場合，その ID やキーワードが異なってもページの種別は同一である．したがって，URL 中にそれらの ID やキーワードが含まれる場合，それを削除（もしくは共通の文字列に置き換え）することにより URL を集約する．**表 1** にその例を示す．

ここで，商品 ID やユーザ ID はサイトごとに付与ルールが異なるため，変換ルールはサイトごとに調整する必要がある．多くの場合は，連続する数字列，URL エンコードされた文字列，および “?” 以降の文字列（CGI 引数）を削除するというルールで対応可能である．

4.4.2　集約方法 2：対応ルールベースによる Web ページの集約

本方法は EC サイトごとに人手により，ページの種別-URL の対応ルールを作成し，正規表現でマッチすればその種別の行動をしたとしてカウントするものである．**表 2** に対応表の一部を示す．表 2 のルール 1，ルール 2，... に記載の正規表現に一致する URL のページ群が集約され，それぞれが行動ベクトルの一次元の要素となる．したがって，次元数は作成したルールの種類数となる．方法 2 は方法 1 に比べ，ページ種別が同じであるにもかかわらず URL がまったく異なるページ群を集約できる分，より少ない次元に Web ページを集約できるという利点があるが，対応ルールベースの作成に多くの人手が必要となる．

[*2] http://www.videoi.co.jp/service/webpac/

© 2013 Information Processing Society of Japan

表 4　実験対象サイトごとのデータセット概要
Table 4　Summary of experimental data.

		サイト A（全カテゴリ）	サイト A（DVD のみ）	サイト B
サイト種別		商品購入サイト		情報収集サイト
ユーザ数	革新的ユーザ [人]	107（訓練用 74, テスト用 33）	22（訓練用 15, テスト用 7）	40（訓練用 28, テスト用 12）
	非革新的ユーザ [人]	789（訓練用 74, テスト用 33）	147（訓練用 15, テスト用 7）	406（訓練用 28, テスト用 12）
期間中対象ユーザの平均アクセス数 [PV]		545	856	570
集約方法 1 の要素数 [種類]		271	-	402
集約方法 2 の要素数 [種類]		-	27	23
経過日数平均（対数）		2.5	2.1	2.3
経過日数標準偏差（対数）		0.7	1.1	0.3
経過日数 -1σ（対数）		1.8	1.0	2.0
経過日数 -0.5σ（対数）		2.1	1.5	2.1

表 2　行動-URL 対応表の例
Table 2　The example of URL-pagetype conversion table.

行動 ID	行動	URL（正規表現）
ルール 1	おすすめ商品ページへのアクセス	.*/gp/yourstore/.* .*/([A-Z0-9])/.*ref=.*_all_.*
ルール 2	商品詳細ページへのアクセス	/exec/obidos/.* .*/gp/([A-Z0-9]10)/.*

表 3　革新的ユーザの推定に用いたアンケート内容
Table 3　Questionnaire for evaluating user's innovativeness.

問：商品購入時の意識について，1（非常に項目 A に近い）～5（非常に項目 B に近い）の 5 段階で回答して下さい．

	項目 A		項目 B
Q1	流行のもの	1 2 3 4 5	自分の考え
Q2	新商品に興味	1 2 3 4 5	新商品には無関心

4.4.3　ベースライン：アンケートの回答によるベクトル生成

Web ページへのアクセスを行動ベクトル化したものを用いた推定との比較対象として，アンケートの回答をベクトル化したものを用いた革新性推定精度を比較対象とした．アンケートの回答が正しければ，革新的なユーザの特徴と一致する回答を行っているユーザは商品採用時期が早く，購買商品の発売日からの平均経過時間も短いはずである．

アンケートのベクトル化は，各設問をそれぞれ一次元とし，回答の数値（1～5 の 5 段階で回答）をそのまま各次元の強さとして利用して行っている．設問は，Rogers が示した，イノベーション採用時期の違いによって異なる行動の特徴に関連する設問のデータを，特性調査データからピックップして用いた．使用した設問を**表 3** に示す．設問内容はビデオリサーチインタラクティブ社の作成による．

4.5　推定モデルの評価

4.3 節で導出した革新性ラベルの正解データと，4.4 節で算出した行動ベクトルデータをデータセットとして，提案手法である SVM による革新的ユーザの推定モデルを評価した．検証はデータセットのうち 70%を訓練用データ，残りをテスト用データとして分割し，訓練用データで学習したモデルのテスト用データの推定精度で評価した．なお，非革新的ユーザの人数が革新的ユーザの人数に対してかなり多いことから，非革新的ユーザのデータを革新的ユーザのデータ数と同数に調整するリサンプリング法を用いている．また，1 人しかアクセスしていない行動ベクトルは，オーバフィッティングを招く可能性が高いため，行動ベクトルデータから除外している．

各サイトのデータセットの概要を**表 4** に示す[*3]．

SVM の実装は統計解析用の言語・環境である R[*4]のライブラリ kernlab [14] を用い，モデルは C-SVM（ソフトマージン SVM），カーネルは Radial Basis kernel function（ガウシアンカーネル）を指定して使用した．各ベクトル要素の値はユーザごとのそのサイトへのアクセス総数に占める割合で正規化して使用している．

評価指標には，検索分野でよく利用される Accuracy（正確度），Precision（精度），Recall（再現率），F-measure（F 値）を用いた．各指標は以下の式で計算される．

$$Accuracy = \frac{(\text{正確に判別したラベル数})}{(\text{テストしたラベル総数})} \tag{2}$$

$$Precision = \frac{(\text{“革新的ユーザ” と正解した数})}{(\text{“革新的ユーザ” と判定した数})} \tag{3}$$

$$Recall = \frac{(\text{“革新的ユーザ” と正解した数})}{(\text{実際に “革新的ユーザ” の数})} \tag{4}$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

4.6　結果

4.6.1　採用時期算出結果

ユーザごとに求めた，商品の発売日から該当商品への初回アクセス日への経過日数平均の分布を**図 3** に示す．横

[*3] サイト A については，集約ルール作成の手間の関係で，集約方法 1 は全カテゴリ，集約方法 2 は DVD カテゴリのみを対象としてデータセットを作成している．
[*4] http://www.r-project.org/

© 2013 Information Processing Society of Japan

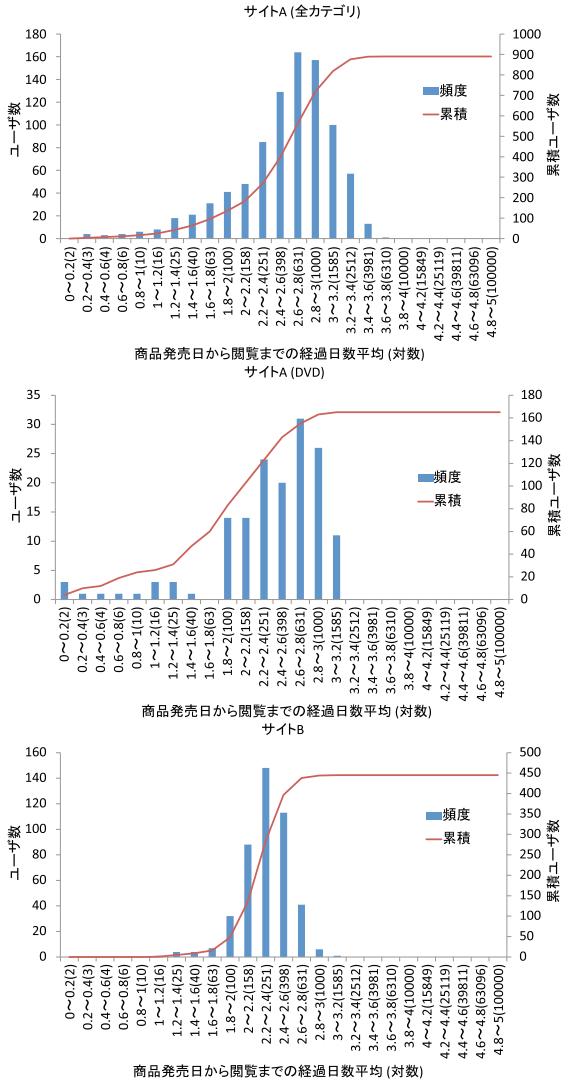

図 3　革新性分布

Fig. 3　Histgram of innovation characterstics.

軸はユーザごとの商品発売開始から閲覧までの経過日数の平均を対数変換したもの，縦軸は各経過日数平均ごとのユーザ数である．分布は正規分布に近い形状をしており，Rogers のイノベータ理論で示されている採用時期ごとのユーザ数分布と一致していることが分かる．

ただし，サイト A（DVD カテゴリ）については，ユーザ数の母数が少ないため度数分布にバラつきが大きく，特に平均経過日数の短いユーザが多い傾向にあるように見える．今後の課題として，より母数の大きなデータを用いた検証が必要と考える．

対象サイト別に革新的ユーザと非革新的ユーザのアクセス人数上位商品のランキングを求めたものを**表 5** に示す．どちらのサイトでも発売年に大きな差はないものの，サイト A では革新的ユーザが発売前のゲームソフトの情報や，人気が出る前のアイドルグループの CD，DVD へのアクセスが多いのに対し，非革新的ユーザはすでに人気の確立したシリーズものの後継作ゲームソフトや書籍へのアクセスが多いという差が見られる．また，サイト B でも革新的ユーザは携帯電話機（半年ごとに新機種が発売され変化が早い）が多い傾向にあるのに対し，非革新的ユーザは電子辞書や FAX 機といったすでに機能が確定した変化の少ない商品へのアクセスが多い傾向となっている．

以上をまとめると，商品の採用時期によるユーザ分類の結果，(1) Rogers の分析と一致する分布が得られていること，(2) 各ユーザ群の選択している商品の革新性が異なっていることが分かった．したがって，経過日数平均が早いか遅いかによりユーザ分類することで，革新的ユーザのラベル付与ができていることが分かる．

4.6.2　革新性推定結果

革新的ユーザとしてラベル付与する対象を，平均経過日数が -1σ 以下のユーザ，および -0.5σ 以下のユーザに広げた場合の 2 パターンについての評価結果を**表 6** に示す．

表 6 から，集約方法や対象とするサイトごとに推定精度の値が大きく異なり，ルールベースで Web ページを集約する集約方法 2 がすべての指標値において，URL を引数変換のみで集約する集約方法 1，およびアンケートからの推定を上回る値を示しており，集約方法 2 により Web ページ集約して推定モデルを構築する方法が有効であることが分かった．

また，革新的ユーザとして判定する平均経過日数の閾値を -1σ から -0.5σ に広げることにより，すべての方法において推定精度が低下することが分かった．この要因として，範囲を広げることで革新的ユーザでない行動特性のユーザまで革新的ユーザとして含まれてしまうことになり，結果として革新的/非革新的ユーザ間の行動特性の差が小さくしてしまったためであると考えられる．したがって，閾値として Rogers の提唱する -1σ を設定することが有効であるといえる．

5.　考察

結果で述べたように，集約方法によって推定精度に違いがあり，URL を引数変換のみで集約する方法 1 はルールベースで集約する方法 2 より全般的に精度が低い．実際に方法 1 で作成したベクトルデータを見ると特定ユーザのみが 1 となるベクトルが多く，行動特性をとらえるには集約が不十分であったと考えられる．

また，ルールベースにより集約する方法 2 はすべてのサイトにおいてベースラインを上回る精度を示しているが，サイトごとに推定の特性に差がある．このことは，集約ルールが推定精度に与える影響が高いことを示唆しており，推定精度を高めるための効果的なルールベース作成方法の定式化を検討する必要があるといえる．

一方で，アンケートによるベースラインの推定精度はランダム推定と変わらない値を示しており，対象商品群を特定しない設問において判断した革新性からは，ユーザの商

© 2013 Information Processing Society of Japan

表 5　アクセス人数上位商品
Table 5　Access ranking.

サイト A

	革新的ユーザ		非革新的ユーザ	
順位	商品概要（発売年）	人数	商品概要（発売年）	人数
1	ゲーム機本体（2006 年）	9	エステ関連書籍（2006 年）	9
2	ゲームソフト（2008 年発売予定）	8	シリーズ 4 作目ゲームソフト（2007 年）	9
3	女性シンガーの音楽 CD（2007 年）	7	携帯型音楽再生機（2007 年）	9
4	男性アイドルのライブ DVD（2007 年）	7	シリーズ 4 作目ゲームソフト（2007 年）	9
5	男性アイドルの音楽 CD（2007 年）	7	シリーズ 8 作目ゲームソフト（2007 年）	9
6	男性バンドの音楽 CD（2007 年）	7	シリーズ最終作ベストセラー書籍（2007 年）	9
7	女性シンガーの音楽 CD（2007 年）	6	シリーズ 9 作目ゲームソフト（2007 年）	9
8	男性アイドルのライブ DVD（2007 年）	6	音声合成ソフト（2007 年）	9
9	男性アイドルのライブ DVD（2007 年）	6	ベストセラー書籍第 6 巻上下巻セット（2006 年）	9
10	自己啓発本（2007 年）	6	シリーズ 3 作目ゲームソフト（2006 年）	9

サイト B

	革新的ユーザ		非革新的ユーザ	
順位	商品概要（発売年）	人数	商品概要（発売年）	人数
1	携帯電話機（2007 年）	6	除湿機（2007 年）	9
2	ゲーム機本体（2006 年）	6	コンパクトデジタルカメラ（2007 年）	9
3	携帯電話機（2007 年）	5	DVD ドライブ（2006 年）	9
4	ゲーム機本体（2007 年）	5	電子辞書（2006 年）	9
5	ポータブルナビゲーション（2007 年）	5	アンチウィルスソフト（2006 年）	9
6	携帯電話機（2007 年）	5	ハードディスク（2007 年）	9
7	コンパクトデジタルカメラ（2007 年）	4	イヤホン（2006 年）	9
8	携帯電話機（2007 年）	4	空気清浄機（2006 年）	9
9	携帯電話機（2007 年）	4	プリンタ（2006 年）	9
10	コンパクトデジタルカメラ（2007 年）	4	FAX 機（2007 年）	9

表 6　各種法の推定精度の評価結果
Table 6　The evaluation result of various methods.

		平均経過日数 $< -1\sigma$ を革新的ユーザに定義				平均経過日数 $< -0.5\sigma$ を革新的ユーザに定義			
行動ベクトルの作成方法と評価サイト		Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
方法 1	Site A (All)	0.56	0.56	0.55	0.55	0.47	0.47	0.36	0.41
	Site B	0.58	0.58	0.58	0.58	0.57	0.56	0.63	0.59
方法 2	Site A (DVD)	0.71	0.64	1.00	0.78	0.56	0.53	1.00	0.69
	Site B	0.88	0.91	0.83	0.87	0.62	0.61	0.63	0.62
アンケート	Site A (All)	0.46	0.46	0.42	0.44	0.46	0.45	0.42	0.44
	Site B	0.59	0.58	0.64	0.61	0.43	0.43	0.39	0.41

品採用時期を正しく推定できないことを示しており，3.2 節の仮説を裏付ける結果が得られたと考えられる．

先述のとおり，本実験で試したいずれの Web ページ集約手法もデータがスパースになりがちで，ほとんどのユーザにおいて $v_{ij} = 0$ となる要素が多数存在するため，ページの種別ごとの集約ではユーザの行動特性を表す指標値としてはまだ適当でないと考えられ，より行動特性に即した集約手法が求められる．たとえば，Moe は，ユーザの EC での行動タイプ分けを行うのに有効な尺度としてセッションに占める行動の割合，カテゴリやブランドのバラエティ等に着目した 14 尺度を提案しており [15]，本尺度をベクトル要素とする方が精度が向上する可能性がある．今後の課題として検証していきたい．

6. まとめ

本稿では，アンケートのようなユーザの明示的な入力に頼らず革新的ユーザを抽出する手法を提案した．本手法では，EC サイトで自動的に記録される購買データとアクセスログを用いて，商品発売からユーザが商品購入を行うまでの経過時間の平均を革新性推定の教師データとする．そして，Web ページへのアクセス回数を行動ベクトル化したものを用いて機械学習を行い，革新性を推定する手法を提

© 2013 Information Processing Society of Japan

案した．

商品購入に関する 2 つの Web サイトを対象にして検証を行った結果，提案手法はベースラインであるアンケートを用いた推定を上回る精度を導出する効果があることを確認した．本提案方式は，取得が容易な Web アクセスログを用いて，簡単な集約処理で革新的ユーザを抽出することができる点に特徴を持つものである．本方式により簡単に大規模な革新的ユーザの行動観察が実現できる．

今後は，より推定に有効なベクトル要素となる行動尺度を探ることで精度向上を図るとともに，多くの商品購入サイトに適用し，提案手法の一般性に関して継続的に検証していく予定である．

参考文献

[1] 総務省情報通信国際戦略局：平成 22 年通信利用動向調査の結果, 総務省（オンライン）, 入手先 ⟨http://www.soumu.go.jp/johotsusintokei/statistics/statistics05.html⟩（参照 2012-02-13）.

[2] 経済産業省商務情報政策局：平成 22 年度電子商取引に関する市場調査, 経済産業省（オンライン）, 入手先 ⟨http://www.meti.go.jp/policy/it_policy/statistics/outlook/ie_outlook.htm⟩（参照 2012-02-13）.

[3] Ichikawa, Y., Nakamura, M., Hataand, K. and Nakagawa, T.: Provision of Services According to Individual User Preferencesover a Cross-section of Sites Implemented with “Personalized-servicePlatform”, *NTT Technical Review*, Vol.6, No.4 (2008).

[4] 市川裕介, 小林　透：大量の行動履歴情報を扱うプラットフォーム技術，情報処理，Vol.51, No.1, pp.18–21 (2010).

[5] Kotler, P. and Armstrong, G.: *Marketing: An Introduction, 4th Ed.*, Prentice-Hall (1996). 恩蔵直人（監訳），月谷真紀（訳）：コトラーのマーケティング入門第 4 版，ピアソン・エデュケーション (2000).

[6] 平久保仲人：消費者行動論，ダイヤモンド社 (2005).

[7] Rogers, E.: *Diffusion of Innovations, 5th Edition*, Free Press (2003). 三藤利雄（訳）：イノベーションの普及，翔泳社 (2007).

[8] 斎藤　隆：食品市場の創造—ヒット商品開発装置 Japan-VALS の提案，東急エージェンシー (1994).

[9] Song, X., Tseng, B.L., Lin, C.-Y. and Sun, M.-T.: Personalized recommendation driven by information flow, *Proc. 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’06*, pp.509–516, New York, NY, USA, ACM (2006).

[10] Kawamae, N., Sakano, H. and Yamada, T.: Personalized recommendation based on the personal innovator degree, *Proc. 3rd ACM conference on Recommender systems, RecSys ’09*, pp.329–332, New York, NY, USA, ACM (2009).

[11] 市川裕介, 小林　透：ユーザの Web アクセス履歴のべき乗分布傾向に着目した属性推定手法の提案，情報処理学会論文誌，Vol.52, No.3, pp.1195–1203（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110008507955/⟩ (2011).

[12] Vapnik, V.N.: *Statistical Learning Theory*, Wiley-Interscience (1998).

[13] 村田剛志：Web 視聴率データからのユーザコミュニティの発見，知能と情報：日本知能情報ファジィ学会誌，Vol.18, No.2, pp.213–222（オンライン），DOI: 10.3156/jsoft.18.213 (2006).

[14] Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A.: kernlab – An S4 Package for Kernel Methods in R, *Journal of Statistical Software*, Vol.11, No.9, pp.1–20 (online), available from ⟨http://www.jstatsoft.org/v11/i09/⟩ (2004).

[15] Moe, W.W.: Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream, *Journal of Consumer Psychology*, Vol.13, No.1, pp.29–39 (2003).

市川 裕介　（正会員）

1994 年慶應義塾大学理工学部計測工学科卒業．1996 年同大学大学院修士課程修了．同年日本電信電話株式会社入社．以来，通信履歴活用サービスの研究開発に従事．情報処理学会山下記念研究賞（2005 年）受賞．

岸本 康成　（正会員）

1989 年九州大学理学部物理学科卒業．1991 年同大学大学院総合理工学研究科修士課程修了．同年日本電信電話株式会社入社．以来，ディレクトリ・システム，課金システム，データマイニング等に関する研究開発に従事．現在，NTT ソフトウェアイノベーションセンタ研究主任．

小林 透　（正会員）

1985 年東北大学工学部精密機械工学科卒業．1987 年同大学大学院工学研究科修士課程修了．同年日本電信電話株式会社入社．以来，ソフトウェア生産技術，情報セキュリティ，データマイニング，次世代 Web 技術等の研究開発に従事．2013 年から長崎大学大学院工学研究科教授．IEEE，電子情報通信学会各会員，博士（工学）．

© 2013 Information Processing Society of Japan