行動データ分析のためのノイズ除去手法に関する研究

２０２２年　３月

田中　孝昌

行動データ分析のためのノイズ除去手法に関する研究

田中　孝昌

システム情報工学研究科

筑波大学

２０２２年　３月

目次

図目次

表目次

第1章　緒論

近年，AIの社会への導入が進んでいる．我が国の内閣府が，2018年に社会の成長戦略として作成したSociety5.0[1]では，様々な産業において，人がデータを分析し，提案する社会から，AIがデータを分析し，AIが人に提案する社会に進化する，と述べられている．例えば，交通分野では，自動車に取り付けられたセンサー情報，天気，交通，ホテル，飲食店といったデータをAIが分析し，AIが好みに合わせた計画を人に提案する，と述べられている．AIに，そのような自動の提案を可能にさせている要素には，膨大なデータと情報推薦アルゴリズムが挙げられる．情報推薦アルゴリズムは，Kaggleに代表されるような世界的なコンペティションが行われており，非常に活発な研究分野である[2]．その一方で，情報推薦アルゴリズムへのインプットデータとなる膨大なデータを効率的に管理することも同じく重要な研究分野である[3]．

今日のようなウェブサイトの自動推薦機能の普及以前から，ウェブサイトの運営において，データを管理することは非常に重要な作業であった．ユーザーに対して商品などのコンテンツの検索機能を提供するウェブサイトの場合，ユーザーの検索クエリによって大量の商品の中から希望の商品を引き当てることができるように，商品の特徴を表すデータを商品の識別子と紐付けて管理しなければならない[4]．具体的な商品として，例えば不動産物件を取り扱う場合は，所在地，賃料や販売価格，部屋の面積，最寄り駅からの距離といったデータが，それに該当

する. このような商品の情報を管理することは, ウェブサイト上でユーザーが商品を探す, という最低限の機能を提供するために必須の内容である.

だが, 近年では, ウェブサイトの最低限の機能というよりも, AI を活用した商品の自動推薦機能といった, ウェブサイトの魅力を飛躍的に高める機能を提供するために, 膨大なデータを管理する重要性が増している[5]. そして, ウェブサイトのユーザーの行動履歴を記録した膨大なアクセスログは, そのような管理するべきデータの中心的な役割を果たしている[6]. 一般にアクセスログとは, ユーザーの情報端末とウェブサイトを構成するサーバの間で交換された情報の履歴データである. 具体的には, 通信パケットに含まれる HTTP メッセージに記述したヘッダ情報などを記録したデータのことを指す[7]. ただし, 今日のウェブサイトの運営におけるアクセスログの位置づけは, ウェブサイトが稼働した結果として偶然に生まれる副産物ではない. 今日のアクセスログの開発担当者は, 例えば商品推薦システムへのインプットデータの開発という明確な目的のために, ユーザーの購買や商品の比較検討といった様々な行動を正確に捕捉するための識別子を設計して, ユーザーのウェブサイトへの訪問から離脱までの行動を漏れなく記録する機能を作り上げている. いわば, 今日の企業においてアクセスログは, 目的達成のために緻密に設計した主産物として扱われている.

他方で, アクセスログに代表される行動データの管理の重要性は, ウェブサイトだけに留まらず, 物理的な実店舗を構えるスーパーマーケットのような業態でも, 同様に高まっている. 今日のスーパーマーケットには, 顧客の買い物体験を便利にするための専用の情報端末を設置している店舗が存在しており, 様々な行動データが収集されている. 行動データを分析することで, 様々な情報端末に割引クーポンの配信が行われていたり, 限られた優良顧客に向けて特別なダイレクトメッセージを郵送されていたりと, 様々なマーケティング活動が行われている. 業界団体の調査によれば, 2020 年時点で 8 割を超える店舗にポイン

トカードが導入されており，我が国のスーパーマーケット業界に関して言えば，すでに多くの店舗において顧客の行動データの収集とそれを活用したマーケティング活動が行われている，と言える[8]．このような企業の行動データを活用したマーケティング活動は，企業経営の補助的な役割を超えて，企業経営の根幹を成す活動となってきている．

しかし，行動データを分析することの企業経営における重要性が高まる一方で，行動データの分析の精度を落とし，目的達成の大きな障害となるような，行動データのノイズ混入が重大な問題として発生している．

例えば，住宅情報ポータルサイトにおいては，大量のインターネットボットのデータがアクセスログに混入することで，集客効果の分析精度が大きく損なわれている問題が発生している．住宅情報ポータルサイトは，住宅検討者と住宅情報のマッチングを多く生み出すために，テレビ CM のような幅広い層にイメージを訴求するマス広告や，リスティング広告のように住宅情報の検索を行った限られた層を直接集客するターゲッティング広告など，性質の異なる広告を使い分けながら集客を行っている[9]．このような集客活動は，住宅情報ポータルサイトの生命線となっている．もし，サイトを訪問したユーザーが，大量のボットと区別がつかなければ，現在の集客効果を正しく把握できないため，ボットによる行動データのノイズは，住宅情報ポータルサイトにとって経営の根幹を揺るがす問題に直結してしまう．だから，住宅情報ポータルサイトには，このようなボットを行動データからノイズとして除外する方法が必要になる．また，例示したような行動データのノイズの問題は，企業が行動データ活用を試みると，企業の外部環境，内部環境，両方から様々な形で発生してくる．そのように様々な要因から発生する，すべてのノイズの問題を単一のノイズ除去手法で解決することは不可能である．

そこで，本研究では，企業との連携によって，3 つの行動データのノイズの問題を抽出し，その問題を従来手法よりも効率的に解決する新たなノイズ除去手法の提案を行う．

1つ目の問題は，上述の住宅情報ポータルサイトに訪れるインターネットボットを行動データからノイズとして除去する問題を取り扱う．住宅情報ポータルサイトのシステムから，複数の技術方式を用いて異なる行動データを取得することで，単一の行動データからは抽出できないボットの知識を獲得する．このような手法を用いない限りは，ボット判別は，ボットのアクセスの量をもとに，異常を判別せざるを得ない．だが，提案手法では，獲得知識を用いた行動パターン，属性に関する知識を活用することで，アクセスの量だけではなく質も合わせて評価が可能となっている．

2つ目の問題は，同じく住宅情報ポータルサイトの行動データに混入する不正識別子をノイズとして除去する問題を取り扱う．住宅情報ポータルサイトは，ユーザーに使いやすい機能を実現するために日常的にサイトの変更作業を行っている．高頻度の変更作業の中で，間違った形式の不正識別子が行動データに記録されるような，誤った変更をしてしまう問題がある．この問題の対して有効な既存手法は，ウェブサイトの自動テストの既存手法が有効だが， この問題に適用するには自動テストのためのコード生成の作業が大きな負担であった．本研究は，自動テストのためのコードを自動生成する新たな手法を提案する．

3つ目の問題は，スーパーマーケットチェーンの行動データにおける大量の一般顧客から少数の優良顧客を抽出する問題を取り扱う．複数店舗を運営するスーパーマーケットチェーンでは，少数の優良顧客を自動で抽出して，個別のマーケティング活動を行う必要があった．この問題に有効な既存手法では，抽出精度と抽出後の顧客に関する知識獲得に課題があった．本研究は，既存手法を拡張手法として，スーパーマーケットの店舗間の異質性を考慮して，優良顧客抽出と知識獲得を実現する新たな手法を提案する．

尚，本研究で扱う行動データのノイズの問題は，連携企業から抽出するが，特定企業の固有の問題ではなく，性質の近しい企業や業界で発生する普遍的な問題に焦点を絞る．提案手法についても，連携企

業の業界におけるウェブサイトのシステム構成や開発方法，商品バラエティといった普遍的な構造を活用しながら，類似の業種に対して汎用的に適応可能な手法として設計している．また，提案手法の一部は，企業で実際に使われている情報システムの上で実証実験行うことで有用性を確認し，情報システムへの導入を完了している．このことは，本研究の提案する新たなノイズ除去手法の信頼性を支持している．

本論文は 6 章で構成される．2 章では，まず企業から抽出した行動データの活用におけるデータのノイズの問題を説明し，その問題に対して有用な関連研究について述べる．3 章では，住宅情報ポータルサイトのインターネットボットの問題に対応した新たなノイズ除去手法を提案する．ウェブサイトから異なる技術方式を用いて複数の行動データを収集，分析することで，単一の行動データからは知り得ないボットの知識を獲得し，その知識を用いて既存手法の精度を超える自動判別の評価実験の結果を示す．4 章では，住宅情報ポータルサイトの継続的な変更開発によって発生する行動データのノイズとしての不正識別子の問題を説明し，行動データの不正識別子を検知するテストコードを自動生成する新たなアプリケーションを用いて，既存手法よりも効率的にテストケースを完了させる評価実験の結果を示す．5 章では，複数店舗を運営するスーパーマーケットチェーンにおいて大量の一般顧客の行動データがノイズとなって少数の優良顧客が困難となる問題を説明する．店舗間異質性を利用した提案手法が，既存の手法よりも判別精度と優良顧客に関する知識獲得において，優れているという評価実験を示す．最終の 6 章では結論を述べる．

第2章　行動データ分析のノイズの問題と対策

本研究の目的は，企業の行動データ分析のノイズの問題に対して，新たなノイズ除去手法を提案することにある．本章では，本研究の中で確認した，企業の直面するノイズの問題について説明し，関連研究を交えて本研究で使用するノイズ除去手法について述べる．

2.1　住宅情報ポータルサイトにおけるインターネットボットの問題

住宅情報ポータルサイトをはじめ，今日のウェブサイトにはユーザーからだけではなく，プログラムによって大量の処理をウェブサイトに要求するボットからのアクセスが多く発生している．そして，ボットの一部は，ウェブサイトの運営者にアクセスを遮断されることを避けるために，様々な情報を偽装しているため検知が難しい．

ボットのアクセスログは，過剰なノイズとして分析の精度を下げるリスクがある．例えば，ウェブサイト上で商品が確認されたあとに購入される遷移率を分析する場合を考える．ボットは，商品を紹介するページを回遊するログは残すが，実際にフォームに情報を入力して購入することはないため，遷移率が実際よりも低く計算されてしまう．また，ボットからのアクセスが特定の商品に集中した場合は，見られても購入には至らないため，何らか問題のある商品として分析結果が出てしまうかもしれない．他にも，商品のレコメンデーションシステムにおいては，アクセスログ上で同一ユーザーから合わせて検討されている商品を類似の商品として学習するアルゴリズムがよく利用される[10]．だが，ボットのログからそのような学習をしてしまうと，実際のユーザーの嗜好に合わないレコメンデーションが行われてしまうリスクがある．

本節では，このような行動データとしてのアクセスログの利活用，提案手法と関連の強いアクセスログの抽出技法，アクセスログに潜むボット検知方法を，関連研究を交えて説明する．

2.1.1 アクセスログの利活用

ウェブサイトにおけるアクセスログの利活用は，2 つの用途に整理できる．

1 つ目の用途は，ウェブサイトの現在の情報の正確な把握である[11]．例えば，多くのウェブサイトは異なる機能を持つ複数の画面で構成されている．不動産物件を探すためのウェブサイトであれば，地図上から物件を検索する画面，賃料や面積で更に検索結果を絞り込む画面，販売店に問い合わせを行う画面などがある．それらの画面に対して，ユーザーからのアクセスがどれくらいの量で発生しているか，機能を利用したユーザーがどれくらいの割合で実際に希望の物件を見つけられているか，といった情報は，ウェブサイトの魅力を現在よりも更に高めていく開発計画を立てるための貴重な材料である．

2 つ目の用途は，ウェブサイトの未来等の不確定な情報の予測，それによるより高度な機能開発である．例えば，ウェブサイトの扱う商品の購買傾向に季節性や周期性がある場合，それに従ってユーザーのウェブサイトへのアクセス量は変動する．ウェブサイトのアクセス量が増えるならば，予めウェブサイトを構成するサーバーを増設しておかないと，サーバーの計算リソースの過負荷により機能を提供できなくなるリスクがある．だが，ユーザーの過去のアクセス量と時系列モデル[12]を用いて，季節性や周期性を学習し，未来のアクセス量を予測することができれば，予めサーバーを適切な規模に増設することで，そのようなリスクに対処できる．また，もうひとつ例を挙げる．既に EC サイト等で一般的な機能である，ユーザーそれぞれに異なる商品をウェブサイト上で表出するレコメンデーションシステム[13]は，アクセスログからユーザーの嗜好を

予測することでその機能を実現している．より高い予測精度を実現すれば，それだけユーザーの検索の手間を減らすことができる．だから，今日のウェブサイトの運営では，レコメンデーションシステムへの投資が積極的に行われている．

2.1.2　　アクセスログの抽出技法

これらの用途のために，実際にアクセスログを活用するには，用途に合わせてデータを望ましい状態に維持する，データ管理の活動が必要である．データ管理の活動の体系的な整理は数多く存在するが[14][15][16]，アクセスログを含むビッグデータの演算処理は，ETL フレームワークを用いて整理することができる[17][18]．ETL フレームワークとは，データを抽出し，変換し，活用のためのデータベースやファイルシステムに装填する，という 3 段階にデータ管理の活動を整理している．アクセスログのデータ管理には多くの課題が存在するが，本研究では，ETL フレームワークの Extraction の部分としてアクセスログの抽出，抜粋の課題に焦点を当てる．

アクセスログを分析して，ウェブサイトの現在の情報の把握や，未来の情報やユーザーの嗜好など不確定な情報の予測を行う場合，アクセスログをできるだけ過不足なく抽出することは，分析の精度にとって非常に重要である．ここでは，本研究が提案するアクセスログを過不足なく抽出する手法を説明するために，アクセスログを抽出するための基礎となる技術方式を簡単に説明する．

アクセスログを抽出する技術方式から代表的な方式[19]を，図 2-1 に抽象化した構造を示しながら説明する．

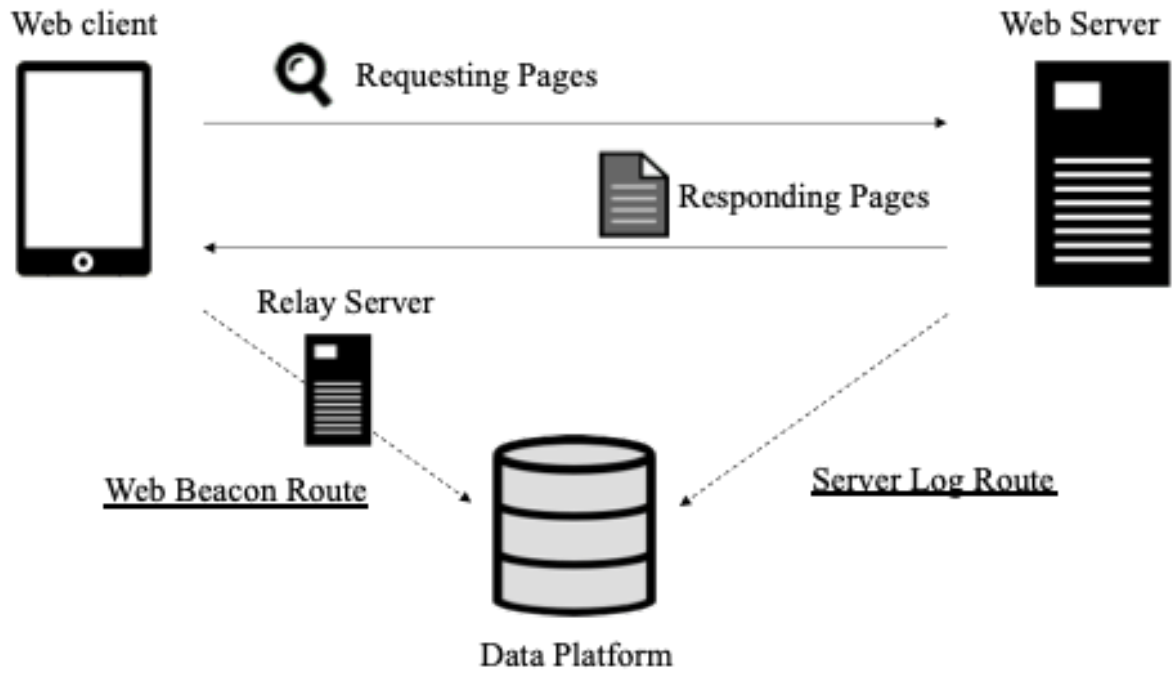


図 2-1: アクセスログ抽出のシステム構成図

1 つはユーザーのウェブクライアントに読み込んだページから抽出する方式，そしてもう 1 つはウェブクライアントからアクセスしているウェブサイトを構成するウェブサーバーから抽出する方式である．前者の技術方式は，ウェブビーコン型と呼称することが多い[20]．実現方法の例を説明する．ウェブページを構成する HTML コードの中に，アクセスログの 1 レコードに相当するウェブビーコンを送信するための Javascript コードを埋め込み，ユーザーの行動に対して，任意の行動対象およびタイミングで Javascript コードを発火させて，ウェブビーコンを送信する．ウェブビーコンは，セキュリティを確保するための中継サーバーを経て，データ分析のためのシステムであるデータ基盤に送信する．一方，後者の技術方式は，サーバーログ型と呼称する．こちらも実現方法の例を説明する．ユーザーは，スマートフォンなどの情報端末にインストールしたウェブブラウザを用いて，ウェブサイトを構成するサーバーからウェブページの情報を取得する．このときにサーバーがユーザーとやり取りをした履歴をサーバーにログとして残す．そして，任意のタイミングで，デ

ータ分析のためのデータ基盤に移送する．この 2 つの方式のメリットとデメリットを表 2-1 に整理する．

表 2-1: アクセスログ抽出方法の長所と短所

Access Log Extraction Method	Pros	Cons
Web Beacon	Including any actions	Frequent data loss
Server Log	No data loss	Including only page loading actions

ウェブビーコン型のメリットのひとつは，任意の行動とタイミングに対してアクセスログを取得できることである．例えば，1 ページの文字数が多く，何度かスクロールしてすべての内容が読めるようなページの場合を考える．その場合は，ウェブビーコン型を用いて，ページの中間と最後までスクロールしたときにそれぞれアクセスログを送信することで，中間まで読んでくれたユーザー，最後まで読んでくれたユーザーを区別して数えることが可能になる．一方，サーバーログ型は，同一ページのスクロールのようなサーバーへのアクセスを伴わないユーザーの行動は取得することができない．だから，ページを読み込んだあとでのユーザーの細かな動きを分析する場合は，ウェブビーコン型が向いている．だが，ウェブビーコン型には，通信パケットの欠損によるデータの消失が起きやすい，というデメリットがある．ウェブビーコン型では，ウェブブラウザから情報を，いわば一方的に中継サーバー，データ基盤に送りつける．著者の経験では，およそ 2%程度の欠損を確認することが多い．通信パケットの欠損が起きたとき，ウェブブラウザとウェブサーバーの間の通信の場合は，ウェブブラウザからの要求とウェブサーバーからの応答，という双方向の通信の中で欠損を検知，自動で修復できる．一方，ウェブブラウザから中継サーバーへの通信は，中継サーバーは前提となるウ

ェブブラウザとWebサーバの間の通信を知らない．だから，通信パケットが欠損しても，中継サーバーは，それを検知できない．そして，中継サーバーは，ウェブブラウザに再送を要求することができず，アクセスログは欠損したままになる．サーバーログ型では，このような欠損は起きないため，ページを遷移して，どのような商品を確認した，といった大まかな行動の分析に向いている．以上から，ウェブサイトの単一ページの中の詳細な行動分析はウェブビーコン型，ウェブサイト全体を俯瞰する分析はサーバーログ型が向いている，と考えることができる．

2.1.3 アクセスログに潜むボットの検知方法

アクセスログに含まれるボットは，DDos 攻撃[21]のようなウェブサイトに急激なアクセスを集中させる有害なボットと，検索エンジンが，検索サービスの提供のために，ウェブサイトに緩やかなアクセスをしながら情報を収集する無害なボットの大きく 2 つに分けることができる．一般に無害なボットは，HTTP メッセージの内容に，自身が検索エンジンからのボットであることを示すユーザーエージェント情報を記録している．一方，有害なボットは，自身がボットであり，同一の主体からの大量アクセスであることを隠すために，HTTP メッセージの内容を偽装することが多い．偽装されたボットは，ウェブサイトの負荷になるだけでなく，アクセスログを用いたウェブサイトの分析にとって有害なノイズとなる．このようなアクセスログに潜むボットを判別する手法として，様々な先行研究が行われている．

先ず，著者の先行研究[22]では，ボットに関する知識を用いたボットの判別モデルを提案している．本研究は，先行研究の拡張として，先行研究の判別手法を用いてボットを除外したアクセスログに対して，新たに異常検知の手法を用いて残存したボットを判別している．新たな手法を取り入れながら，先行研究で除外しきれなかったボットを除外することで，レコメンデーションの精度向上などの有用な成果を確認している．

また，本研究の提案手法を構築する上で，共通の考え方を取り入れている関連研究を列挙する．Zhen ら[23]，Kheir[24]，Kitts ら[25]，Grill ら[26]は，アクセスログに潜むボットの判別のために，ユーザーエ

ージェント情報に含まれる文字列のパターンを分析している．本研究も，ユーザーエージェント情報をボット判別に活かしており，考え方に共通性がある．但し，本研究では，ユーザーエージェント情報のみを活用するよりも，ページビュー数や様々な行動パターンを組み合わせて活用したほうが精度の高い判別が行えることを確認している．同じように部分的に共通性のある関連研究を挙げると，Loyola ら[27]は，行動パターンの解釈を判別に活かしている．更に，Stassopoulou ら[28]，Alhosseini ら[29]，Kouvela ら[30]は，そのような行動パターンを機械学習モデルに学習させて，判別を行っている．また，Mitterhofer ら[31]，Masud ら[32]は，複数のアクセスログを取得することで，そのような行動パターンをより細かく分析している．本研究は，これらの関連研究の考え方を統合して用いることで，ウェブサイトの運用にそのまま実用できる実践的な手法を提案している．表 2-2，2-3，2-4，2-5，2-6 に，ボット検知に関する先行研究を一覧としてまとめる．

表 2-2: ボット検知に関する主な研究(1/5)

ユーザーエージェント情報の分析	Zhang ら 2015[23]	HTTP メッセージに含まれるユーザーエージェント情報の文法チェックによって，Web サイトを攻撃するボットを判別した
	Kheir2012 [24]	音楽サイトのアクセスログに含まれるクロスサイトクリプティング，SQL インジェクション等の攻撃のデータを分析して，ボットの特徴を示すユーザーエージェント情報を抽出した
	Kitts ら 2013[25]	クリック詐欺のボットのユーザーエージェントを分析し，判別を行うシステムを提案した
	Grill ら 2014[26]	マルウェアに感染したホストコンピューターからのボットの HTTP 通信を，ユーザーエージェント情報の分析によって判別した
行動パターンの分析	Loyola ら 2016[27]	WEB サイトでの行動パターンから，人間と人間以外のボットを判別するモデルを提案した
	Stassopolou ら 2009[28]	WEB アクセスログを用いて，1 回の訪問単位の行動パターンを分析することで，人間とクローラーの行動パターンを判別する機械学習モデルを提案した
	Alhosseini ら 2019[29]	ソーシャルネットワークにおけるスパムボットをユーザー間の繋がりのグラフ構造を分析することで判別した

表 2-3: ボット検知に関する主な研究(2/5)

行動パターンの分析	Kouvela ら 2020[30]	Twitter におけるフィッシング詐欺やフェイクニュースを拡散を行うボットを解釈性に優れた機械学習モデルで判別した
	Yu ら 2010[33]	検索エンジンに不正なクエリを発行するボットをクエリのログから判別を行った
	Cai ら 2017[34]	ソーシャルネットワークにおけるフェイクニュースを拡散するボットを，ソーシャルネットワーク上での発言内容のテキストを分析することで判別した
	Yang ら 2020[35]	Twitter におけるボット判別において，学習データのラベル付けの工夫によって，従来よりも高い精度で判別を行うモデルを提案した
	Daya ら 2019[36]	ゼロデイ攻撃を行うホストを，ホスト間の通信関係をグラフ構造に表現しながら，2 段階の教師あり学習，教師なし学習によって判別を行った
	Tao ら 2018[37]	オンラインゲームの不正ユーザーを，知識を用いた教師あり学習，異質な行動を探し出す教師なし学習の二段階の判別を行った

表 2-4: ボット検知に関する主な研究(3/5)

行動パターンの分析	Kang ら 2013[38]	オンラインゲームの不正ユーザーの行動パターンを分析，特定の行動の回数にしきい値を設ける等のルールベースの判別を行った
	Beskow ら 2018[39]	ソーシャルメディアにおけるボットを，ユーザー間のコミュニケーションネットワークを機械学習によって分析し，判別を行った
	Chen ら 2010[40]	オンラインゲームの不正ユーザーの意思決定パターンを分析，人間とボットとの判別を行った
	Efthimion ら 2018[41]	Twitter におけるボットを，ユーザー名の長さ，応答率，感情表現，相互フォロー率などの特徴量を用いて教師あり学習による判別を行った
	Heidari ら 2021[42]	Twitter におけるボットに対して，ツイートの感情表現を特徴量として用いた判別を行った
	Mohammad ら 2019[43]	ソーシャルメディアのユーザーアカウント情報から，畳み込みニューラルネットワークを用いて判別モデルを構築した

表 2-5: ボット検知に関する主な研究(4/5)

行動パターンの分析	Heidari ら 2020[44]	自然言語処理エンジン Google BERT を用いて，Twitter のツイートの感情表現を分析して，ボットの判別を行った
	Beatson ら 2021[45]	ソーシャルメディアのボットのルールベース判別を提案した
	Acien ら 2020[46]	ウェブサイトのカーソル操作の軌道を用いたボットの判別を提案した
	Lee ら 2015[47]	オンラインゲームのボットの行動パターンについて，大量データを用いた詳細パターンの分析によって判別を行った
	Varol ら 2017[48]	Twitter におけるボットを，友人の内容，ツイートの感情表現，人間関係のパターン，行動の時系列情報などの特徴量を用いて教師あり学習による判別を行った
	Chu ら 2018[49]	ウェブサービスのボットを，クライアントサイドで発生させたログをサーバーに転送して，サーバーで判別を行うシステムを提案した

表 2-6: ボット検知に関する主な研究(5/5)

行動パターンの分析	Jiら 2014[50]	ウェブシステムに訪れるボット群の連動関係に着目し，ボット単体に着目した判別に比べて，優れた精度で判別が可能であることを示した
	Phamら 2022[51]	ソーシャルメディアのボットをユーザー間の繋がりのネットワークから分析，アカウント情報やツイート情報を使わずに高い精度で判別が可能であることを示した
	Cabriら 2018[52]	ウェブシステムに訪れるボットをアクセスが続いている状態で，訪問のセッションすべての情報を用いずに，即座に判別する手法を提案した
複数のアクセスログの分析	Mitterhofer ら[31]	オンラインゲームにおいて不正なスクリプトを用いて操作を行うユーザーを，サーバーサイドのログを分析することで判別した
	Masudら [32]	ボットによるアクセスをネットワーク機器のログとサーバーのログを合わせて分析することで判別した

2.2　住宅情報ポータルサイトの行動データへの不正識別子の混入問題

本章では，住宅情報ポータルサイトにおける行動データへの不正識別子の混入の問題を説明する．住宅情報ポータルサイトの行動データにおける識別子とは，特定の住宅情報やページ，情報照会のための問い合わせボタンなどの機能などを，アクセスログ上で一意に識別するために用いられる．住宅情報ポータルサイトでは，このような識別子を用いて，ユーザーの行動をアクセスログに正確に記録して，住宅情報の自動推薦システムのインプットデータに用いている．また，このようなアクセスログの識別子は，ウェブサイトを開発する際に，ある機能が動いたときに規定の文字列ルールに従ってレコード中に記録が残るようにプログラムで定義することで実現している．

本研究で取り扱う行動データの不正識別子の混入の問題は，継続的なサイト開発に伴う膨大な回帰テストに起因する．住宅情報ポータルサイトは，競合サイトとの厳しい競争の中で，ユーザー体験の改善を目的に，高頻度で，継続的な開発を行っている．例えば，ページの改修において，行動データの生成機能を担うJavascriptの変数名が競合してしまったとき，既存の変数の名前を変えると，行動データの一貫性が損なわれる，といったデグレードバグが発生する．このようなデグレードバグを防止するためには，グローバル変数の利用を控えることが定石となる．しかし，ユーザーのサイト上のあらゆる行動履歴を記録するという行動データの性質上，あらゆる機能の開発から変更の影響を受けてしまう．その結果，回帰テストの量は膨大になってしまう．テストの量が膨大になると，開発のスピードが上がらず，ユーザー体験の改善が進ま

ない．開発のスピードを優先して，テストの量を減らせば，行動データ，その先の AI，住宅情報の推薦システムに必要なインプットデータの品質が下がり，結果としてユーザー体験の改善が進まない，という悪循環に陥ってしまう．本研究では，継続的なサイト開発の中で行動データの識別子の文字列ルールを管理し，不正識別子の混入を効率的にチェックするためのシステムを提案する．

本研究の提案するシステムは，ウェブサイトの自動テストのフレームワーク Selenium をシステムに組み込みながら，行動データのテストコードそのものの生成を自動化する．関連のある研究を調査した．Shariff ら[53]は，仕様の異なる様々なウェブブラウザにウェブページをロードする速度確認の膨大なテストを効率化するために，Selenium を使った自動テストのシステムを構築している．また，Chen ら[54]は，ウェブページに埋め込んだ Javascript のコードを効率的にリファクタリングするために，Selenium を使ってテストコードを自動で生成するシステムを構築している．本研究では，行動データの出力のタイミングと内容のテストに焦点を当てており，ページのロード速度やコードとしてのメンテナンスのしやすさなど，出力機能の具体的な実装方式には焦点を当てていないが，関連する重要なテスト観点を扱った研究として報告する．Castro ら[55]は，ウェブアプリケーションを Selenium で稼働させ，ウェブアプリケーションが更新したデータベースの内容を，自動でテストするシステムを構築した．手動で全ての工程をテストした場合と比較して，92％の作業時間を短縮できたことを報告している．本研究も，ウェブアプリケーションの稼働結果をデータベース上で確認している点で関連のある研究である．Iyama ら[56]は，設計書から画面遷移のテストコードの一部を自動生成する手法を研究している．本研究では，設計に関する情報を人手でデータベースに登録することで，テストコードの全てを自動生成しており，関連のある研究である．Fard ら[57]は，ウェブサイトのクローリングしたログ情報と開発者のウェブサイトの目的に関する知識を組み合わせて，テストケースを効率的に生成する手法を研究している．開発者の

知識とクローリング技術を組み合わせている点で，本研究と関連のある研究と言える．Nagarajan ら[58]は，ウェブサイトのテストへの入力データを自動生成するシステムを研究している．Mirshokraie ら[59]は，DOM 情報を自動で収集することで，Javascript コードのテストケースを自動生成する手法を提案している．Marshall ら[60]は，Selenium による自動テストの実行速度を上げるために，Selenium の設定情報の分析し，成果を報告している．Leotta ら[61]は，Selenium が利用する WebDriver などの周辺ツールの評価を行っている．Neto ら[62]は，家庭の電力管理システムを題材に，テストケースの自動生成を，Selenium を活用して，実現している．Stocco ら[63]は，テストしやすいウェブページの構造を Selenium で自動生成する手法を提案している．Lim ら[64]は，RDBMS の GUI によるテストを Selenium によって効率化する手法を提案している．Martinez ら[65]は．UI のテストを効率的に行うためのミドルウェアの構造を提案しており，その構造の中に Selenium を組み込んでいる．Taneja ら[66]は，回帰テストに焦点を絞って，テストケースを自動生成する手法を研究している．Bures ら[67]は，Selenium の WebDriver を自動テスト用に拡張する手法を提案している．Boni ら[68]らは，クラウド環境の計算資源を活用できるように WebDriver を拡張して，自動テストを効率化することを提案している．Kirinuki ら[69]は，過去に生成したテストコードを回帰テストで再利用する際に，他の改修の影響に自動で対応する機能を提案している．本研究も，テストコードの編集を不要する手法を提案しており，関連のある研究と言える．表 2-7，2-8，2-9 に，自動テストに関する先行研究を一覧としてまとめる．

表 2-7: 自動テストに関する主な研究(1/3)

自動テストシステムの提案	Shariff ら 2019[53]	仕様の異なる様々なウェブブラウザにウェブページをロードする速度確認の膨大なテストを効率化するために，Selenium を使った自動テストのシステムを構築した
	Chen ら 2013[54]	ウェブページに埋め込んだ Javascript のコードを効率的にリファクタリングするために，Selenium を使ってテストコードを自動で生成するシステムを構築した
	Castro ら 2013[55]	ウェブアプリケーションを Selenium で稼働させ，ウェブアプリケーションが更新したデータベースの内容を，自動でテストするシステムを構築した
	Lim ら 2019[64]	RDBMS の GUI によるテストを Selenium によって効率化する手法を提案している
テストコードの自動作成	Iyama ら 2018[56]	設計書から画面遷移のテストコードの一部を自動生成した
テストケースの自動作成	Fard ら 2014[57]	ウェブサイトのクローリングしたログ情報と開発者のウェブサイトの目的に関する知識を組み合わせて，テストケースを効率的に生成した

表 2-8: 自動テストに関する主な研究(2/3)

テストケースの自動作成	Neto ら 2016[62]	家庭の電力管理システムを題材に，テストケースの自動生成を，Selenium を活用して，実現している
	Mirshokraie ら 2013[59]	DOM 情報を自動で収集することで，Javascript コードのテストケースを自動生成する手法を提案した
	Taneja ら 2011[66]	回帰テストに焦点を絞って，テストケースを自動生成する手法を研究している
	Nagarajan ら 2017[58]	ウェブサイトのテストへの入力データを自動生成するシステムを構築した
自動テストのコンポーネントの研究	Marshall ら 2019[60]	Selenium による自動テストの実行速度を上げるために，Selenium の設定情報の分析し，成果を報告した

表 2-9: 自動テストに関する主な研究(3/3)

自動テストのコンポーネントの研究	Leotta ら 2013[61]	Selenium が利用する WebDriver などの周辺ツールの評価を行った
その他のテスト効率化の研究	Stocco ら 2015[63]	テストしやすいウェブページの構造を Selenium で自動生成する手法を提案している
	Martinez ら 2014[65]	UI のテストを効率的に行うためのミドルウェアの構造を提案しており，その構造の中に Selenium を組み込んでいる
	Bures ら 2016[67]	Selenium の WebDriver を自動テスト用に拡張する手法を提案している
	Boni ら 2018[68]	クラウド環境の計算資源を活用できるように WebDriver を拡張して，自動テストを効率化することを提案している
	Kirinuki ら 2019[69]	過去に生成したテストコードを回帰テストで再利用する際に，他の改修の影響に自動で対応する機能を提案している

2.3　スーパーマーケットチェーンの行動データからの優良顧客抽出の問題

本章では，スーパーマーケットチェーンの行動データにおける大量の一般顧客から少数の優良顧客を抽出する問題を説明する．スーパーマーケットの業界では，優良顧客に絞り込んだマーケティング施策を行うにあたって，大量の一般顧客の行動データが，分析のノイズになってしまう問題がある．そこで，本研究では，スーパーマーケットチェーンの優良顧客を抽出するため，伝統的顧客抽出手法である RFM 分析を拡張した新たな手法を提案する．RFM 分析とは Recency，Frequency，Monetary の 3 の指標を用いて顧客を表現する手法である[70][71][72]．Recency はより最近来店していること，Frequency はより頻繁に来店していること，Monetary は総購入額をそれぞれ示す．RFM 分析は取り組む課題に対応して様々な拡張が提案されており 2-3-1．節でそれらを関連研究として説明する．また，我が国のスーパーマーケットの経営課題を 2-3-2 節で，スーパーマーケットの経営課題への対策を行った関連研究を 2-3-3 節でそれらを説明する．2-3-4 節では，前節までに取り上げた関連研究と本研究の位置づけについて説明する．

2.3.1 RFM 分析の拡張に関する研究

Chen ら[73]は，物流業界の顧客の解約予測問題に対して RFM 分析の拡張モデルを提案している．RFM 指標に顧客の会員期間の長さ(Length)，計上した利益(Profit)の 2 つを加えることで LRFMP モデルとし，解約予測精度とモデルから得られる示唆において改善が見られたことを報告している．この他にも RFM 指標に期間の観点を加えた研究として，Bizhani[74]らは，銀行の顧客セグメンテーションに RFM 指標と顧客としての期間を組み合わせた手法を提案している．また，Wu ら[75]は，小児歯科患者の分析において RFM 指標，患者としての期間，性別や年齢等のデモグラフィック情報,ベイジアンネットワークを利用した．また，指標を足すだけでなく機械学習手法を組み合わせた研究も報告されている．Chan ら[76]は，自動車販売店の顧客セグメンテーションに RFM 分析と群知能を組み合わせた手法を提案している．Kim らは，特許の利用データから RFM 指標を抽出し，決定木分析を行うことで，将来重要となる特許のパターンを発見した[77]．Poel ら[78]は，新聞の解約顧客予測に対して，RFM 指標とブランドイメージ等のアンケート結果，ベイズモデル，分位点回帰を組み合わせることを提案している．その他にも RFM 指標自体の抽出方法の拡張として，Zeng ら[79]は家庭用生地販売店の顧客セグメンテーションに対して，RFM 指標を月あたり来店回数の平均，最大値，最小値など 10 指標に細分化して利用する手法を提案している．表 2-10 に，RFM 分析の拡張に関する先行研究を一覧としてまとめる．

表 2-10: RFM 分析の拡張に関する主な研究

著者	適用した業界	拡張した情報や手法
Chen ら 2015[73]	物流	会員期間の長さ, 累積利益
Bizhani ら 2010[74]	銀行	会員期間の長さ, 累積利益
Wu ら 2012[75]	小児歯科	患者としての期間の長さ, 性別, 年齢などのデモグラフィック情報, ベイジアンネットワーク
Chan ら 2016[76]	自動車販売店	群知能
Kim ら 2012[77]	特許管理	決定木
Poel ら 2013[78]	新聞販売店	ブランドイメージ等のアンケート結果, ベイズモデル, 分位点回帰
Zheng ら 2015[79]	家庭用生地販売店	RFM 指標を期間や基礎統計量に分割

2.3.2　　我が国のスーパーマーケット業界の状況

日本のスーパーマーケット業界は長期的な売上減少傾向にある．図 2-2 には日本チェーンストア協会がチェーンストア販売統計として収集したデータ[80]から作成した 1992 年から 2015 年に至るまでの日本のスーパーマーケットの売上推移を示す．

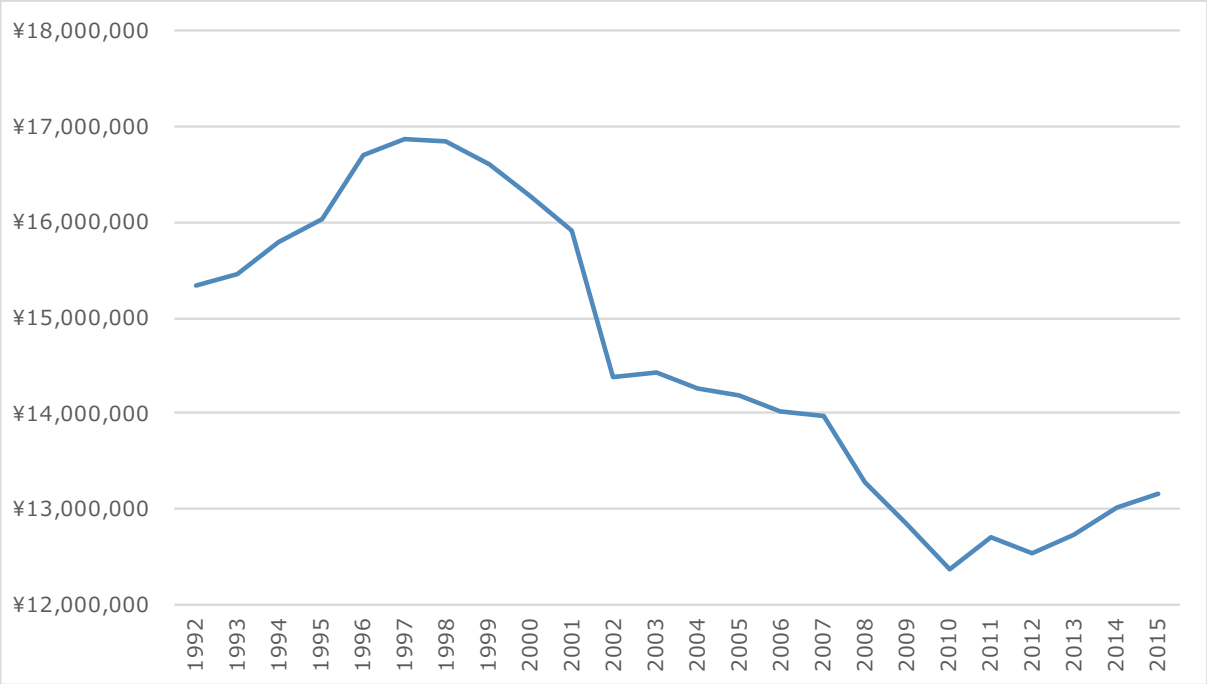
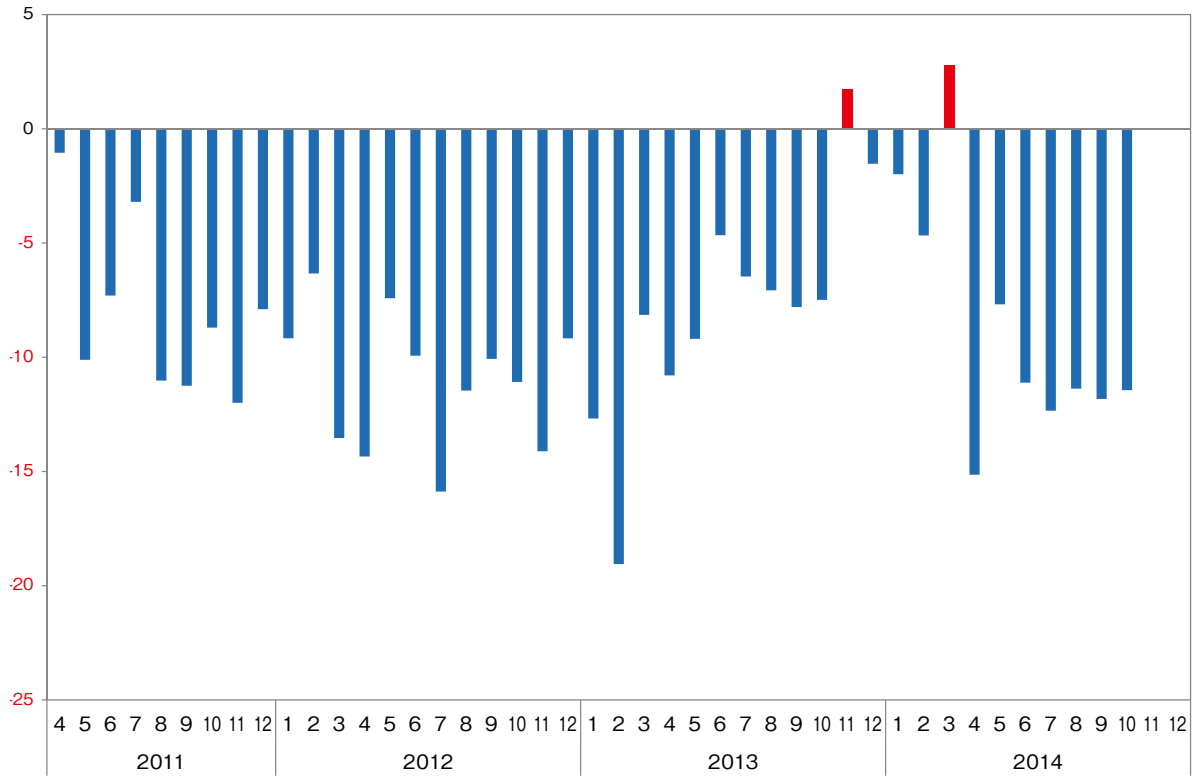


図　2-2：日本のスーパーマーケットの売上推移

1997 年をピークに 2010 年まで売上の減少が続いており，それ以降小幅な成長が見られるもののピーク時の 70%程度の水準に落ち込んでおり長期的な売上減少傾向が確認できる．以降では売上減少の背景にある来店顧客数の減少，食品市場の縮小，食品購買手段の多様化について説明する．

売上減少の背景には来店数の減少がある．図 2-3 には新日本スーパーマーケット協会が協会加盟企業 125 社へ実施したアンケート調査[81]から，前月との来店数の増減について確認した結果を示す．

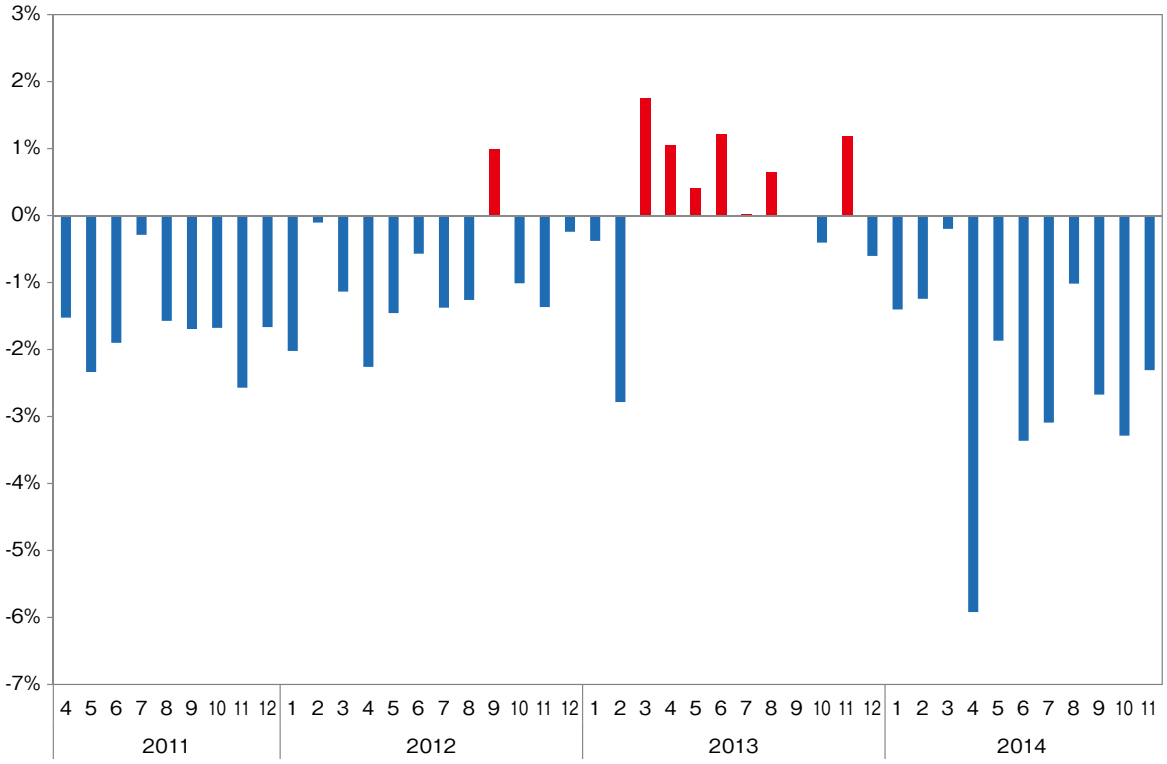


出典：2015 年度版スーパーマーケット白書

図 2-3: 日本のスーパーマーケットの来店数の増減調査

調査期間の 2011 年 4 月から 2014 年 10 月までの 43 ヶ月の内 41 ヶ月で来店数の減少が確認でき，来店数の増加があった 2 件のうち 2014 年 3 月の来店数増加は消費税率引き上げの影響と予想できる．

一方，図 2-4 には同じく新日本スーパーマーケット協会が小売業従業者を除く 20 代から 60 代の男女 31,332 名に実施したアンケート調査[81]から，前月との来店回数の増減について確認した結果を示す．



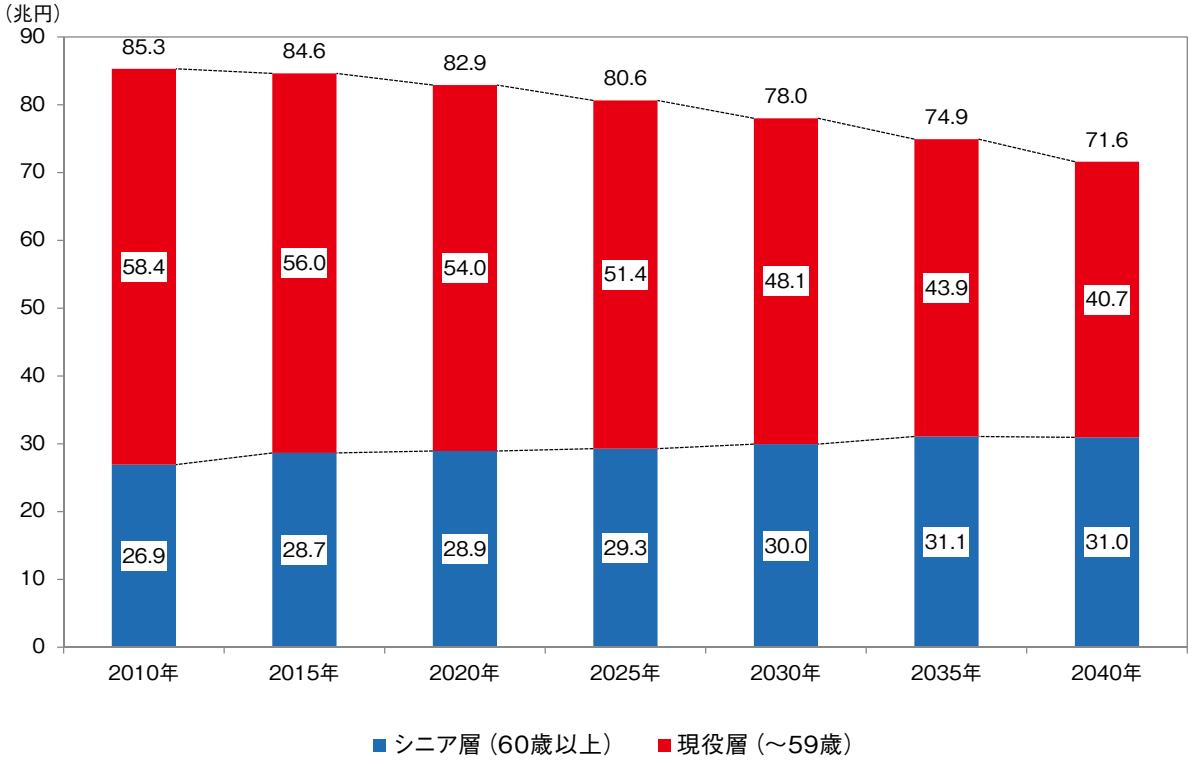
出典：2015 年度版スーパーマーケット白書

図 2-4: 日本のスーパーマーケット顧客の来店回数の増減

2012, 2013 年に増加を示す月が見られるものの調査期間の 8 割で減少が確認できる. この 2 つのアンケート調査の結果から, 日本のスーパーマーケットの来店数減少の実態を需要側と供給側の双方の回答から確認できた. 本研究の提案モデルは顧客管理の指標のひとつとして, 顧客の将来価値を予測することで離脱に近い顧客を予め把握し, 離脱を防止する対策に必要な情報を提供する.

スーパーマーケット業界は売上減少を恒久的に取り組むべき課題と捉えるべきか, そうでないかを検討する上で日本の人口動態変化が与える食品市場の縮小に着目する. 総務省の調査によれば日本の 65 歳以上の高齢者人口は平成 25 年 9 月 15 日時点で 3186 万人, 総人口 1 億 2726 万人の 25.0%に達している[82]. また, 総務省の推計によれば平成 47 年までに総人口は 1 億 1212 万人に減少, 高齢者

人口の割合は 33.4%に達すると予測されている．一般に高齢者は若い世代と比較して食品の消費量が少ないと考えることができ，図 2-5 には新日本スーパーマーケット協会が公表している人口減少，高齢化の影響を踏まえた将来の食品市場の推計結果を示す[81]．

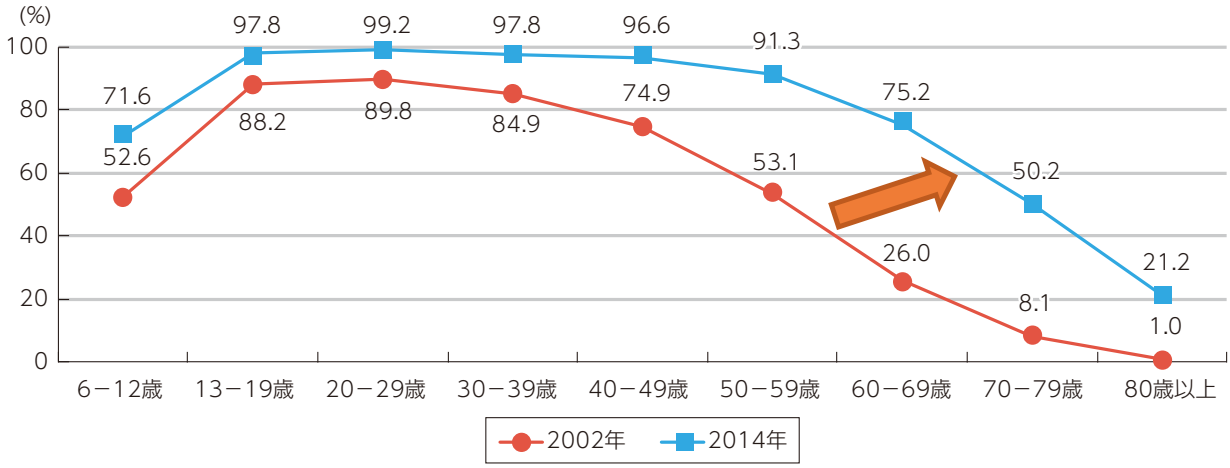


出典：2015 年度版スーパーマーケット白書

図 2-5: 食品市場規模の推計

この推計結果によれば日本の食品市場は 30 年で 20%程度の縮小が見込まれ，スーパーマーケット業界の売上および来店数の減少が今後も継続するであろうことを支持した結果となっている．

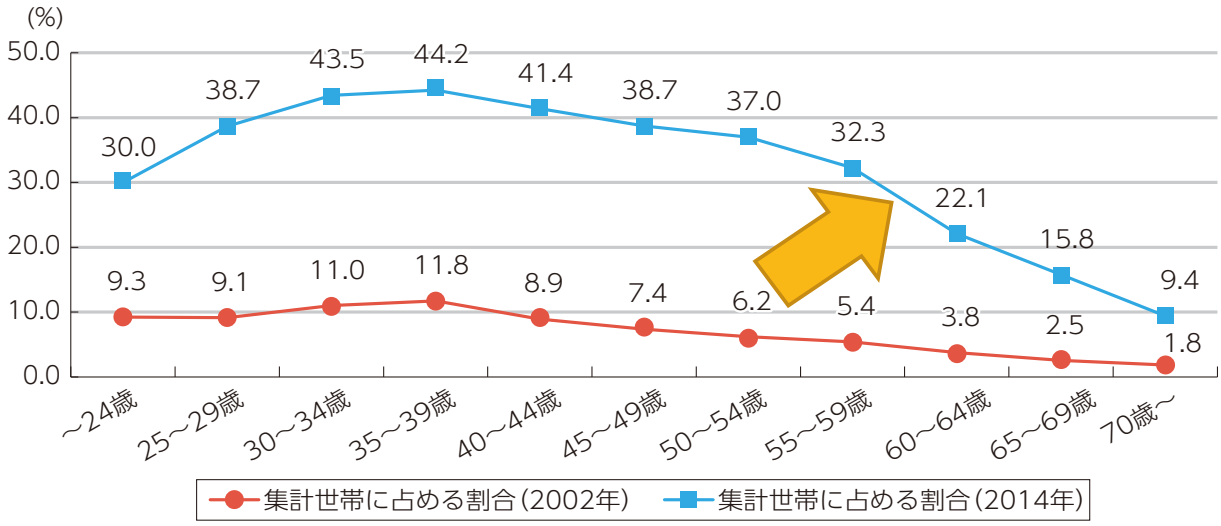
近年，インターネット利用およびインターネット通販が普及しており，その利用者は全世代に，購買対象は食品に及んでおり，スーパーマーケット業界の売上減少の一因となっている．図 2-6 には総務省が調査した日本の世代別インターネット利用率を示す．



出典：平成 27 年度情報通信白書

図 2-6: 日本の世代別インターネット利用率

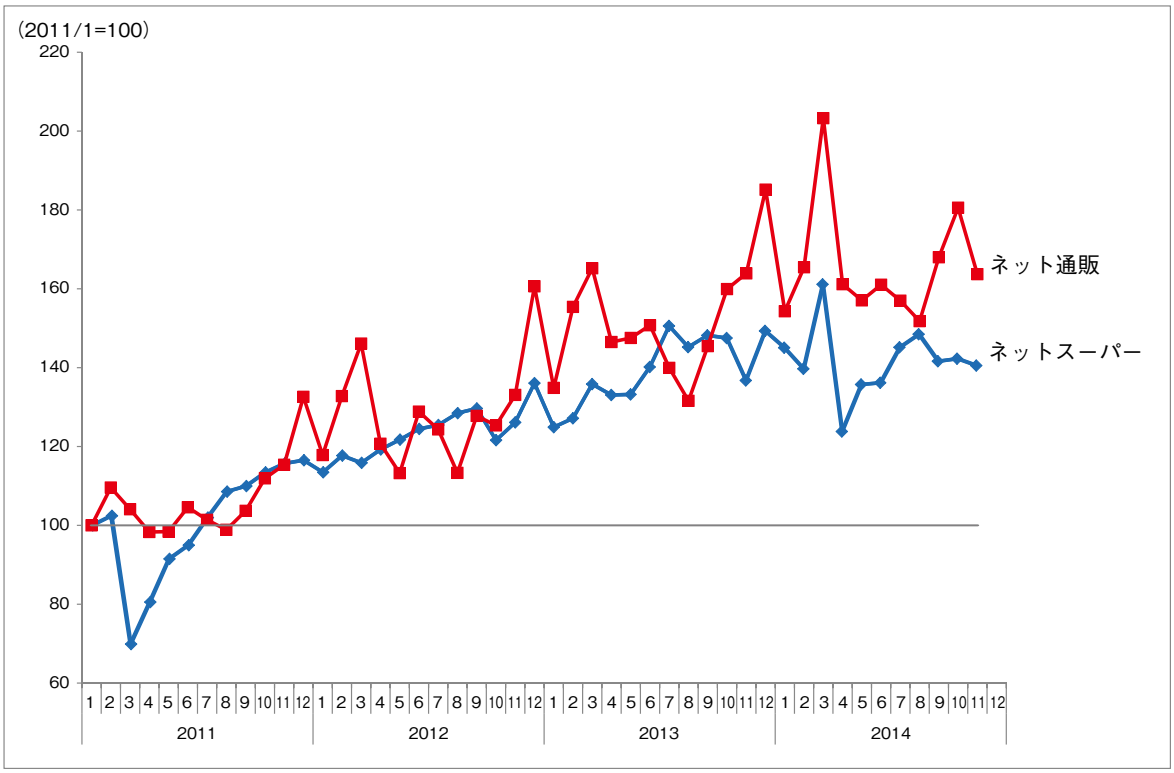
2002 年と 2014 年の調査結果を比較すると，調査対象の全ての世代においてインターネットの利用率が上昇していることが確認できる．図 2-7 には同じく総務省が調査した日本の世代別ネットショッピング利用率を示す．



出典：平成 27 年度情報通信白書

図 2-7: 日本の世代別ネットショッピング利用率

同じく 2002 年と 2014 年の調査結果を比較すると，相対的に多忙で，インターネットを使い慣れている若い世代だけでなく高齢の世代も含めてネットショッピングが利用されていることが確認できる．そして図 2-8 には新日本スーパーマーケット協会が調査したインターネットを通じた食品購入頻度の時系列推移を示す．



出典：2015 年度スーパーマーケット白書

図　2-8：インターネットを通じた食品購入頻度

この調査ではネットショッピングの購買対象を食品に限定し，またネット上の店舗を，生鮮食品など実際のスーパーマーケットと同じ商品目を取り扱ったネットスーパーとそれ以外の一般的なネット通販に店舗形態を分けて集計している．いずれの形態でも調査対象の 2011 年から利用の普及が確認でき，インターネットショッピングの対象は食品に及んでいることが確認できる．但し，食品購買手段の多様化としてのインターネットショッピングの普及は必ずしもスーパーマーケット業界の売上減少に繋がるわけではない．表 2-11 には大手スーパーマーケット企業であるイトーヨーカドーのネットスーパーの業績を示す．

表 2-11: イトーヨーカドーのネットスーパー業績

年	売上高	店舗数	会員数
平成19年	50億円	80店舗	17万人
20年	130億円	89店舗	33万人
21年	210億円	118店舗	60万人
22年	300億円	133店舗	86万人
23年	350億円	137店舗	116万人
24年	420億円	145店舗	ー

出典:平成 27 年度情報通信白書

平成 19 年からの 5 年間で売上は 5 倍，会員数は 10 倍に増加していることが確認できる．つまりスーパーマーケット企業は自社でネットスーパーを展開することで売上を拡大することができる．しかし，スーパーマーケットの実店舗にとって食品購買手段の多様化，ネットスーパーの普及は脅威であり，実店舗だからこそ可能な価値提供を追求しなければならない．本研究は実店舗の経営支援を目的としている．

2.3.3 スーパーマーケットの経営改善のための研究

情報通信機器を活用した販売促進に関する研究として，Nurmi ら[83]は，店舗の買い物かごに付属させたモバイル端末上で商品推薦を行い，顧客の購入履歴の解析結果に基づく推薦内容がランダムに選択した場合よりも売上向上に寄与することを示した．同じく売上向上を目的に購入商品に着目した研究として，大澤らは，文書解析の手法として開発されたアルゴリズムである Key Graph を用いてスーパーマーケットの POS データを分析し，顧客の潜在的需要に基づいた販売に注力すべき商品を提示した[84]．飯塚らは，購入金額による会員のランク分けを行い,ランクごとの来店回数や購買品目を分析することで，高ランクの会員の特徴および低ランクの会員を高ランクに成長させるための施策を示した[85].その他に商品に着目した研究として，高橋らは，商品の欠品を課題として捉え,新品目の将来の売れ行きの予測に有用な目利き会員の判別を行った[86].また，商品ではなく店舗環境に焦点を当てた研究にも実績がある．店舗のレイアウトに関する研究として，Chen らは，膨大な種類の商品を店内にどのように配置することが顧客の移動効率上最適か，顧客の購入履歴と遺伝的アルゴリズムを利用して解析した[87]．岸本らは，同様の問題に対してエージェントベースシミュレーションを用いて取り組んだ[88]．同じく店舗環境に関する研究として，川田らは，店舗の音環境に焦点を当て，来店者の意識する音,不快に感じる音の調査を行った[89]．

2.3.4 実務家が用いる顧客管理の手法

スーパーマーケットの経営に有用な顧客管理手法として，表 2-12 に実務家が用いる顧客管理手法を示す[90].

表 2-12: 実務家が用いる顧客管理手法

手法	分析指標	実施概要
デシル分析	購買額	購買額で顧客を 10 段階程度にランク付けし，上位 3,4 ランクに割引・DM 送付等のマーケティング投資を振り分ける
RFM 分析	購買額，来店頻度，直近来店日	購買額に加え，来店頻度と直近来店日を考慮し，顧客が生存しているかどうかを判別する
顧客ベース管理	生存顧客数	会社の資産としての生存顧客数を管理する
カスタマー・エクイティ管理	顧客の将来価値	顧客の将来価値を管理する

本研究では，将来の優良顧客の判別実験としてデシル分析を用いて顧客のランク付けを行い，基準の時点よりも過去の購買行動から将来の上位ランクを予測する．過去の購買行動の分析は提案手法にRFM 分析を組み合わせて行う．また，本研究では，生存顧客の判別実験を合わせて実施し，顧客ベース管理の一方法として提案する．本研究は，顧客の購買行動の符号化に文書符号化手法の概念を用いて表現した購買商品目を用いている点で新規性があるが，顧客の将来価値を優良顧客，生存顧客として予測している点で表 2-12 に挙げた 4 つの顧客管理手法と関連がある．

2.3.5 本研究の位置づけ

本研究は，RFM 分析，購入商品目とその店舗別異質性，機械学習手法を組み合わせる点で RFM 分析の拡張に関する研究のひとつである．本研究では，現在の店舗売上を生み出しているだけでなく，将来に渡って店舗売上を生み出す顧客を優良顧客と定義する．そして，市場縮小局面にある日本のスーパーマーケット経営者に対し，最優先に維持すべき優良顧客の抽出情報と優良顧客が顕著に購入している商品目情報の 2 つを提供する． RFM 分析の拡張研究としても，スーパーマーケットの経営課題を扱った研究としても，著者が調査した限り同様の研究は存在しない．

第3章　マルチソースアクセスログ分析によるボットの検知

本章では，住宅情報ポータルサイトに訪れるインターネットボットを，集客分析などポータルサイトのマーケティング課題への対応として行動データから除去する新たな手法を提案する．住宅情報ポータルサイトのシステム構成の特徴を活かして，複数の技術方式によって抽出したアクセスログからボットの知識を新たに獲得して，ボット検知に活用する．主要な効果としての従来手法と比べたボット検知精度，副次効果としてのリコメンドシステムへの性能の改善効果を報告する．

3.1 はじめに

住宅情報ポータルサイトには，ネット検索事業者や競合ポータルサイトによる調査のためのボット，悪戯のためか同一コンテンツに大量アクセスを行ってサーバーを攻撃するボットなど，非常に多様なボットが訪れ，それらのボットは，住宅検討者の分析にとって深刻なノイズとなる．行動データを人の目で確認すると，住宅情報ポータルにおける不自然な振る舞いについて考え，判断することができるが，毎日発生する膨大なデータを人の目で確認することは困難であるため，機械化が必要である．伝統的な異常検知の手法では，ウェブサイトへのアクセス規模が外れ値であるか，といった検知しかできず，検知精度に課題がある．本研究の手法は，住宅情報ポータルサイトのシステム構造を活かして，実際に訪れているボットの知識を獲得する手法，それを活かして異常検知の精度を向上させる新たな手法として提案する．

3.2 ボット検知の方法

本節では，マルチソースアクセスログ分析によるボット検知の方法を説明する．

3.2.1 マルチソースアクセスログ分析の概要

まず，提案手法が用いるアクセスログの内容を説明する．ボットからの大量アクセスが発生したこと自体はわかっているが，レコード単位ではボットの判別ができていないアクセスログを用意する．アクセスログは，ウェブビーコン型とサーバーログ型の 2 つをそれぞれ用意する．2 つのアクセスログは，同一期間のアクセスログである．取得できている HTTP ヘッダ中の情報は共通であるため，サーバーへの要求 ID を用いて，アクセス主体の同一アクションによるアクセスログのレコードであることを照合することが可能である．

サーバーログ型は，前章で述べたように，抽出時の欠損リスクが無いため，網羅性が高く，ウェブサイト全体を俯瞰する分析に適した性質を持っている．一方，ウェブビーコン型は，技術方式として，一部のボットからのアクセスの影響を受けない特徴を持っている．その特徴を説明する．

一般的に，ウェブサイトにアクセスするボットは，ウェブサイトのページ構造を調査するクローリング[91]とページ中のテキスト情報などページの内容を抽出するスクレイピング[92]の 2 つの活動を主に行う．ボットは，サイトのページにアクセスするための URL 情報を得るために，クローリングを行うので，サーバーログ型のアクセスログは必ず出力される．一方，ウェブビーコン型のアクセスログは必ずしも出力されない．その理由を説明する．ウェブビーコン型のアクセスログを出力するために HTML コードに埋め込まれている Javascript のコードは，ボットがページの情報をスクレイピングによって取得するために，実行する必要がな

いからである．また，アクセスログの出力する目的以外にも，今日の殆ど全てのウェブサイトには，ページ内の動作機能を担うJavascriptコードが埋め込まれている．だが，情報収集を目的としたボットは，そのようなJavascriptコードを実行する必要がない．また，ボットの目的が，DDos攻撃などのウェブサイトの計算リソースの大規模な消費であっても，ページ中のJavascriptコードはウェブサーバー側ではなくウェブクライアント側で稼働するので，攻撃としての効果は見込みづらい．とはいえ，Curlコマンド[93]などLinuxの標準的な機能を用いて，Javascriptを稼働させるボット，稼働させないボットは，容易に作り分けることができる．その容易性を踏まえると，Javascriptコードが稼働していればボットではない，と判断するのは早計である．

しかし，今日のウェブサイトは，ボタンの押下などページ内の動作機能をJavascriptコードが多く担っている．ユーザーはJavascriptコードを，ウェブブラウザ上で稼働させなければ，サイトを正常に利用することはできない．だから，Javascriptコードが稼働していなければ，そのアクセスはボットである可能性が非常に高い，というのが本研究の考えである．

つまり，サーバーログ型に含まれるアクセスから，ウェブビーコン型を用いてJavascriptコードが稼働していないアクセスを判別することができれば，そのアクセスはボットのものとして除外することができる．この除外を本研究では，ルールベース判定の1段階目のボット除外として扱う．更に，ウェブビーコン型と組み合わせて分析することで，ボットからのアクセスと判定したサーバーログ型のアクセスログを抽出できる．そこからは，ボットの特徴を分析することができる．その分析から得られたボットの特徴を用いて， 2段階目のボット除外としてモデルベース判定を行う．このルールベース判定とモデルベース判定の2段階のボット除外の全体像を図 3-1 に示す．

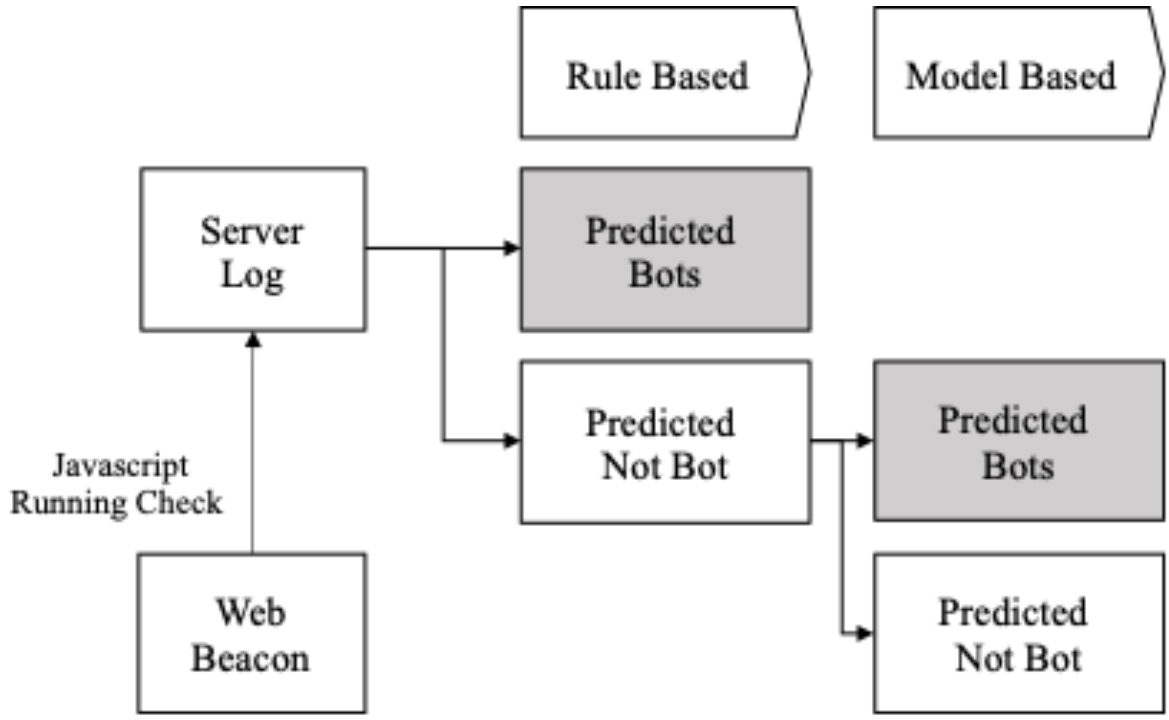


図 3-1: ボット除外の全体のフロー図

図 3-1 にサーバーログとして表記したアクセスログは，住宅情報ポータルサイトのウェブサーバへの 1 日分のアクセスから抽出している． 前節で取り上げた検索エンジン経由の無害なボットを除外し，および特定の住宅商材の関連コンテンツへと絞り込みを行った 9,783,466 レコードを分析に用いる．アクセスログの内容は，一般的な HTTP メッセージ形式として，リクエスト URL やリクエストパラメータの値を含む．そのような値を解釈することで，様々なユーザーの行動を分析できる．例えば，ユーザーが住宅情報の概要を示すページを参照したのか，細かな住宅仕様などの詳細情報を示すページを参照したのか，といった参照ページの種別をできる．あるいは，そのページを日中に見ていたのか，深夜に見ていたのか，といった時間帯も分析できる．更に，それは，数

秒の流し読みだったのか，数分かけてじっくり読み込んだのか，といった時間の長さも，ユーザーの関心を表現する重要な情報として分析することができる．

3.2.2　　ウェブサイトでの振る舞い情報による分析

1 段階目の除外では，Javascript 稼働有無によるルールベースの判定を行う．具体的には，サイトを動作させることに必要な Javascript コードが稼働していない殆どがボットであるという仮説の予測ボットデータと，それ以外の予測非ボットデータに，アクセスログを二分する．

この二分したデータを用いて，ウェブサイト上での振る舞い情報とユーザーエージェント情報に含むテキスト情報の分析を行う．この分析によって，予測ボット，予測非ボットそれぞれの仮説の確からしさを評価する．そして，ボットの特徴を分析することで，ボットの特徴を発見する．
分析には，ボットの特徴，非ボットであるユーザーの特徴に関する仮説を用いる．仮説の内容を表 3-1 にまとめる．

表 3-1: ボットの特徴として検証する仮説

Point of View		#	Features of sessions from Bots
attribution		1	User Agent strings include bot suspicious words
		2	User Agent change at one session
		3	IP address change at one session
		4	Cookie change at one session
		5	Frequent Cookie data loss
Behavior	Visit	6	Early Morning, late at night
		7	From no referrer
		8	From no Ad Banners
	Stay	9	Large amount of Page Views
		10	Mean Interval of time between Page Views are very small (almost 0)
		11	Variance Interval of time between Page Views are very small (almost 0)
		12	Duration of stay is too long
		13	Viewing limited types of pages

ここで，アクセスログの分析におけるレコードのグルーピング，ウェブサイトへの 1 回の訪問の区切り方を説明する．アクセスログ中の 1 回の訪問とみなす単位は，Cookie 中に含まれる同一ユーザーを示す ID からのアクセスのうち，アクセスの間隔が 30 分以内のものは 1 回の訪問として数える．このような手続きは，アクセスログの分析の一般的な手法として知られている[94]．そして，その 1 回の訪問の中で，ユーザーとボットの違いを表す特徴量を抽出し，分析を行う．

尚，本節による検証は，特徴量の統計量および可視化結果は，Javascript 非稼働データはボットであるとする仮説を支持するかどうか，本研究としての一次評価を行う．本研究としての最終評価は，ユーザ

ーまたはボットを示す評価用データを用いた定量評価を行うが，その内容は次章にて説明する．

表 3-1 の仮説について説明する．観点は，アクセス主体そのものが持つ属性とウェブサイト上での振る舞いに分けている．属性は，アクセス主体が持つユーザーエージェント情報，IP アドレス，Cookie 情報など HTTP アクセスを構成する基礎的な情報を用いる[95]．例えば，ユーザーエージェント情報は，1 回の訪問の中でアクセス主体のウェブブラウザや情報端末の OS が変化することは，通常考えられないため，ボットの偽装による変化を疑う特徴量となっている．

次に，振る舞いは，訪問が日中か深夜かを示す時間帯や，1 回の訪問において確認したページの量，ページを確認する時間の間隔などを特徴量として定義している．ウェブサイトを訪問する時間帯の仮説を検証した内容を図 3-2 に示す．

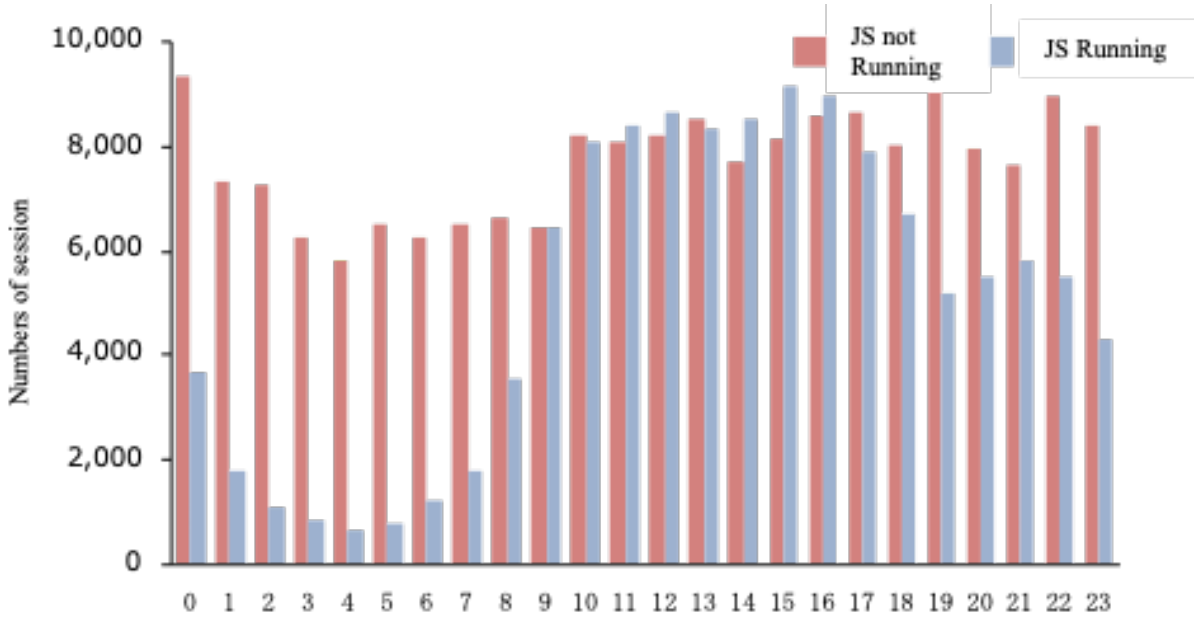


図 3-2: ボットとユーザーのアクセスしている時間帯の違い

Javascript コード非稼働のボットと思われる訪問は，24 時間の中で断続的に発生している．その一方で，Javascript コード稼働のユーザーと思われる訪問は，日中に明確な訪問のピークがあり，深夜になるにつれて訪問が少なくなっている．この内容は，明確に Javascript コードの稼働有無でボットとユーザーを分ける仮定の妥当性を支持しているといえるであろう．

図 3-3 には，1 回の訪問におけるページを確認する間隔時間の平均値の自然対数をとったものを，ヒストグラムとして示す．

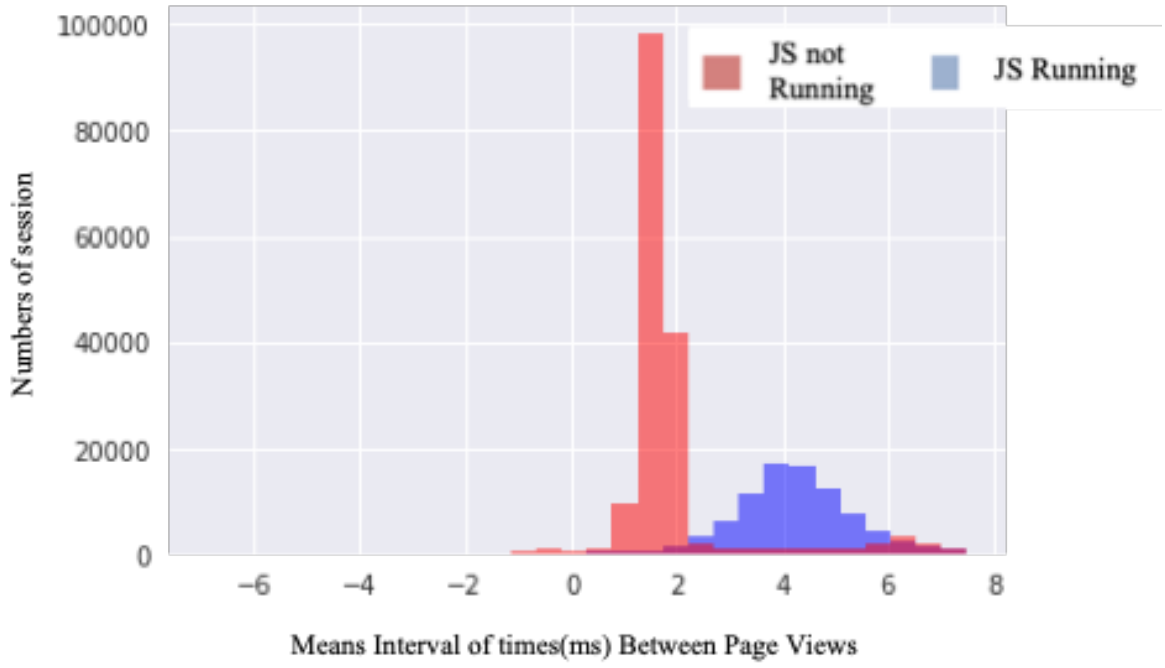


図　3-3：ボットとユーザーの滞在時間の違い

Javascript コード稼働と非稼働の訪問で，ピークとばらつきが異なっている．ボットと思われる Javascript コード非稼働の訪問の間隔時間は，圧倒的に短い上に個体によるばらつきが小さく，ボットによる機械的なアクセスの傾向に当てはまる．一方，ユーザーと思われる Javascript コード稼働の訪問の間隔時間は，長いうえに，個体によるばらつきがあり，ボットに比べたときの人間的な特徴を表現している．この内容も，Javascript コードの稼働有無がボットとユーザーを分けるル

ールとしての有用性を支持している．更に，これらの特徴量は，ボット検知に有用な特徴として，二段階目のモデルベース判定にも用いることができる．

3.2.3 ユーザーエージェント情報による分析

次に，アクセス主体の属性の観点としてのユーザーエージェント情報の分析について説明する．表 3-1 の#1 は，ボット，またはユーザーのユーザーエージェント情報だけに，頻繁に現れる単語が存在するのではないか，という仮説である．我々は，伝統的な手法である bag-of-words 表現[96]を用いて，ユーザーエージェント情報の文字列を分析した．基礎データに含まれる 4,930 ユニーク数のユーザーエージェント情報の文字列から，691 単語を抽出した．尚，数値とみなせる文字列は除外し，アルファベットの大文字，小文字は区別しない．単語の抽出のイメージを図 3-4 にしめす．

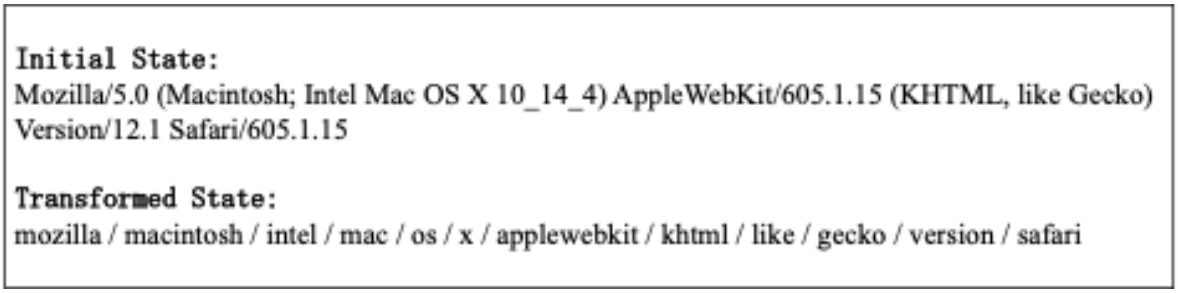


図 3-4: ユーザーエージェント情報の抽出方法

このように抽出した 691 単語のうち，Javascript コードの稼働，非稼働のアクセスを特徴づける単語は一部である，と考えられる．そこで，単語に絞り込みをかけ，分析の目的に不要な単語を除去する．そして，単語の内容そのものがボットまたはユーザーを表す内容であるか定性的な評価を行う．更に次節では，絞り込んだ単語を用いたボット判別モデルを構築することで単語の有用性について定量的な評価を行う．

ここでは，単語の絞り込みのために，ロジスティック回帰モデル[97]の構築を行い，Javascript コードの稼働有無を特徴づける単語を，定量的に評価する．ウェブサイトへの訪問を 1 レコードとして扱い，目的変数は Javascript コードの稼働有無，説明変数には訪問の主体が持つユーザーエージェント情報における 691 単語の出現有無を用いる．この変数設計を用いる場合，各レコードの目的変数の出現を示すデータは殆ど 0 となり，疎データの扱いを工夫する必要がある．そこで，今回は，一般的なロジスティック回帰ではなく，Lasso 回帰モデル[98]を用いることで，偏回帰係数が 0 になる説明変数をモデルから排除する L1 正則化を行うことで対策する．正則化項のチューニングによる単語の絞り込み，および偏回帰係数の推定結果を図 3-5 に，正則化項の値ごとのモデルの実データへの当てはまりを示す AUC[99]を表 3-2 に，その他の当てはまり指標を表 3-3 に示す．尚，表 3-3 に示す当てはまり指標の評価においては，予測モデルの推定に用いる訓練データ，推定した予測モデルの頑健性を確認するための検証データの 2 つを用いて，交差検証を行う．2 つのデータは，アクセスログを分割することで作成しており，それぞれ重複するレコードは存在しない．

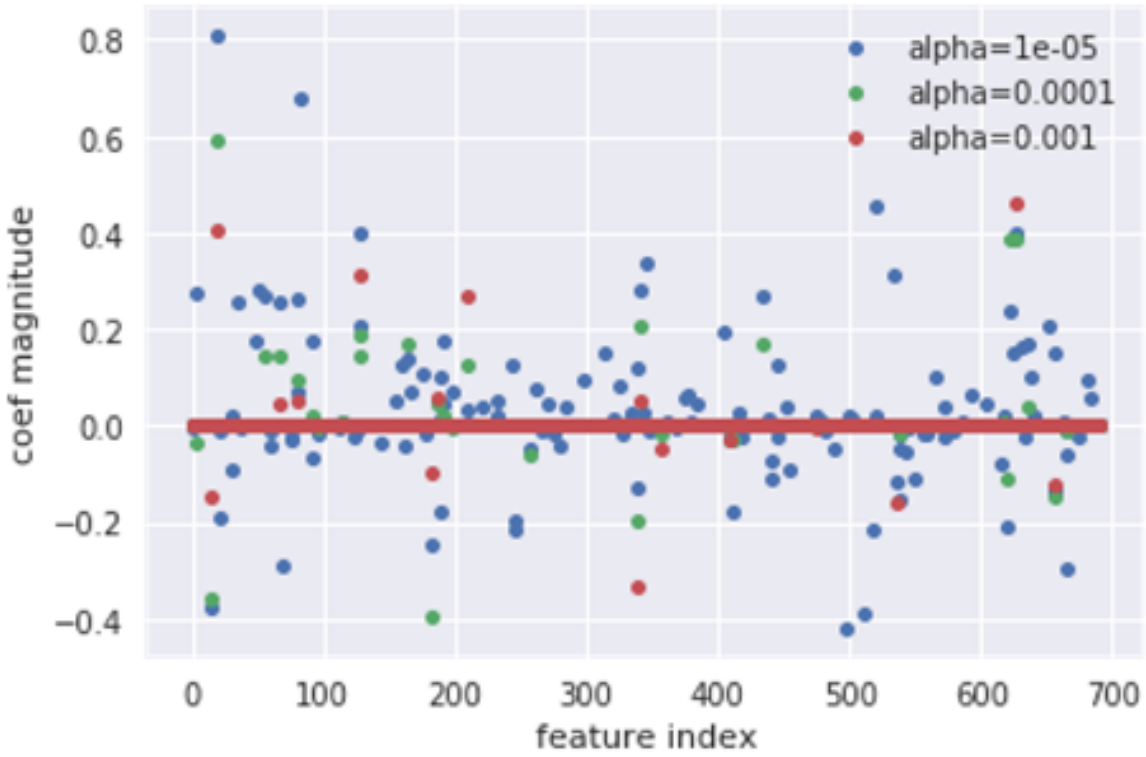


図 3-5: 単語抽出と L1 正則化

表 3-2: 正則化パラメータと単語抽出数と AUC の関係

Value of Regularization	10^{-5}	10^{-4}	10^{-3}
Extracted Words	151	35	17
AUC	0.934	0.934	0.933

表 3-3: 正則化パラメータと単語抽出数とAUCの関係

Performance Measures	Training Data	Validation Data
Accuracy	0.901	0.902
Precision	0.998	0.997
Recall	0.810	0.813

図 3-5 からは，正則化項の値が大きくなることで，罰則の値が多くなり，説明変数としての単語が減り，一部の単語で大きく推定されていた偏回帰係数の値が小さく評価されていることが確認できる．また，表 3-2 では，そのように正則化項の値を大きくして，単語数を減らしても，AUC の値は，殆ど同じ水準にあることが確認できる．そして，表 3-3 は，17 単語を用いたモデルについて，過学習は発生しておらず，頑健なモデルであることを示している．以上の定量的な評価は，Javascript コードの有無を特徴づける単語は 691 単語から 17 単語に絞り込めることを支持している．この評価に基づいて，抽出した 17 単語と推定した偏回帰係数を表 3-4 に示す．

表 3-4: ボットの特徴量を表す単語の偏回帰係数

Features of Bot			Features of User		
#	Word	Coef.	#	Word	Coef.
1	ubuntu	0.4605	1	like	0.3318
2	apple	0.4077	2	rv	0.1613
3	facebookexternalhit	0.3104	3	android	0.1438
4	http	0.2675	4	win	0.1242
5	go	0.0603	5	gecko	0.0975
6	com	0.0534	6	mac	0.0455
7	linux	0.0494	7	mobile	0.0313
8	chrome	0.0453	8	os	0.0056
9	applebot	0.0000			

ボット傾向のある単語として抽出された ubuntu, linux と, 非ボット傾向のある単語として抽出された android, win, mac を比べると, 一般的なユーザーが利用する情報端末のオペレーションシステムの単語が, 非ボット傾向の単語として選択されている. また, ボット傾向のある単語として選択された http は, 悪意のない検索エンジンからのボットが自身の身元を明示するためにユーザーエージェント情報に記載することが一般的にわかっている. これらの内容は, Javascript コードの稼働有無によってボットを判別するルールの有用性を支持している.

本研究では，マルチデータソースのアクセスログを用いることで，ウェブサイトへのアクセスの Javascript コードの稼働有無の判定を可能にした．そして，本節の分析結果は，Javascript コードの非稼働のアクセスがボットからのアクセスと判定することを支持した．これにより，本研究の提案するルールベース判定に，有用性があることを確認した．次節では，本節の分析ではユーザーと判定したアクセスに含まれるボットを，本節の分析で得られた知見を活用しながら，モデルベース判定を行う手法を提案する．

3.2.4 1クラスサポートベクトルマシンの学習

本節では，多くがユーザーからのものと考えられるJavascriptコード稼働のアクセスログから，異常検知モデルを用いて，さらにボットを取り除く．ユーザーまたはボットであるアクセス主体の特徴量には，前節の分析で得た知見を用いる．訪問の時間帯や1回の訪問あたりのアクセスしたページ数などの振る舞い情報に関する内容，およびユーザーエージェント情報から抽出した17単語の出現有無を，特徴量として投入する．

異常検知モデルには，伝統的な手法である1クラスサポートベクトルマシンを用いる．1クラスサポートベクトルマシンについて，本節では提案手法の理解のために最低限必要な内容を説明する．理論的な詳細および具体的な解法は文献[100][101][102]が詳しい．

提案手法の説明のために，先ずは2値分類を行うサポートベクトルマシンについて説明する．この手法は，教師あり学習に相当し，レコード単位に2クラスのラベルを付与した学習用データから2クラスの境界線とサンプルとのマージンを最大化するように境界線を導出する．例えば，アクセスログのレコードにボットを示すフラグを付与した学習データが入手可能であれば，この手法を選択することで，ボットとユーザーの境界線を示す判別モデルを構築できる．

しかし，教師あり学習に必要なレコード単位の2クラスのラベルが入手しづらい問題も存在する．本研究で取り組むアクセスログに潜むボットの判別は，まさにそのような問題である．ウェブサイトのアクセス量の時系列の変化などから，その日のアクセスログにボットが一定規模で混入

していることは解釈可能である．しかし，アクセスログのレコード単位にボットであることを示す証拠は存在しない．

そこで，本研究では，2 値分類を行うサポートベクトルマシンを応用した教師なし学習の手法である 1 クラスサポートベクトルマシンを採用する．この手法は，レコード単位の異常はわからないが．ある割合で異常レコードが混入していることがわかっている場合に，ハイパーパラメータで与えた割合の異常レコードとそれ以外の正常レコードとの境界線を推定する．そして，境界線との距離から．レコード単位に異常度を導出する．本研究では，その異常度を用いてレコード単位にボットを判定する．図 3-6 を用いて更に手法を説明すると，1 クラスサポートベクトルマシンは，高次元空間におけるサンプルを，異常のサンプルほど原点に近くなるように写像した上で，原点と 1 クラスの境界線とマージンを最大化する境界線を求めている．図 3-6 では，わかりやすさのために，高次元空間を 2 次元で簡易に表現している．今回の分析におけるサンプルとは，アクセスログのレコードを，前述のウェブサイト上の振る舞い情報および単語出現有無を用いて高次元ベクトル化したものを指す．マージンを最大化する問題は，図 3-6 におけるサンプルを囲む球の半径 R を最小化する問題と変換することができる．最小化問題を式 1 に示す．

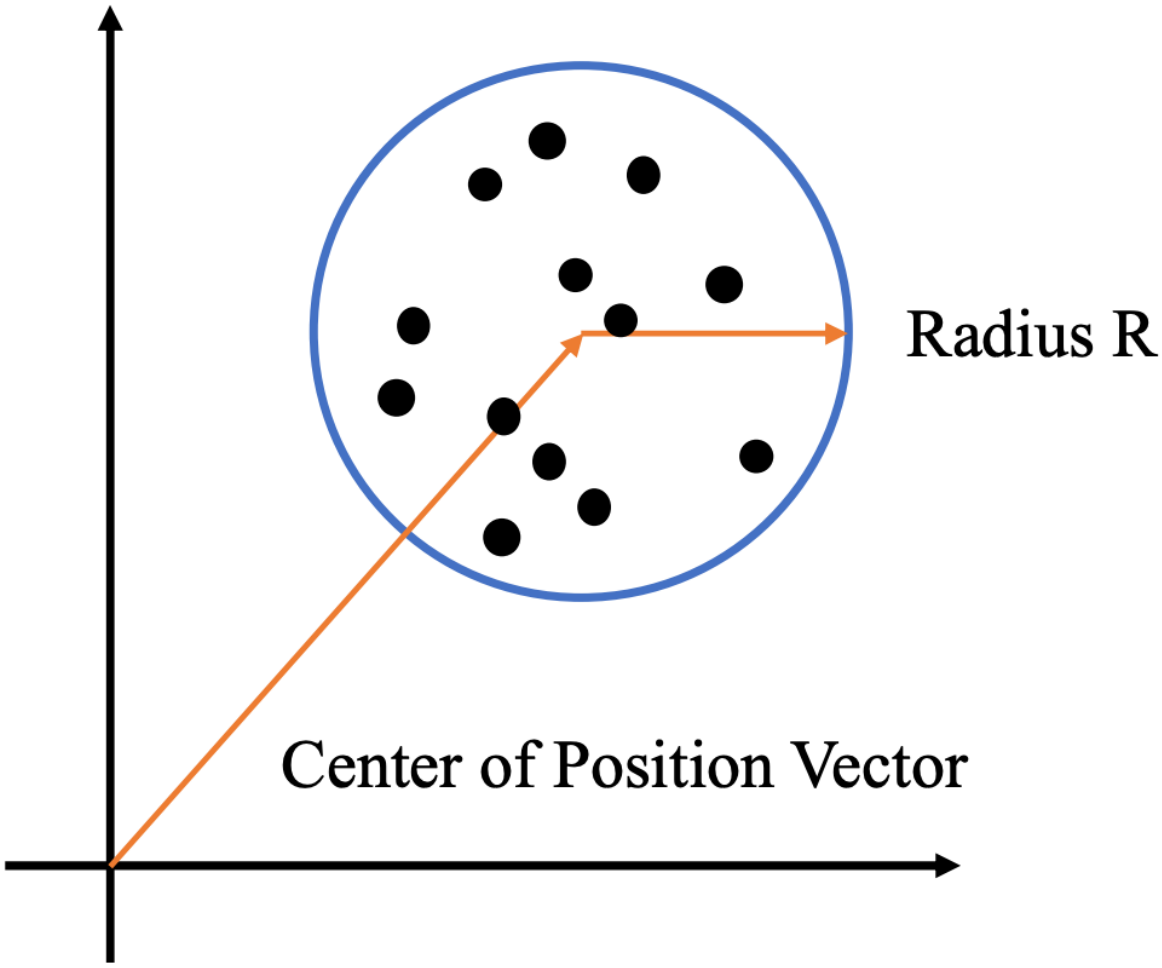


図　3-6: 1 クラスサポートベクトルマシンのイメージ

$$\min_{R^2,b,u}\left\{R^2+C\sum_{n=1}^{N}u^{(n)}\right\}\text{subject to}\left\|x^{(n)}-b\right\|^2\le R^2+u^{(n)}$$

……(1)

式 1 の最適化問題の解を(R^2,b^*,u^*)としたとき，サンプルの異常度は球からはみ出した長さとして式 2 のように定義できる．

$$a(x')=\left\|x^{(n)}-b^*\right\|^2-R^{2*}\quad\cdots\cdots(2)$$

それぞれのパラメータを説明する．先ず，b は球の中心の位置ベクトルを示す．u は，遊び，スラック変数などと訳されることが多いが，半径 R の球をはみ出したサンプルを許容する変数としての役割を持つ．式 1 において，u の総和に C を掛けた項が最小化問題へのペナルティ項として取り込まれていることがわかる．また，C は，サンプル全体における異常値の割合を決める定数で，1 クラスサポートベクターマシンにおける正則化定数と呼ばれる．1 クラスサポートベクトルマシンを世の中の異常値検知の問題に実用するためには，この定数は何らかの方法で決定しなければならない．この課題に対して，本研究は，実際に得られたデータへの判別モデルの当てはまりからハイパーパラメータを探索する，という手法を選択する．手法をより詳細に解説するためには，モデル評価を行うためのデータが必要であるため，モデル評価の説明部分で解説する．

本節では，ウェブサイトの振る舞いと単語の出現有無を用いてベクトル化したサンプルの異常度を，1 クラスサポートベクトルマシンを用いて算出する方法を説明した．そして，次章にて，本研究における評価用データの作成方法を示した上で，評価用データを用いて， 1 クラスサポートベクトルマシンによるボット検知手法の評価を行う．更に，同じ評価用データを用いて，実用のための課題として提起した正則化定数の決定方法を説明する．

3.3　ボット検知手法の評価

3.3.1　人による判別の再現の実験

本節では，人が実際にアクセスログの内容を目で確認して，ボットと判断した内容を評価用データとして用意した．そして，提案手法によって評価用データと同等の判定の再現が可能か確認することで，提案手法の定量評価を行った．

評価用データの作成のためのアクセスログは，1000 回の訪問を抽出し，アクセスログの内容を 5 人で，それぞれ目視を行うことで，ボットを判定した．その判定方法を説明するために，判定者が実際に確認したアクセスログの項目およびレコードを抜粋したものを表 3-5 に示す．

表 3-5: 実験のためのアクセスログのサンプル

#	Visit time	Duration of stay	Jsessionid Uniqueness	List Pageviews	Detail Pageviews
1	01:13	25.3	457	588	379
2	04:03	1,144.3	1	199	5,938
3	07:57	808.0	1	0	1,061
4	07:55	102.0	2	27	53
5	13:47	56.7	2	6	44
6	20:41	14	1	3	3

判定のための観点を説明する．表 3-5 の#1 から#3 のレコードは，いずれも目視によってボットとして判定したレコードである．#1 は，1 回の訪問の中で Cookie に記述するセッション ID を 457 回，自ら切り替えを行っている．何らかの意図をもって，別人からのページビューであることを装っているボットではないか，と疑われた．このような値は事前知識をもとに集計を行うことで，人間による判別を支援し，判別精度を高める工夫をしている．次に，#2 は，深夜 4 時から 1,144 分もの長時間，約 6,000 ページという膨大なページを閲覧していることからボットであることが疑われた．#3 は，住宅商材を一覧で紹介したページには 1 度もアクセスすることなく，住宅商材の詳細情報を示すページを 1,061 ページも閲覧している．詳細ページにたどり着くには，一覧ページに掲載されている詳細ページへのリンク URL をクリックする必要がある．何らかの手段で，詳細ページのリンク URL を入手，生成することで大量アクセスを発生させているボットであることが疑われた．以上のような疑いをもとに，判定者はボットのレコードに対して投票を行い，最終的に 1000 レコードのうち，228 レコードをボットとして判定した．全体の約 97%にあたる 221 レコードは全員一致でボットに投票された．投票結果の分かれた 3%については，IP アドレスの切り替えなど疑わしい要素はあるが，ページビュー数は 60 ページ程度で一般的なユーザーの量と変わらない，といったレコードが挙げられた．これらは，最終的に多数決を用いて判定したが，今回用意できたデータソースを用いた特徴づけにおいては，これ以上の精度での判定は困難であった．尚，表 3-5 の内容に相当するレコードを，本提案手法のモデルにベクトルとして投入している．

1 クラスサポートベクトルマシンの学習は，前節の分析で得たボットのユーザーエージェント情報に含まれる単語等を特徴量として追加しながら，複数のモデルを構築した．複数のモデルによる予測結果を，評価用データの判定内容と比較を行い，提案手法の有用性としての予測精度を示す．予測精度を評価した内容を表 3-6 に示す．

表 3-6: ボット検知モデルの性能評価

#	Patterns of Features	Precision	Recall	Accuracy
1	Pageviews	0.744	0.982	0.919
2	Pageviews, Behavior	0.833	0.982	0.951
3	Pageviews, Behavior, Words	0.953	0.982	0.985

ボットの閲覧したページ数を投入したモデル#1 に対して，先述の分析で発見した時間帯や訪問の間隔時間を投入したモデル#2，ボットに含まれる単語の出現有無を投入したモデル#3 が優れた予測精度を記録した．そして，いずれの精度指標においても，同率を含む 1 位を獲得したモデル 3 が最も効率良く人による判別を再現できる，という実験結果を得た．

更に，本実験にて，提案手法を用いて実際にボットの判定を行うにあたり，先述の課題として述べた，式 1 におけるハイパーパラメータである正則化定数 C を決定する方法を説明する．モデル#3 の正則化定数 C を変動させたときの精度変化の計測結果を図 3-7 に示す．

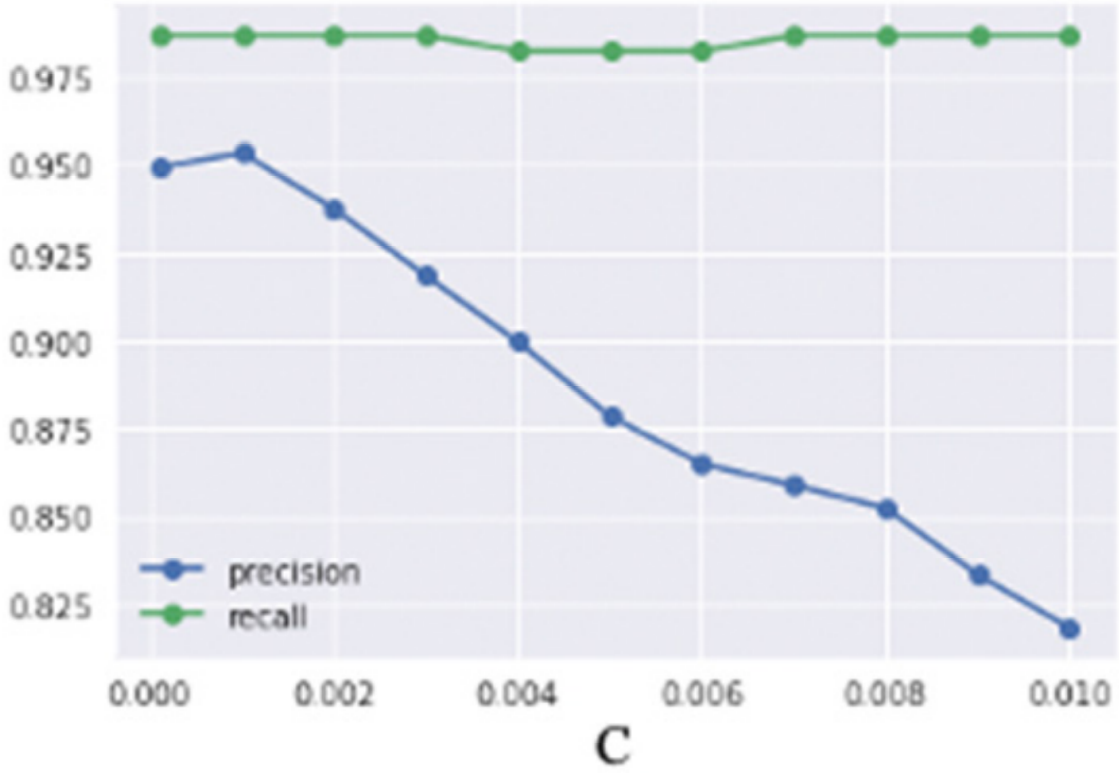


図　3-7: 1 クラスサポートベクトルマシンのパラメータチューニング

正則化定数 C を 0.001 よりも大きくすることで，評価用データにおける異常値を漏れなく異常と判定できた割合を示す recall 指標は変化していない．つまり，いずれの値の場合でも，異常値は異常値として漏れなく判定できている，と解釈できる．

一方で，異常値としての判定の正答率を示す precision 指標は下がり続けている．これは，C を小さくすればするほど，誤判定が増えていることを示している．recall 指標の解釈を踏まえると，異常値は既に殆ど全て検出できている状態から，C を小さくすることで，正常値を異常値として判定する規模が増えている，と解釈することができる．

今回のデータにおいては，0.001 が最も当てはまりが良いことが確認できた．C の値は，ウェブサイトにアクセスのあるボットの質と量によって適当な値が変動する．そして，ボットの質と量は，ウェブサイトの運営者にはコントロールできないため，本手法を実際に運用する場合は，定

期的な正則化定数 C の再探索によるモデルの再推定が必要になる．運用における課題は後述する．

尚，今回の実験データには 2%程度のインターネットボットが含まれていた．実験データにおける含有率が上昇すると，図 3-6 の 1 クラスサポートベクトルマシンの球の外の点の量を多く許容するようなモデル導出を行うことがわかっている．但し，実際のウェブサイト運用において発生するモデルの含有率の変動は，モデルの判別性能と関係がない．それは，あくまでも，モデル性能は，モデルの捉えているボットの特徴をもとに決定されるからである．つまり，モデル導出時に実験データに含まれていたボットと同質なボットが大量に訪問した場合，モデルの判別性能は維持できる．一方で，異質なボットが訪問した場合は，その量に関わらず，モデルの判別性能は維持できない．このようなモデルの特性に起因する運用の課題については，今後の課題を扱った節でさらに考察する．

3.3.2 レコメンデーションの予測精度向上の実験

本節では，アクセスログに潜むボットを検知し，ノイズとして除去することで，ユーザーの行動履歴からユーザーの嗜好を予測するレコメンデーションシステムの精度向上を実現する実験を行う．レコメンデーションシステムのアルゴリズムとして一般的に利用されている手法に協調フィルタリング[103]がある．商品そのものの属性を用いてレコメンデーションを行うコンテンツベースフィルタリング[104]に対して，協調フィルタリングは，ウェブサイトの全ユーザーの嗜好を用いてレコメンデーションを行う．もし，全ユーザーからボットを除くことができれば，協調フィルタリングによる学習の効率が上がり，レコメンデーションの精度向上が期待できる．

評価実験は，ユーザーの商品の閲覧有無を示す行列を，ボット除去前とボット除去後の 2 つを用意する．それぞれのデータは，ユーザー単位に更に 2 つに分割する．ユーザーの 80%を学習用の教師データ，残りの 20%のユーザーをモデル評価用の検証データとして分割を行い，商品の閲覧有無を予測する．検証データは，いくつかのユーザーと商品の組み合わせのデータを隠すことで，予測の正誤を確認する．予測に用いるアルゴリズムは，一般的な機械学習ライブラリである Scikit-learn[105]に用意された協調フィルタリングのパッケージ Scikit-surprise に採用されている全てのアルゴリズムを採用した．予測精度は RMSE[106]で評価した．評価結果を表 3-7 に示す．

表 3-7: レコメンデーション精度に与えるボット除去の効果

Algorithm	Before Bot Cleansing	After Bot Cleansing
User-based CF	0.149	0.129
Item-based CF	0.149	0.143
Matrix Factorization	0.144	0.138
Slope One	0.140	0.132
Co-Clustering	0.141	0.131

全てのアルゴリズムにおいて予測精度の改善が確認できた．今回の実験における改善規模は必ずしも大きなものではないが，協調フィルタリングの学習を妨げるノイズ規模に従って，改善規模は変動するものと考えられる．また，協調フィルタリングのアルゴリズムには，ノイズとなるようなユーザーの行動，例えば全ての商品を閲覧しているユーザーの影響を小さくするような工夫が行われている．今回，特定のアルゴリズムではなく，全てのアルゴリズムにおいて改善が確認できたことは，アルゴリズムの個別の工夫で対応しきれなかったノイズを除去できたことが，小規模ではあるが確認できた，と評価することができる．

3.3.3 業務効率改善の評価

住宅情報ポータルサイトで行われていた人の目による判別とボット除外作業を，本ボット検知手法を用いて自動化を行った．2021 年 12 月現在まで 2 年間運用した結果，1 ヶ月あたり 160 時間を要していたボット除外作業が，本研究が提案するモデルを実装したシステムによって完全に自動化することができている．

従来のボット除外作業は，表 3-5 で示したアクセスログの目視確認を行い，アクセスログを分析する前に，アクセスログのレコードを示す IP や Cookie 情報，ユーザーエージェント情報を指定して除外するプログラムを作成していた．アクセスログの目視確認，IP や Cookie 情報やユーザーエージェント情報のブラックリスト出力は，本研究が提案するモデルから自動出力が可能となった．自動出力したブラックリストを，データ分析の前処理として定常的に自動適用することで，人手の作業は不要となった．

但し，誤検知の監視として，実際にメールアドレスなどの情報を入力して，住宅情報に問い合わせを出したユーザーを，本システムがボットとして判定していたら通知する機能を開発して，運用している．運用開始以降，月に数件程度は発生しているが，集客投資効果の分析やリコメンデーションシステムの構築などを目的としたボット除外品質としては問題ない，と判断している．

3.3.4 手法の汎用性の評価

本節では，提案手法の持つ汎用性について説明する．提案手法は，分析用データの抽出方法，その後の分析方法ともに，今日のウェブサイトが一般的に用いている技術を用いて，特殊な技術の導入なしに実現できる手法である．具体的な説明として，提案手法の重要な特徴であるマルチソースアクセスログの取得手法について述べる．インターネットボット検知を目的に，本研究と同じくマルチソースアクセスログを取得するアプローチを探索した先行研究[32]では，アクセスログの一部をネットワーク機器から抽出している．今日のウェブサイトの構成要素として，ファイヤー・ウォール，ルーター，スイッチといったネットワーク機器を含まないウェブサイトは存在し得ない．よって，マルチソースアクセスログの一部としてネットワーク機器を活用する手法には一定の汎用性が認められる．しかし，今日のウェブサイトの運営事業者の置かれている技術動向を踏まえると，先行研究には汎用性の観点の課題がある．

近年，クラウド・コンピューティングの普及を受けて，ウェブサイトを運営する事業者は，自社専用のネットワーク機器を保有する必要性が下がってきている．ネットワーク機器は，ユーザーとウェブサイトの間の通信機能を実現する重要な機器であるが，その操作は米国大手のシスコ・システムズ社の IOS に代表される製造事業者の独自のオペレーションシステム上で行う必要がある場合が多い．そのようなオペレーションシステムは，今日のウェブサイトのウェブサーバーやデータベースサーバなどに汎用的に用いられている Linux 等とは操作方法が異なる上に，用途は通信経路の制御に専門に用いられている．このように手法

を構成する技術においては，ネットワーク機器の操作に依存する先行研究の手法は，汎用性の観点で課題があるといえる．一方，提案手法が，マルチソースアクセスログとして提案しているウェブビーコンログ方式はウェブページの HTML ファイル，ウェブアクセスログ方式は Linux 等の汎用的なオペレーションシステム上で稼働させられるウェブサーバーを操作することになる．HTML ファイルに関しては，ウェブサイトの画面の定義情報そのものであり，この資材を保有，操作しない事業者は想定しづらい．ウェブサーバーに関しては，先述した Linux 等の汎用的なオペレーションシステム上で稼働する Apache HTTP Server[107]に代表されるウェブサーバーを構成するミドルウェアを操作することになる．この操作も，ウェブサイトを起動させる作業そのものに相当するため，この処理を操作しない事業者も想定しづらい．このような技術で構成した提案手法は，ウェブサイトを運営する事業者にとって扱いやすい技術を用いた汎用的手法と評価できる，と考えている．

3.3.5 今後の課題

本節では，提案手法をウェブサイトの運用に実用し，ウェブサイトのユーザーの分析の精度を高めていくにあたっての今後の課題について説明する．実用化にあたっては，ウェブサイトのアクセスログ全体に占めるボット含有率の監視が重要になる．今回の実験データの含有率 2% は永続的に続くものではなく，攻撃者の行動によって大きく変動しうる．そして，実際には，真のボット含有率そのものは計測できない，という問題にも対処しなければならない．そのため本手法の実用にあたって，誤検知の対策に有効な対策は，予測したボットのアクセスログ全体における含有率の時系列変化の監視である．日々の予測ボットの含有率を分布として計測し，含有率にも異常検知のアプローチを行うことで，攻撃者の行動変化，モデル再構築の必要性について検知することができる．

また，ボットと判定したアクセス主体が，個人情報の入力といった人間でしか発生しない行動を取った場合は誤検知と判定し，その検知規模の時系列変化を確認する手法も有効な可能性がある．人間でしか発生しない行動とはどのような行動か探索することは，本研究が対象とするインターネットボットの検知に重要な知見の獲得に繋がるため，このような知見の収集は今後の研究課題となる．

これらのモデルの関連指標に大きな変化が現れた場合は，ボット判別モデルの再構築を検討する必要がある．本手法を実用する上で，ボット判別モデルの更新頻度は重要な検討対象になるが，同一サイトのアクセスログに含まれるユーザーとボットの振る舞いや属性情報の傾向

が，短期間で大きく変わることは想定し辛い．だが，その傾向を，未来永劫に不変であると仮定して，モデルの更新を想定しない運用を採ることは明らかに不適切である．このようなモデル再構築のタイミングの問題に対しては，先述のボット含有率やアクセスログの規模を常に時系列で監視することで，傾向の変化を検知し，モデルの更新の判断根拠とすることが，有効なアプローチと考えられる．仮に，ユーザーとボットの質的な特徴の変化が頻繁に発生する場合は，判別モデルをそれに合わせて高頻度に更新する必要があるだろう．そのような問題に対応するためは，本論文で示した手法には拡張が必要である．特に，特徴量の生成，1 クラスサポートベクターマシンのハイパーパラメータ探索，頑健性の評価といったモデル構築の手続きについて，作業効率向上，作業の自動化といった追加的な研究が必要である．

3.4 おわりに

本研究では，ウェブサイトの運営におけるアクセスログの管理，利活用を支援するためのボット検知の手法を提案した．アクセスログに潜むボットの検知，除去に焦点を当て，アクセスログを複数の方式で取得することで，ボットの特徴を発見した．この特徴を利用したルールベースと伝統的な異常検知手法を組み合わせた新たな手法を提案した．提案手法を用いれば，人手によるボットの判別と同程度の精度で，判別を自動化することが可能となる．人手によるボットの判別は，日々発生する膨大なアクセスログ全体に適用することは難しいため，この自動化には有用性があると言える．さらには，レコメンデーションシステムの予測精度においても，改善の効果が確認できた．

これらの検証結果は，住宅情報ポータルサイトに活かすために，当該事業を営む企業の情報システム上の実験を行った．検証によって，ウェブサイトを構成する汎用的な技術を組み合わせることで，高い精度のボット判別が実現できることがわかった．当該事業ドメインに特化した新たな知見は得られなかったが，情報提供を行うウェブサイトに広く有効な汎用的な手法が導出できた．

以上のように，本論文で提案したマルチデータソースの分析によるボット検知手法は，ウェブサイト運営において，アクセスログに含まれるボットの自動識別が実現できることを確認した．一般に，データの分析には，分析の目的に合致したデータを用意することが非常に重要である．ウェブサイトのユーザーをより深く知り，ウェブサイトの改善を検討するための分析用データに，ユーザー以外の情報が含まれていることは決して望ましい状態とは言えない．そこで，提案手法を用いることで，ウ

ェブサイトのユーザーを分析する目的に合致した，ユーザーのアクセスログに絞り込んだ分析用データを自動で提供することが可能になる．

今後は，前節にて扱った今後の課題，特にインターネットボットの質の変化に柔軟に対応するためのモデル再導出の効率化，自動化といった手法の探索に取り組んでいく所存である．

第4章　文字列照合アルゴリズムの自動テストシステムへの適用

本章では，住宅情報ポータルサイトの行動データのノイズとなる不正識別子を除去する新たな手法を提案する．不正識別子の除去作業をテスト作業と捉え，テスト作業の効率を改善する自動テストシステムの実現方法を探索する．提案する自動テストシステムの実現方式，および共同研究を行った連携先企業の情報システム上で行った作業効率の改善検証の結果を報告する．

4.1　はじめに

住宅情報ポータルサイトの行動データには，前章で取り扱ったインターネットボットだけでなくウェブサイト開発時の作業ミスなど，様々な要因によって，データ分析の目的の妨げになる不正な形式の識別子がノイズとして混入する．ウェブサイト開発の作業ミスでいえば，その背景には，住宅情報ポータルサイトにおける重要なマーケティング活動であるユーザビリティ向上のための継続的インテグレーション[108]がある．具体的には，住宅検討者が好みの住宅を探しやすい，使いやすいサイトを探求し，開発を継続する活動である．住宅情報ポータルサイトに限らず，今日のウェブサイトの運営事業者は，このような活動に取り組んでいるため，ウェブサイト開発の作業ミスがなくなることはない．よって，作業ミスによって生まれる行動データの不正識別子は発生し続ける．行動データの識別子は，データ分析において，ユーザーがウェブサイトで接触するコンテンツや，ユーザーが取った行動の内容を，一意性をもって識別するために設計される文字列である．より具体的には，文字列のルールは，連続する特定文字数のアルファベットや数字，その羅列の組み合わせとして，設計される．例えば，商品の識別子が，このような文字列のルールを逸脱すると，商品を分析するためのプログラムが正しく動作しなくなるため，ユーザーがどのような商品を併行検討したのか，といった行動データが分析できなくなる．そして，そのような不正識別子の除去作業は膨大な量になり，今日の事業者の負担となっている．行動データの不正識別子を効率的に除去するシステムがあれば，このような課題を解決できる．

具体的なアプローチとしては，対象作業をテスト作業と見立て，自動テストシステムを提案する．自動テストシステムは，多くの先行研究が存在しているが，それらに共通していることは，自動化する対象作業を，システム化が可能な繰り返し作業に絞り込んでいることである．自動テストシステムの研究におけるその代表的な作業は，テストケースの作成とテストコードの作成である．提案手法は，行動データからの不正識別子の除去作業のためのテストケース作成およびテストコード作成に焦点を絞った．テストケース作成に対応する GUI システム，テストコード作成に対応する文字列照合アルゴリズム，これらを組み合わせた新たな自動テストシステムとして設計した．本章では，提案手法の実現方式を報告するとともに，提案手法の開発にあたって共同研究を行った連携先企業の情報システムの上での実用性の検証結果を報告する．

4.2　自動テストシステムの機能

行動データの不正識別子の除去に対応する自動テストシステムの検討にあたって，実現手法を探索すべき要素は大きく 2 つに分けることができる．1 つは，テストケースの作成，もう 1 つはテストコードの作成である．それぞれについて提案手法を解説する．

4.2.1 テストケース作成の GUI システム化

ソフトウェアテストにおいて，テストケースとはどのような場面でどのような作業を行うか，を定義したものである．通常は，テスト作業者に期待する作業を文章で表現することが多い．先行研究[53]では，指定のウェブブラウザを立ち上げて，指定のウェブページをロードし，ロードにかかった時間を計測する，といった指示相当の文章がテストケースに該当する．

行動データの不正識別子の除去作業においては，ウェブページのどのページで，どの識別子が，どのような文字列ルールに合致した形で生成されているか確認する，といった指示文書がテストケースに該当する．このようなテストケースを作成するにあたっての課題は，ウェブページ，識別子，文字列ルールの数が膨大になってしまうために，文章として抜け漏れなく記述，管理するための作業が困難になる．膨大な作業に長い時間がかかる，という量の問題だけでなく，定義すべき内容に抜け漏れが起こりやすい，という質の問題も合わせて解決しなければならない．この課題に対応するために，予め設計された画面遷移プロセスに従って作業者を誘導しながら作業を進める GUI システムが有効なアプローチとなる可能性がある．このアプローチを，行動データの不正識別子の除去作業にあてはめて，設計した画面と機能の組み合わせを表 4-1 に記載する．

表 4-1: 自動テストシステムの画面と機能の組み合わせ

画面	機能
テスト要件登録	ページ操作登録
	変数登録
	文字列パターン登録
テスト実行指示	ページ操作実行
	変数出力結果収集
テスト結果確認	テスト結果確認

行動データは，ウェブページの操作に合わせて生成されるため，テスト要件登録の画面では，まずどのようなページ操作のときの作業であるか対象を特定する．生成された行動データには HTTP ヘッダー情報相当の大量の変数が分析用の識別子として記録され，それぞれの識別子をチェック対象として登録を行う．登録した識別子には，チェックすべき文字列パターンが存在するため，その内容もテスト要件として登録する．以上が，テスト要件の登録である．この作業を一度抜け漏れなく登録できれば，以後は抜け漏れなくテスト要件が実行されることを担保できる．

次に，テスト実行指示の機能では，登録しておいたページ操作を実行する．この実行は，繰り返し行うことを想定している．一般的な自動テストは，過去に実行したものと同じテストをなるべく効率的に再現する構造をとっており，提案手法も同じ構造である．そして，テスト結果確

認では，テスト要件どおりの行動データが生成されたか，差分はどのようなものだったか，確認を行っている．

テスト要件登録の画面イメージを，図 4-1 に記載する．

図 4-1: 行動データのテスト要件登録の画面イメージ

テスト要件登録の画面設計では，作業者が迷うことなく各識別子の名称を確認し，必要なチェック内容をタブから選択できるような機能の提供を試行した．チェック内容をタブから選択すると，予め登録したテストコードが呼び出される実現方式をとっているが，テストコードの内容については，次項 4.2.2 にて説明する．次に，テスト結果確認の画面イメージを図 4-2 に記載する．

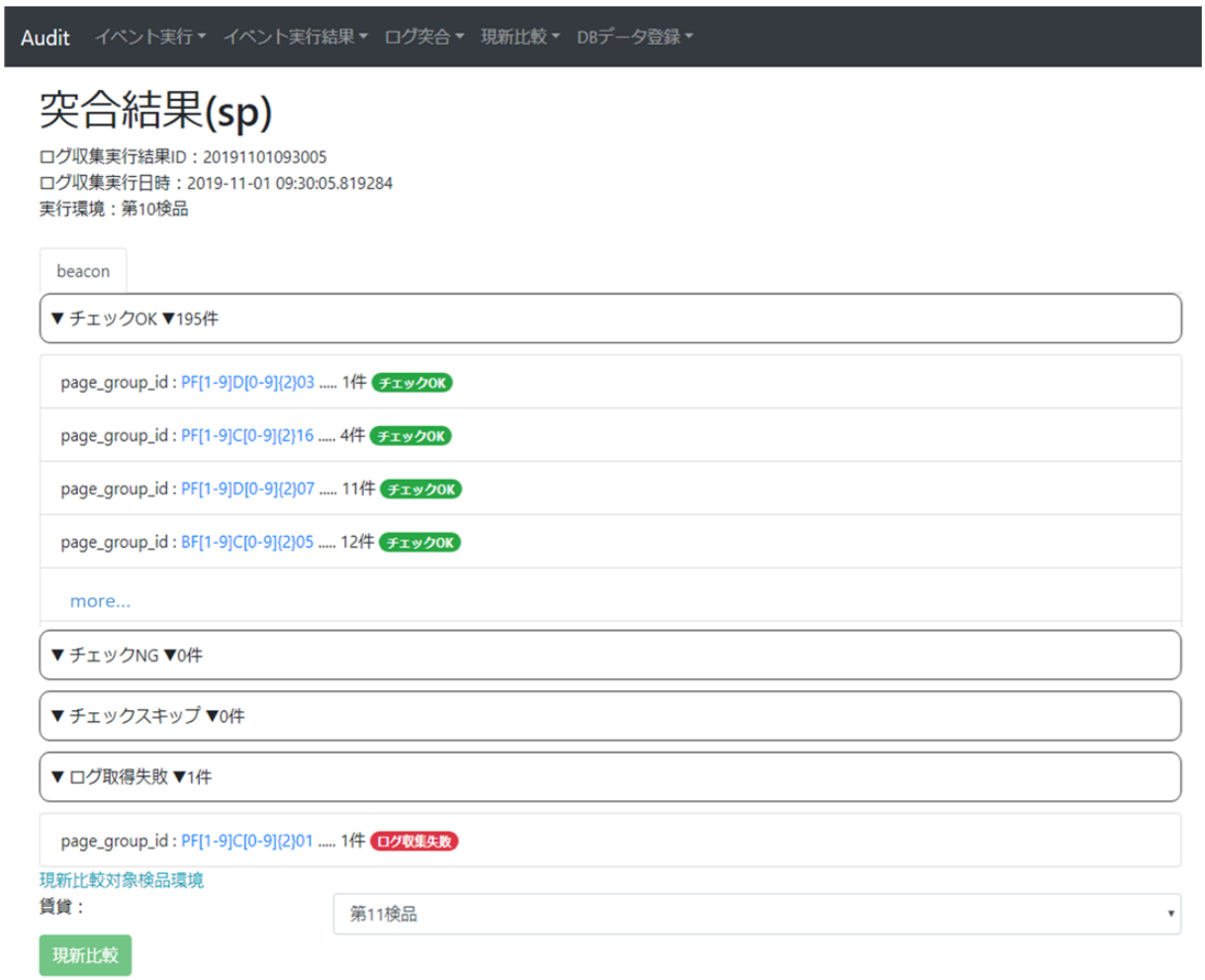


図 4-2: 行動データのテスト結果の確認画面

テスト結果の確認画面では，エラーとなった識別子を不正識別子として GUI システム上で見逃すことがないような表示方法を試みている．また，エラーになった内容が，具体的にどのような文字列ルールに違反したのか，GUI システムの操作によって特定できるような設計を試みている．このような手法を検証することで，利用者に特別なスキルを要求せず，システムの操作に慣れれば誰でも作業が実施できる簡便性の実現を試行する．

4.2.2 文字列照合アルゴリズムを活用したテストコード作成の効率化

本項では，テストコードの自動作成の手法について説明する．実現手法を説明するために提案手法を実現するためのシステムスタック図を図 4-3 に示す．

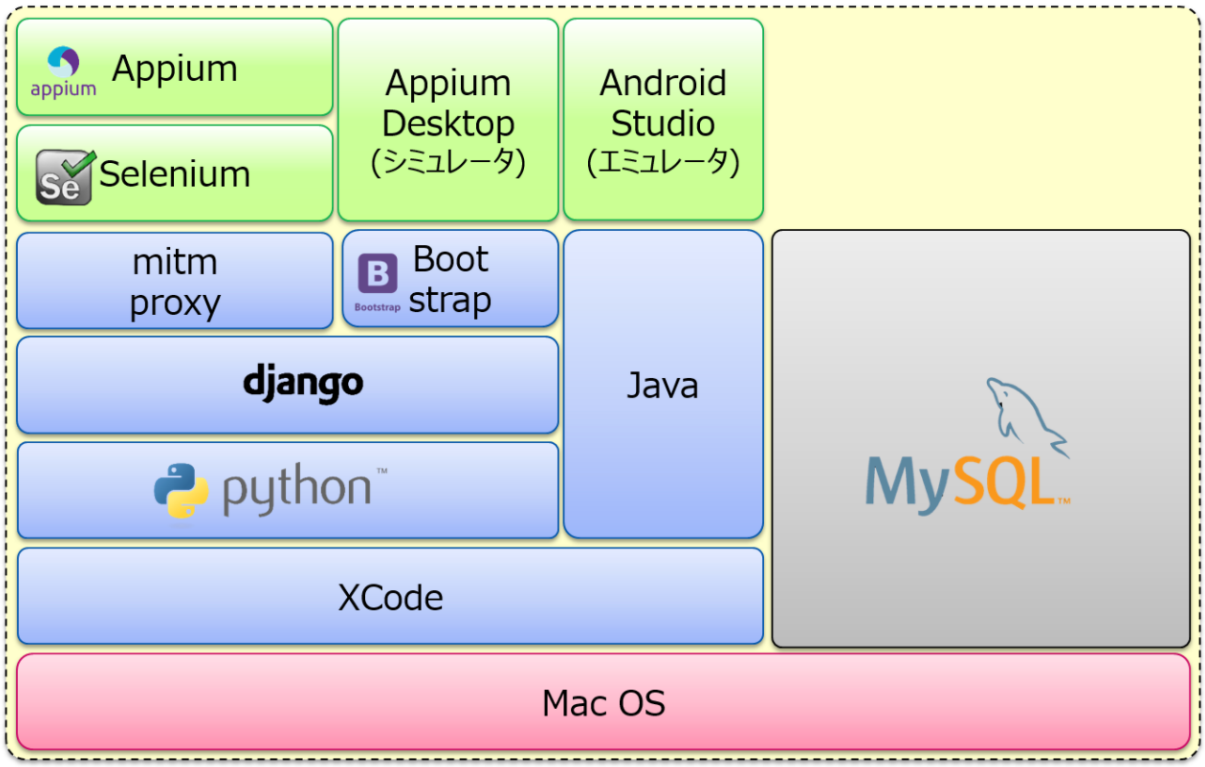


図 4-3: 自動テストシステムのシステムスタック図

テストコード作成に関連の深い技術要素を説明する．システム層の上流にある Selenium[109]は，ウェブサイトの操作を Python プログラムで効率的に記述するためのソフトウェアフレームワークである．自動テスト

システムに関する多くの先行研究が本コンポーネントを活用しているが，提案手法も同様に Selenium の活用の可能性を探索した．

Selenium の稼働内容を定義する実行パラメータをリレーショナルデータベースである MySQL に格納し，Selenium を使ったプログラムから参照することでテストコードとして完成させて，稼働させる方式を検討した．このような方式を採用することで，テストコードの作成作業を図 4-1 のテスト要件作成画面で完結させられる設計を試行している．また，この実行パラメータは，図 4-1 のテスト要件登録画面で入力された情報をデータベース格納用に変換したものである．データベースに保存された Selenium の実行パラメータを表 4-2 に示す．

表 4-2: 行動データのチェック内容を定義した Selenium 用実行パラメータ

Page Name	Event Name	Action of Selenium	Correct Logs Specification
Detail Info	Landing Page	1) driver.get("target URL")	Page_ID=PF[1-9]C[0-9]{2}02&wkat=TPV&templateId=SP_SUUMO&…
Detail Info	Impression Image	1)driver.get("target URL") 2)document.querySelectorAll('{target_DOM}')[0]. scrollIntoViewIfNeeded()	Page_ID=PF[1-9]C[0-9]{2}02&wkat=TPV&templateId=SP_SUUMO&event_data=ImgImp01&…
Detail Info	Impression Button	1)driver.get("target URL") 2)document.querySelectorAll('{target_DOM}')[0]. scrollIntoViewIfNeeded()	Page_ID=PF[1-9]C[0-9]{2}02&wkat=TPV&templateId=SP_SUUMO&event_data=Impbutton01&…
Detail Info	Click Button	1)driver.get("target URL") 2)document.querySelectorAll('{target_DOM}')[0]. scrollIntoViewIfNeeded() 3) {target_DOM}.click()	Page_ID=PF[1-9]C[0-9]{2}02&wkat=TPV&templateId=SP_SUUMO&event_data=clickbotton01&…

行動データの識別子は，特定文字数のアルファベットや数字の羅列，その組み合わせである．そのような文字列に対して照合処理を効率的に行うためには様々なアルゴリズムが提案されている[110].そのようなアルゴリズムを容易に実装可能な手法として，正規表現プログラム[111]を Selenium の実行パラメータとして受け付けられる設計を提案している．正規表現プログラムの文字列照合を表した状態遷移図を図 4-4 に示す．

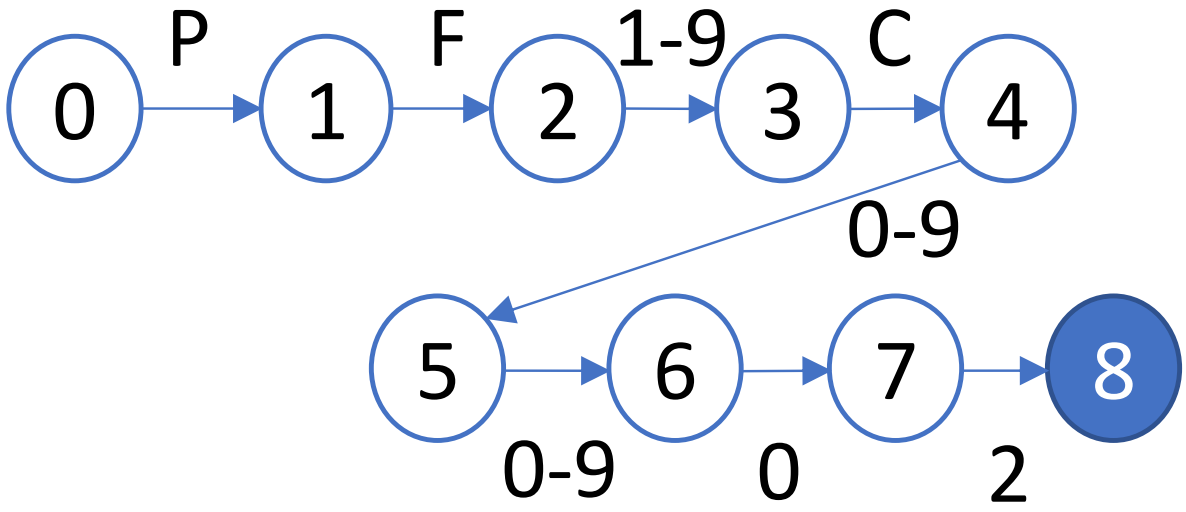


図 4-4: 正規表現プログラムを変換した状態遷移図

この文字列照合アルゴリズムは，表 4-2 の 4 列目に例示されている正規表現プログラムで記述された文字列ルール”PF[1-9]C[0-9]{2}02”を表したものである．処理内容を簡単に説明すると，1 文字目は P，2 文字目は F，3 文字目は 1 から 9 までの数字 1 桁，4 文字目は C，5 文字目と 6 文字目は合わせて 0 から 9 までの数字 2 桁，7 文字目は 0，8 文字目は 2，これで文字列が終わっている場合に True を返し，それ以外は False を返す照合処理である．このような記述方法は，数あるプログラミング技術の中でも習得しやすい技術であり，この技術を習得すれば提案手法を用いて，行動データの不正識別子を除去するテストコードは作成できる．このような技術方式によって，事業者にとって扱いやすい，操作の簡便な自動テストシステムの実現を試行した．

4.3　行動データの自動テストシステムの導入効果

本研究の提案するシステムの効果を確認するために，不動産広告サイトの機能の開発における行動データのテストの作業時間を計測する実験を行った．実験は下記の 3 パターンのテスト手法を比較した．

(a) すべて手動でテスト
(b) Selenium による自動テスト
(c) 提案システムによるテスト

(a)は，全てのテストの作業を手動で行う．(b)は，一般的な自動テストとして Selenium を用いてテストコードを手動でコーディングし，自動テストを行う．(c)は，本研究が提案するシステム上でコーディングを行うことなく，テストを実行する．図 4-3 に作業フローを示す．

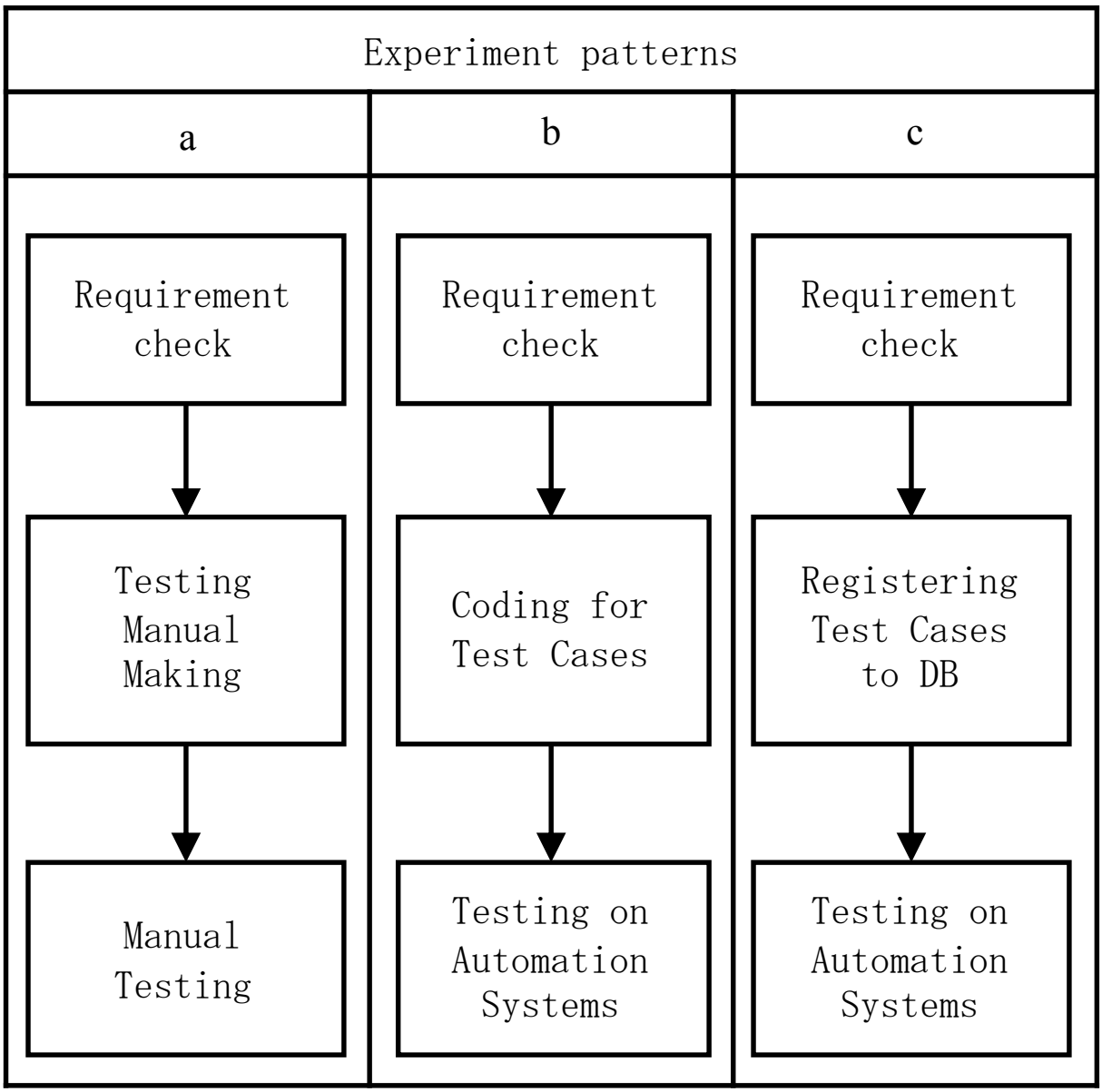


図 4-3: 行動データのワークフローの比較

実験は，作業フローの 1 番目の作業の要件確認を除く 2 つの作業を，3 つのテスト手法を用いて，それぞれ別の作業者が実行した．また，実験は，異なる機能開発を対象に 3 回実施した．結果を表 4-2 に示す．表の単位は時間数である．

表 4-2: 自動テストシステムの実験結果

Pattern	1st Trial		2nd Trial		3rd Trial	
	Work2(Testing Manual Making, Coding for Test Cases, Registering Test Cases to DB)	*Work3 (Manual Testing, Testing on Automation Systems)*	*Work2(Testing Manual Making, Coding for Test Cases, Registering Test Cases to DB)*	*Work3 (Manual Testing, Testing on Automation Systems)*	*Work2(Testing Manual Making, Coding for Test Cases, Registering Test Cases to DB)*	*Work3 (Manual Testing, Testing on Automation Systems)*
a. Manual Testing	1.0	1.2	1.0	1.2	1.2	1.5
b. Automated test with Selenium, but without proposed system	6.5	1.0	4.3	1.0	8.0	1.0
c. Automated test with Selenium and proposed system	0.1	0.1	0.1	0.1	0.2	0.2

テストコードのコーディングとテストの実行をシステム上で作業したｃが，最も短時間で作業を行うことができた．考察として，本研究が対象とする行動データのテストは，ボタンのクリックや画像の表出，コード値のチェックなど，パターンが限られるため，システム上での定型機能として用意し，操作させることに適していた．また，ｂのようなテストコードのコーディングの場合は作業者のコーディングスキルの差が作業効率に影響しやすいことに対し，提案システムでは，システム上の操作となるため作業者による差が生まれにくい，と想定できる．実験結果から，本研究が提案するシステムが，ウェブ行動データのテストの効率化にとって有効であることを確認した．

4.4 おわりに

本研究では，行動データの不正識別子を効率的に確認するための自動テストシステムを提案した．従来の手動テスト，および Selenium を用いた一般的な自動テストよりも，GUI システムによるテストケースとテストコードの自動作成を実現した提案システムが，作業効率を大幅に改善できることを確認した．

また，提案手法は，作業者のプログラミング技術になるべく依存せずに作業を完了できる簡便なシステムとして設計を行った．共同研究を行った連携先企業では，実際にプログラミングスキルの低いエンジニア層を作業担当者に据えることができており，提案手法の信頼性を支持する検証を行うことができた．

第5章　不均一データ分析アルゴリズムの優良顧客抽出への適用

本章では，スーパーマーケットチェーンの行動データにおける大量の一般顧客から少数の優良顧客を抽出する問題について，伝統的な顧客のランク付け手法の RFM 分析とテキストマイニングアルゴリズムを組み合わせた新たな手法を提案する．スーパーマーケットの業界では，優良顧客に絞り込んだマーケティング施策を行うにあたって，大量の一般顧客の行動データが，分析のノイズになってしまう問題がある．スーパーマーケットのマーケティング活動では，価値の高い優良顧客に対する特別な宣伝活動として，豪華なダイレクトメッセージや値引きクーポンなどを送る施策が行われている．スーパーマーケット事業の特徴として，20%の優良顧客が 80%の売上を生むことが通説となっており，優良顧客を効率良く分析するためには多数の一般顧客の行動データがノイズになる．提案手法が，先行研究と比べて高い精度で優良顧客を抽出できることを示す．

5.1 はじめに

スーパーマーケットチェーンでは，少数の優良顧客のデータに対して，より詳細な分析を行い，優良顧客に絞った広告宣伝費の投資を行うことは，マーケティング活動として重要である．顧客のランク付けには，有用な先行研究として RFM 分析が存在する．だが，RFM 分析は，様々な事業に適用可能な汎用性を持つ一方，わずか 3 つの指標でランク付けを行うため，精度の課題がある．より高い精度で優良顧客を自動抽出することができれば，多数の店舗を運営するスーパーマーケットチェーンの広告投資の効率を上げることができる．

5.2 優良顧客の抽出

本研究では，将来の店舗売上の大部分を生み出す顧客を優良顧客として定義し，現在の顧客情報からそのような優良顧客を予め抽出するモデルを提案する．

5.2.1 分析データ

本研究では，18 店舗を含む 1 つのスーパーマーケットチェーンの ID-POS データ 2 年分を利用する．2 年分のデータは，前半 1 年を現在の顧客情報，後半 1 年を将来の顧客情報として分割して利用する．又，本研究では，抽出モデルに投入する顧客の特徴量として，ID-POS データに予め設定された 167 品目の類型化された商品情報を利用する．この対応により，JAN コード等を利用した個別商品単位で分析を行う際に発生するデータのスパース(疎)性の問題に対処する．表 5-1 と表 5-2 に商品情報のマスタを示す．

表 5-1: 商品目マスタ(農産，水産，畜産)

農産	水産	畜産
果菜	丸物	和牛
葉菜	切身	国産牛
茎菜	貝	豪州産牛
根菜	鮮魚盛合せ	米国産牛
山菜	鮮魚_他	輸入牛
豆類	冊類	通常牛
きのこ	刺身	牛肉関連
発芽野菜	たたき	牛肉_他
妻物	水産生食	銘柄豚
野菜盛合せ	刺身類盛合せ	国産豚
野菜関連	刺身類_他	輸入豚
野菜_他	ボイル魚	通常豚
野菜水煮	冷凍魚	豚肉関連
冷凍野菜	味付魚	豚肉_他
カット野菜	漬魚	銘柄鶏
野菜加工品_他	塩蔵	国産鶏
季節果物	干物	輸入鶏
輸入果物	小魚	通常鶏
果物盛合せ	魚卵	鴨肉
果物関連	海草	鶏肉関連
果物_他	塩干加工品_他	鶏肉_他
冷凍果物		羊肉
カットフルーツ		馬肉
果物加工品_他		ひき肉
		内臓肉
		畜産生食
		精肉盛合せ

		精肉類_他 味付肉 加工肉 鶏卵 精肉加工品_他

表 5-2: 商品目マスタ(食品，惣菜，嗜好食品，その他)

食品	惣菜	嗜好食品	その他
食用油	揚物半惣菜	製菓材料	たばこ
香辛料	煮物半惣菜	ゼリー・プリン	花類
基礎調味料	焼物半惣菜	アイスクリーム	ギフト・銘菓
加工調味料	和風半惣菜	冷凍菓子	テナント
米飯調味料	洋風半惣菜	半・生菓子	カウンター
スプレッド・ディップ	中華半惣菜	乾菓子	その他_他
トッピング	スナック半惣菜	つまみ菓子	
調味料_他	半惣菜セット物	菓子関連	
粉類	半惣菜_他	菓子_他	
米	揚物惣菜	嗜好飲料	
餅	焼物惣菜	乳系飲料	
麺類	蒲焼惣菜	野菜・果実飲料	
皮生地	和風惣菜	清涼飲料	
パン	洋風惣菜	飲料_他	
シリアル	中華惣菜	ビール類	
穀物類_他	スナック惣菜	リキュール類	
農産乾物	サラダ惣菜	ワイン	
水産乾物	惣菜盛合せ	洋酒	
ドライフルーツ	惣菜_他	日本酒	
乾物類_他	米飯惣菜	焼酎	
乳製品	寿司惣菜	ノンアルコール飲料	
漬物	麺惣菜	酒関連	
水物	パン惣菜	酒類_他	
練物	弁当・セット物		
煮豆・佃煮	弁当_他		

農産加工品 水産加工品 畜産加工品 加工食品_他 即席麺 即席汁物 レトルト惣菜 レトルト米飯 冷凍食品 その他食品 即席食品_他			

5.2.2 優良顧客の定義

スーパーマーケット業界では，顧客の 20%から 30%といった限られた層が店舗の 80%の売上を生む，と言われている．そのような少数の顧客を，デシル分析や RFM 分析を用いて，予め優良顧客として抽出し，ダイレクトメールや割引クーポン等の広告宣伝費を集中させる施策が行われている[90]．本研究の分析対象スーパーマーケットチェーンにおいて，同様の状況が当てはまるか確認するため，前半 1 年分のデータを用いてデシル分析を行った．デシル分析とは，顧客を購入額の多い順に 10 分の 1 ずつのグループに分けることで，グループごとの総購入額への影響を調べる手法である．結果を図 5-1 に示す．

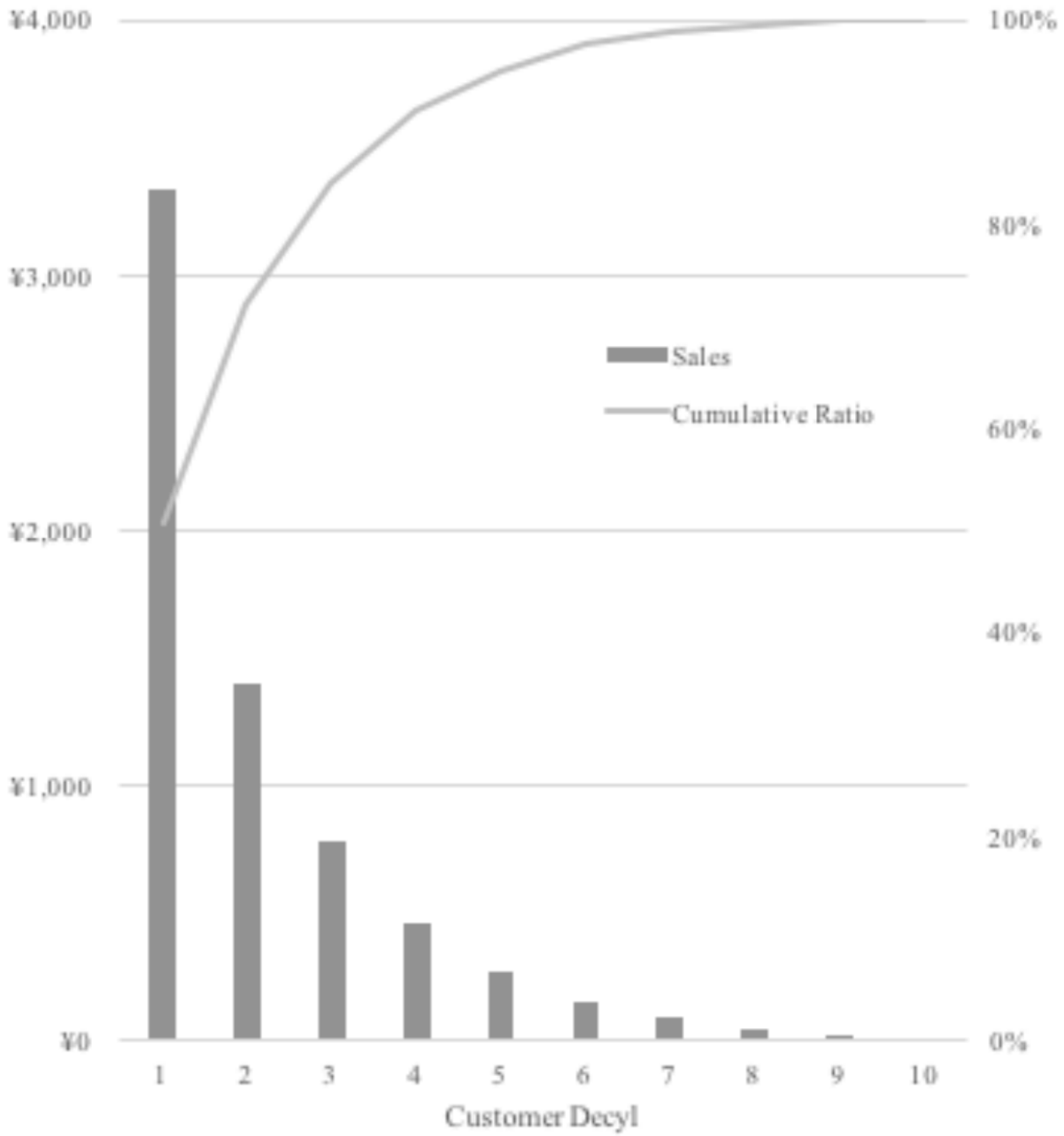


図 5-1: 顧客のデシル分析とパレート図

分析対象のスーパーマーケットチェーンにおいて，顧客デシルの上位 3 グループで売上の 84.0％を生み出していることが分かった．そこで，本研究における優良顧客は，上位 3 グループに該当する顧客と定義する．

5.2.3 優良顧客抽出モデル

本研究の抽出モデルには，伝統的機械学習手法であるロジスティック回帰を用いる．数ある機械学習手法の中からロジスティック回帰を採用する理由は，優良顧客を顕著に特徴づける商品目の情報を得るためである．ロジスティック回帰の利点のひとつに，モデルに投入した説明変数が目的変数に対してどの程度寄与したかを偏回帰係数として定量的に把握できることが挙げられる．

本研究ではその利点を活用する．また，優良顧客抽出モデルの目的変数には，後半 1 年の顧客デシル上位 3 グループであることを示すフラグを用いる．そして，説明変数には RFM 分析の 3 指標と購入商品目を用いる．

顧客の特徴量表現として購入商品目を用いるにあたり 2 つの手法を提案する．そのうち 1 つは店舗の異質性を表現した手法である．
まず，1 つめの提案は，顧客の総購入点数に占める購入商品目の割合表現である item frequency スコア(以下 if スコア)として式(1)として定式化する．

$$if_{i,c} = \frac{n_{i,c}}{\sum_k n_{k,c}} \cdots\cdots (1)$$

$n_{i,c}$は，顧客cが商品iを購入した点数，$\sum_k n_{k,c}$ *は顧客*c*が*購入した全k種類の商品の合計点数を表す．このような定式化ではなく，もし点数を単純に用いた場合，Monetary スコアと高い相関となることが容易に予想される．その場合，多重共線性の問題から，RFM 分析と組み合わせてロジスティック回帰に投入することができない．つまり，RFM 分析

の拡張としては不適切な特徴量表現となる．一方，*if*スコア表現の場合，Monetary スコアの高低に関わらず顧客 1 人あたりの *if*スコアの合計は 1 になり，多重共線性の問題を回避できる．そのため，*if*スコアは，RFM 分析の拡張として適切な手法であると言える．

2 つめの提案は，店舗の異質性を表現した定式化である．顧客の趣向を把握する上で，顧客の所属する店舗全体の販売傾向を考慮し，多くの顧客が購入している商品目の影響を顧客の趣向から取り除く表現手法を提案する．図 5-2 は，分析対象のスーパーマーケットチェーンに含まれる 18 店舗の商品目ごとの販売点数の割合を示したものである．

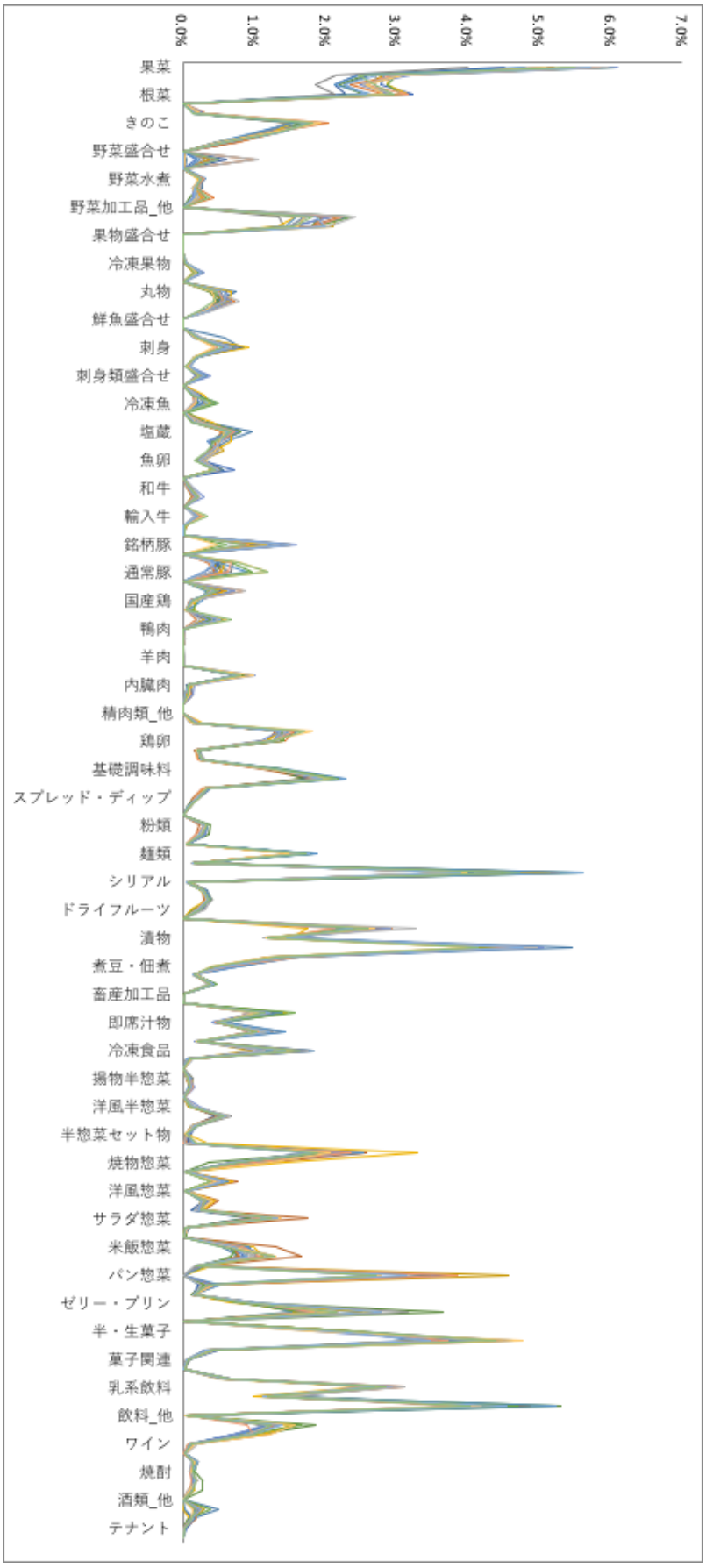


図 5-2: 商品ごとの店舗の販売傾向差

特に店舗間の差異が大きかった上位 3 商品目は，パン惣菜，アイスクリーム，果菜となり，店舗間で 5%から 2%程度の販売点数の割合の差が確認できた．この差を特徴量表現に活用した各店舗の販売商品の逆頻度 *inverse shop frequency* スコア(以下，*isf* スコア)を式(2)として定式化する．

$$isf_{i,s} = log \frac{|C_s|}{|\{c_s : \ni i_s\}|} \cdots\cdots (2)$$

$|C_s|$は，店舗 s における総顧客数，$|\{c_s : \ni i_s\}|$は店舗 s における商品 i を購入した顧客数を表す．*isf* ベクトルは，各店舗の商品ごとに定義される．店舗の顧客に広く購入されている商品は低く，限られた顧客だけに購入されている商品は高く算出される．スーパーマーケットチェーンの全顧客に対して算出した *if* スコアに顧客の属する店舗ごとの *isf* スコアを乗じて *if-isf* スコアとして定義する．例えば，パン惣菜が人気の店舗で，顧客がパン惣菜を購入した場合，*if* スコアは，店舗内の人気に依らず顧客の特徴として上昇する．一方で，*if-isf* スコアは，店舗内の人気を考慮し，上昇を抑える働きをする．逆に，パン惣菜が不人気の店舗で，顧客がパン惣菜を購入した場合，*if-isf* スコアは，店舗内の人気を考慮し，上昇を促す働きをする．本研究の 2 つめの提案として，この *if-isf* スコアを用いて，図 5-2 に示した店舗それぞれの販売傾向の異質性を考慮しながら顧客の特徴量表現を獲得する．

5.3 評価

5.3.1 評価手順

1 つのスーパーマーケットチェーン 18 店舗 2 年分の ID-POS データを，前半 1 年，後半 1 年のデータに分割する．前半 1 年のデータからは顧客ごとの特徴量として RFM 指標，*if* スコア，*if-isf* スコアを抽出する．後半 1 年のデータを利用してデシル分析を実施し，上位 3 グループの顧客に対して優良顧客フラグを付与する．ID-POS データに予め含まれる顧客 ID を利用して，目的変数としての優良顧客フラグと説明変数としての顧客ごとの特徴量を結合し，実験データとする．この結合処理の設計上，実験に用いる顧客の母集団は前半 1 年と後半 1 年のデータそれぞれに来店のあった，少なくとも 2 回以上の来店のあった顧客に限定される．この制約は，一般顧客の中のデータの欠損に相当するが，欠損しても一般顧客の学習には十分なデータ量があること，欠損は一般顧客の学習に質的に致命的なバイアスとならないこと，以上から優良顧客の判別のための学習データとして致命的な問題はない，と考える．

実験データは，モデル構築に用いる訓練データとモデル評価に用いる検証データの 2 つに分割して利用する．実験データに含まれる正例，負例の割合は約 30:70 であり，抽出問題における正例，負例の偏りが大きい際に用いられるオーバーサンプリング，アンダーサンプリングの手法[112]は，本実験においては適用不要と判断する．また，モデル構築にあたり説明変数の投入には強制投入法を採用することで，優良顧客の抽出に対して有意に寄与する RFM 指標，商品目を網羅的に探索する．

実験は 2 段階で構成する．1 段階目は，チェーン全体での抽出の実験である．チェーンの全店舗全顧客から 1 つのモデルを構築し，チ

ェーン全体の優良顧客を抽出する．説明変数の投入は，RFM モデル，RFM 指標に *if* スコアを組み合わせた RFM+IF モデル，RFM 指標に *if-isf* スコアを組み合わせた RFM+IF-ISF モデルの 3 パターンを試行する．モデルの評価は，正確性，適合率，再現率，F 値を用いた抽出精度，および優良顧客の特徴理解の 2 つの観点で行う．

2 段階目は，店舗個別での抽出の実験である．18 店舗個別に 1 段階目と同様の手法を用いて 18 のモデルを構築し，店舗ごとに抽出精度の評価を行う．但し，各店舗の販売傾向の異質性を活用する *if-isf* スコアは，単一店舗に絞ったモデル構築では機能しないため RFM モデル，RFM+IF モデルの 2 つの手法を用いる．評価は，1 段階目の顧客全体を用いて構築したモデルの内で最も抽出精度の良いモデルと，抽出精度比較を行う．スーパーマーケットチェーン経営において，モデル構築を含めたデータ分析には，経営資源の投資が必要であり，店舗ごとに構築した複数のモデルと同程度以上の抽出精度を，1 つのモデルで実現することで提案手法の経営上の有用性を示す．

5.3.2 チェーン全体での評価

まず，分類精度を評価する．構築した 3 モデルの抽出精度の評価を正確性，適合率，再現率，F 値の 4 指標から評価した結果を表 5-3 に示す．

表 5-3: 手法ごとの優良顧客の抽出精度差

	Accuracy	Precision	Recall	F-measure
RFM	88.68%	86.00%	69.76%	0.770
RFM+IF	88.77%	85.95%	70.22%	0.773
RFM+IF-ISF	89.25%	86.52%	71.68%	0.784

全 4 指標において RFM+IF-ISF モデルが優れたモデルとして評価された．また，抽出精度の総合的な評価として正確性，F 値ともに RFM+IF-ISF モデル，RFM+IF モデル，RFM モデルの順に優れたモデルとして評価された．そして，正確性の差の統計的有意性について，二項検定を用いて確認し，RFM+IF-ISF モデルと RFM+IF モデルの間に 1%有意差があることを確認した．

次に,優良顧客の特徴理解の観点で評価する．RFM モデルの構築結果を表 5-4 に示す．RFM 指標は，Recency は前回来店からデータ取得日までの経過月数の逆数，Frequency は来店回数，Monetary は購入額，それぞれ異なる単位で表現している．そのため偏回帰係数の大小関係は比較できないが，符号の正負を評価できる．

表 5-4: RFM モデルの偏回帰係数

Criterion Variables	Coefficients	P Values
Intercept	-1.37E+01	0.0%
Recency	9.61E-01	0.0%
Frequency	4.58E-03	0.0%
Monetary	2.23E-05	0.0%

RFM 指標は，優良顧客の抽出に対していずれも統計的有意に寄与している．偏回帰係数の符号は正で，最近来店しており，多く来店しており，多く購入している顧客を優良顧客と考える直感的な理解と合致する．しかし，それゆえに発見性に乏しく，本研究の目的である優良顧客の維持に繋がる特徴理解が得られたとは言えない．

次に，RFM+IF モデルの構築結果を表 5-5 に示す．尚，誌面の都合から RFM モデルと殆ど同結果が得られた RFM 指標および優良顧客抽出への 5%有意な寄与が確認できなかった説明変数については省略する．また説明変数は偏回帰係数の大きさの降順で記述する．

表 5-5: RFM+IF モデルの商品ごとの偏回帰係数

Criterion Variables	Coefficients	P Values
果菜	6.81E+00	2.0E-16
葉菜	3.27E+00	1.2E-02
茎菜	5.90E+00	2.2E-08
根菜	4.52E+00	2.1E-05
山菜	6.48E-01	6.6E-01
豆類	3.55E+00	1.4E-02
きのこ	4.85E+00	8.6E-05
発芽野菜	3.73E+00	5.0E-04
妻物	2.09E+00	1.2E-01
野菜盛合せ	-6.50E+00	3.0E-01
野菜関連	4.56E-01	7.4E-01
野菜_他	1.14E+00	7.8E-01
野菜水煮	1.56E+00	2.7E-01
冷凍野菜	4.99E-01	6.6E-01
カット野菜	1.48E+00	2.7E-01
野菜加工品_他	-4.12E+00	5.0E-01
季節果物	3.04E+00	1.6E-04
輸入果物	4.00E+00	1.1E-07
果物盛合せ	2.17E+00	6.2E-01
果物関連	-2.57E+00	6.1E-01
果物_他	-1.72E+00	7.6E-01
冷凍果物	5.77E-01	3.5E-01
カットフルーツ	2.08E+00	4.1E-03
果物加工品_他	-2.19E+01	3.1E-01

丸物	1.02E+00	3.6E-01
切身	3.05E+00	4.5E-03
貝	2.01E+00	1.1E-01
鮮魚盛合せ	2.93E+01	2.6E-01
鮮魚_他	-2.17E-01	9.5E-01
冊類	1.07E+00	3.8E-01
刺身	3.09E+00	3.2E-04
たたき	1.68E+00	1.7E-01
水産生食	2.87E-01	8.9E-01
刺身類盛合せ	2.30E+00	2.1E-02
刺身類_他	1.76E+00	5.0E-01
ボイル魚	2.55E+00	2.2E-02
冷凍魚	1.44E+00	1.0E-01
味付魚	1.37E+00	5.1E-01
漬魚	-1.32E+00	3.0E-01
塩蔵	1.48E+00	1.9E-01
干物	1.56E+00	1.3E-01
小魚	4.19E+00	1.6E-03
魚卵	4.10E+00	9.6E-03
海草	1.37E+00	2.4E-01
塩干加工品_他	2.01E-01	9.5E-01
和牛	2.57E+00	1.3E-01
国産牛	1.50E+00	1.0E-01
豪州産牛	NA	NA
米国産牛	1.66E+00	3.6E-01
輸入牛	3.56E-01	8.1E-01
通常牛	-1.45E-01	9.4E-01
牛肉関連	2.92E+00	5.6E-02
牛肉_他	NA	NA

銘柄豚	4.89E+00	4.2E-07
国産豚	1.34E+00	5.7E-01
輸入豚	1.33E+00	1.8E-01
通常豚	3.54E+00	1.7E-02
豚肉関連	-1.20E+02	8.8E-01
豚肉_他	NA	NA
銘柄鶏	4.05E+00	2.1E-04
国産鶏	2.65E+00	5.9E-02
輸入鶏	1.18E+00	5.5E-01
通常鶏	2.27E+00	7.1E-02
鴨肉	1.14E+00	1.3E-01
鶏肉関連	-9.59E-02	9.6E-01
鶏肉_他	2.15E+01	4.7E-01
羊肉	2.45E+00	1.6E-01
馬肉	8.70E-01	7.8E-01
ひき肉	2.37E+00	1.6E-02
内臓肉	-3.34E-01	7.9E-01
畜産生食	4.87E-02	9.7E-01
精肉盛合せ	-7.82E-01	8.7E-01
精肉類_他	-1.24E+01	3.2E-01
味付肉	2.30E+00	7.8E-02
加工肉	6.66E+00	1.6E-07
鶏卵	6.28E+00	1.3E-08
精肉加工品_他	NA	NA
食用油	2.25E+00	1.9E-01
香辛料	3.78E+00	1.1E-03
基礎調味料	5.64E+00	4.9E-08
加工調味料	6.73E+00	2.5E-09
米飯調味料	3.69E+00	4.3E-04

スプレッド・ディップ	2.57E+00	1.9E-02
トッピング	1.20E+00	5.1E-01
調味料_他	2.02E+00	8.6E-01
粉類	1.05E+00	4.9E-01
米	2.09E+00	5.6E-03
餅	2.27E+00	2.2E-01
麺類	4.39E+00	3.9E-05
皮生地	6.19E-02	9.6E-01
パン	3.85E+00	5.3E-06
シリアル	1.24E+00	7.2E-02
穀物類_他	NA	NA
農産乾物	1.75E+00	2.9E-01
水産乾物	3.77E+00	5.4E-03
ドライフルーツ	1.46E+00	7.5E-02
乾物類_他	1.02E+01	1.7E-01
乳製品	7.49E+00	2.0E-16
漬物	6.40E+00	4.8E-11
水物	8.56E+00	2.0E-16
練物	3.88E+00	4.9E-04
煮豆・佃煮	2.11E+00	6.1E-02
農産加工品	2.26E+00	1.1E-01
水産加工品	1.64E+00	2.3E-01
畜産加工品	1.77E+00	2.5E-01
加工食品_他	-2.86E+00	4.2E-01
即席麺	3.43E+00	3.9E-07
即席汁物	1.82E+00	5.4E-02
レトルト惣菜	4.12E+00	1.3E-05
レトルト米飯	1.85E+00	2.1E-02
冷凍食品	3.76E+00	2.9E-12

その他食品	9.53E-01	3.5E-01
即席食品_他	2.38E+00	2.8E-01
揚物半惣菜	5.11E-01	7.5E-01
煮物半惣菜	1.38E-01	9.5E-01
焼物半惣菜	8.92E-01	6.7E-01
和風半惣菜	NA	NA
洋風半惣菜	2.92E+00	1.5E-02
中華半惣菜	2.68E+00	2.2E-03
スナック半惣菜	-1.75E-01	9.3E-01
半惣菜セット物	1.80E+00	3.6E-01
半惣菜_他	6.94E-01	5.0E-01
揚物惣菜	3.35E+00	8.6E-06
焼物惣菜	2.35E+00	2.5E-07
蒲焼惣菜	-1.78E+00	5.2E-01
和風惣菜	2.22E+00	5.6E-03
洋風惣菜	2.95E+00	3.9E-02
中華惣菜	2.98E+00	7.2E-03
スナック惣菜	7.57E-01	5.1E-01
サラダ惣菜	4.14E+00	7.1E-09
惣菜盛合せ	2.42E+00	2.4E-02
惣菜_他	3.08E+00	1.4E-01
米飯惣菜	1.03E+00	1.1E-01
寿司惣菜	2.48E+00	2.0E-04
麺惣菜	2.80E-01	7.9E-01
パン惣菜	3.85E+00	8.9E-15
弁当・セット物	1.78E+00	8.8E-04
弁当_他	NA	NA
製菓材料	4.09E-01	7.4E-01
ゼリー・プリン	4.09E+00	2.4E-09

アイスクリーム	2.99E+00	2.0E-16
冷凍菓子	-9.67E-01	6.6E-01
半・生菓子	4.19E+00	1.3E-05
乾菓子	5.08E+00	3.2E-15
つまみ菓子	1.76E+00	6.6E-02
菓子関連	4.98E-02	9.8E-01
菓子_他	3.01E-01	8.7E-01
嗜好飲料	3.20E+00	1.9E-04
乳系飲料	6.88E+00	2.0E-16
野菜・果実飲料	2.11E+00	1.3E-03
清涼飲料	5.20E+00	2.0E-16
飲料_他	1.81E+00	5.7E-02
ビール類	2.18E+00	2.3E-14
リキュール類	1.94E+00	2.2E-15
ワイン	1.28E+00	3.4E-02
洋酒	2.43E+00	6.8E-03
日本酒	1.41E+00	1.5E-03
焼酎	-1.13E-01	8.6E-01
ノンアルコール飲料	8.05E-01	1.5E-01
酒関連	1.79E+00	5.5E-05
酒類_他	-6.41E-01	9.9E-01
たばこ	NA	NA
花類	1.79E+00	1.7E-04
ギフト・銘菓	2.71E-01	7.2E-01
テナント	-4.62E+00	1.5E-01
カウンター	NA	NA
その他_他	1.17E+00	5.0E-01

説明変数として投入した 167 品目中，13 品目について優良顧客抽出への 5%有意な寄与を確認した．しかしながら，13 品目のうち 11 品目は偏回帰係数の符号が負となり，優良顧客を維持するためには売らないほうがよい商品目である，と示されている．また，その中には通常豚，パン，葉菜，清涼飲料など一般的な商品目を含んでおり，実際の経営者に直感的に受け入れ難く，本研究の目的である優良顧客の維持に繋がる特徴理解が得られたとは言えない．

最後に，RFM+IF-ISF モデルの構築結果を表 5-6 に示す．尚，RFM+IF モデルの評価と同じく RFM モデルと殆ど同結果が得られた RFM 指標および優良顧客抽出への 5%有意な寄与が確認できなかった説明変数については省略する．また，説明変数は偏回帰係数の大きさの降順で記述する．

表 5-6: RFM+IF-ISF モデルの商品ごとの偏回帰係数

Criterion Variables	Coefficients	P Values
水物	8.56	0.0%
乳製品	7.49	0.0%
乳系飲料	6.88	0.0%
果菜	6.81	0.0%
加工調味料	6.73	0.0%
加工肉	6.66	0.0%
漬物	6.40	0.0%
鶏卵	6.28	0.0%
茎菜	5.90	0.0%
基礎調味料	5.64	0.0%
清涼飲料	5.20	0.0%
乾菓子	5.08	0.0%
銘柄豚	4.89	0.0%

説明変数として投入した 167 品目中 66 品目の優良顧客抽出への 5%有意な寄与を確認した. また, 66 品目全ての偏回帰係数の符号が正となり, 優良顧客を維持するために売ったほうがよい商品目として示された. 尚, 表 5-6 に記載した商品目は 66 品目から偏回帰係数の上位 13 品目に絞り込んでいる. RFM+IF モデルが抽出した 13 品目と比較すると, 水物, 乳製品, 果菜, 調味料, 卵など食品スーパーマーケットで購入される一般的な商品目から構成されており, 経営者にとって, 優良顧客向けに販売促進すべき商品として直感的に受け入れ易い内容となっている. この結果から, 本研究の目的である優良顧客の維持に繋がる特徴理解が得られた, と評価する.

5.3.3 店舗個別での優良顧客の抽出

前節では，1 つのスーパーマーケットチェーンに含まれる 18 店舗の ID-POS データを全て利用して，1 つの優良顧客抽出モデルを構築した．本節では，18 店舗それぞれについて店舗個別モデルを構築するとともに，前節で構築したチェーン全体モデルとの精度比較を行う．

まず，店舗個別に構築した RFM モデルと RFM+IF モデルの抽出精度を正確性から比較する．比較結果を表 5-7 に示す．尚，RFM+IF-ISF モデルは，学習データに含まれる店舗間の異質性を活用するため，店舗個別の学習データを用いる本モデル構築では取り扱わない．

表 5-7: 店舗個別に導出した RFM モデルと RFM+IF モデルの精度差

Shop ID	Accuracy		P Values of Binomial Test
	Individual shop RFM Model	Individual shop RFM+IF Model	
1	89.1%	88.2%	11.2%
2	88.8%	88.7%	87.1%
3	88.8%	87.7%	9.4%
4	89.6%	88.4%	2.7%
5	90.0%	88.4%	0.5%
6	86.6%	80.9%	0.0%
7	90.9%	90.4%	29.0%
8	89.7%	88.8%	6.1%
9	90.9%	88.0%	0.0%
10	87.7%	85.8%	2.3%
11	89.5%	87.6%	0.7%
12	88.8%	86.6%	1.6%
13	87.2%	85.7%	3.7%
14	86.9%	83.8%	0.1%
15	90.7%	89.6%	1.3%
16	87.8%	86.4%	1.7%
17	86.9%	83.3%	0.0%
18	90.0%	89.5%	38.9%

全 18 店舗中，12 店舗において RFM モデルの抽出精度が統計的有意に優れていることを確認した．また，有意差が確認できなかった 6 店舗についても，RFM モデルの抽出精度が高い結果を示した．店舗

個別にモデルを構築して各店舗の優良顧客を抽出する場合，商品目を活用した提案手法よりもRFMモデルが優れていることを確認した．

次に，前節で最も優れた抽出精度であったチェーン全体から構築したチェーン全体RFM+IF-ISFモデルを利用して各店舗の優良顧客抽出を行い，上記の店舗個別RFMモデルとの抽出精度比較を行う．結果を表5-8に示す．

表 5-8: 店舗個別に導出した RFM モデルと RFM+IF-ISF モデルの精度差

Shop ID	Accuracy		P Values of Binomial Test
	Individual shop RFM Model	Individual shop RFM+IF Model	
1	89.1%	89.1%	95.6%
2	88.8%	89.4%	18.3%
3	88.8%	88.7%	92.4%
4	89.6%	89.2%	50.8%
5	90.0%	90.1%	85.4%
6	86.6%	86.0%	60.3%
7	90.9%	91.4%	30.5%
8	89.7%	89.9%	75.7%
9	90.9%	91.2%	63.3%
10	87.7%	88.6%	22.5%
11	89.5%	88.8%	32.8%
12	88.8%	88.5%	71.0%
13	87.2%	86.6%	39.0%
14	86.9%	86.9%	100.0%
15	90.7%	90.9%	75.8%
16	87.8%	87.8%	95.7%
17	86.9%	86.5%	61.1%
18	90.0%	90.1%	85.5%

全 18 店舗において店舗個別 RFM モデルとチェーン全体 RFM＋IF-ISF モデルの間に統計的に有意な抽出精度は無かった．単純な大小比較においては，18 店舗中 8 店舗が RFM モデル，10 店舗が RFM+IF-ISF モデルを支持した．しかし，統計的に有意な精度差が確認できなかったことから，総じて同等の抽出精度と評価する．店舗個別

RFM モデルは，18 店舗それぞれについて個別に構築して抽出に用いている一方で，チェーン全体 RFM＋IF-ISF モデルは 1 モデルのみを用いて店舗個別 RFM モデルと同精度で抽出していることから，より優れたモデルである．

5.4　おわりに

本研究では，市場縮小局面にある日本のスーパーマーケット経営にとって，重要な課題である優良顧客の維持に有用な新たな手法を提案した．提案手法は，現在だけでなく将来に渡って店舗の売上の大部分を生み出す優良顧客を予め抽出し，優良顧客が顕著に購買する商品目情報を提供する．また，複数店舗を持つチェーン経営者向けに，店舗間の異質性を考慮した 1 つのモデルを提案した．店舗個別にデータ分析を行い，それぞれで優良顧客を抽出した場合と同程度の精度が確認できていることから，少ない手間で同等の効果が得られるモデルを提案できた．

また，今回の検証では，スーパーマーケット事業の経営に役立てることを目的に，高精度の優良顧客抽出モデルの構築方法，および優良顧客は一般顧客が気づいていないような店舗の隠れた良い商品を買っているといった示唆を獲得した．だが，提案手法は，多くの品目を扱う業界であれば汎用的に適応できる手法であり，その適用可能性については本研究で検証できていない．そのような可能性の探索について，今後の研究課題とする．

第6章　結論

本研究は，我が国の企業における AI の導入，データ利活用の活性化を背景にして，データ分析によく用いられる行動データを，分析に使いやすい状態に整えることを目的に，行動データのノイズを除去する新たな手法を提案した．行動データのノイズは，企業の置かれている競争環境，企業の持つ顧客構造，企業がデータを生成するプロセスなど，様々な要素から生まれてくる．このような研究を進めるためには，企業との密な連携を行いながら，課題の設定から，解決策の検証まで，一連の研究プロセスを行うことが不可欠である．本研究は，行動データ分析のためのノイズ除去手法の研究として，企業から普遍性のある課題を抽出し，企業の保有する実データ，情報システムを用いて解決策の検証を行った．このような企業との密な連携は，本研究の行った検証の信頼性，研究そのものの独自性を支持するものと考えている．

まず，二章では，本研究が企業から抽出した課題と周辺の研究を説明した．1 つ目の課題は，住宅情報ポータルサイトの行動データに混入するインターネットボットの課題を説明した．このようなボットは，検索サイトや住宅情報を調査する競合他社など様々な主体によるものと考えられていた．ボットによる行動データを除去するための普遍的な知識や手法としては，ボットのアクセスのボリュームをもとに異常度を判定する手法があった．だが，このような手法では，住宅情報を調査するボットの振る舞いや属性情報が考慮できておらず，判別精度の課題があった．住宅情報ポータルサイトにおいて，ボットの判別精度は，集客という

経営インパクトの大きな活動の分析に非常に大きな影響があるため，本研究の課題として採択したことを説明した．

2 つ目の課題は，同じく住宅情報ポータルサイトの行動データに混入する不正識別子の課題を説明した．住宅情報ポータルサイトは，競争環境の中で，高頻度で継続的なウェブサイトの機能開発を行っていた．そのように繰り返される機能開発において，行動データを記録する機能のソフトウェアテストが正常に行われず，その結果として行動データに不正識別子がノイズとして混入してしまうことを説明した．ウェブサイトのソフトウェアテストを効率的に行うための研究には，過去に実施したテストをコードとして管理して，自動で実行する手法があり，このような手法はこの課題に対しても非常に有効であった．しかしながら，住宅情報ポータルサイトにおいては，そのようなテストコードを作成する作業者の確保の課題があり，少しでも効率化を進める手法が必要であったことを説明した．

3 つ目の課題は，スーパーマーケットチェーンの顧客管理における課題を取り上げた．スーパーマーケット業界の顧客構造は，売上の多くを生み出す少数の優良顧客とその他の多数の一般顧客という形になっていた．その顧客構造が，行動データにおいては，少数の優良顧客の行動データが多数の一般顧客の行動データに紛れてしまい，分析が行いづらいという問題を生み出していた．一般顧客の行動データをノイズとして除去するためには，顧客のランク付けの研究が有用であった．先行研究には，顧客属性を大きく 3 つに分けて取り扱うことで様々な業界に適用できる手法があったが，優良顧客の抽出の精度においては課題があった．多くの商品を取り扱うスーパーマーケットの特徴を活かして，より高い精度の優良顧客の抽出手法が必要であったことを説明した．

3 章では，ウェブサイトに訪れるインターネットボットを，ウェブサイトを構成する情報システムの構造における異なる構成要素から複数の行動ログを収集することで，単一の行動ログの分析では得られないボット

に関する知識の獲得する方法を提案した．その知識を，先行研究の異常検知モデルに対して，追加的な特徴量として与えたところ，インターネットボットの検知精度が向上することを報告した．

4 章では，ウェブサイトの継続的な開発の中で，行動データに不正識別子を埋め込んでしまう問題に対して，不正識別子を見つけるためのウェブサイトのテストコードそのものを自動で生成するシステムを提案した．複数のテストケースにおいて，従来手法による自動テストと提案手法による自動テストを比較したところ，テストコードそのものを自動生成した分の作業効率化が実現できたことを報告した．

5 章では，スーパーマーケットチェーンの優良顧客抽出の問題に対して，RFM 分析の拡張研究として，Monetary 指標を大量の商品目に分解し，分解した商品ごとの購入点数を文書符号化手法である TF-IDF を用いて一般的に購買されやすい商品の影響を小さく，それ以外を大きく特徴量として加工して，RFM 分析と組み合わせる新たな手法を提案した．従来手法である RFM 分析を用いた抽出精度，RFM 分析の拡張として商品目を購入点数として分解した手法の抽出精度と比較したところ，提案手法が最も良い抽出精度を実現できたことを報告した．

これらの研究を通して，企業がデータ利活用に取り組む際に直面する行動データのノイズの課題に対応するためのノイズ除去手法について，いくつかの新たな解決手法を提案できた．しかし，先に述べたように，行動データのノイズは，企業の置かれている環境など様々な要因から発生していることから，異なる原因を持つ様々なデータのノイズを単一の手法で解決することは困難である．今後も，企業と連携しながら，新たな行動データのノイズの課題に取り組んでいく．

謝辞

博士課程の在学中，公私に渡って大変お世話になった筑波大学大学院ビジネス科学研究科の津田和彦教授に深く感謝いたします．また，本論文の執筆にあたり，津田研究室の方々には，日頃より研究の進め方についての貴重な示唆やご意見を頂戴いたしました．深く感謝いたします．

論文審査を快くお引き受けいただき，的確なアドバイスを頂戴した筑波大学大学院ビジネス科学研究科の倉橋節也教授，木野泰伸准教授，同大学同大学院システム情報工学研究科の古川宏准教授，帝京大学文学部社会学科の藤田昌克教授に深く感謝いたします．

発表の場などで，様々なアドバイスやコメントを下さった筑波大学大学院の教員の方々に深く感謝いたします．

最後に，筆者が在籍した株式会社 NTT データ，日本 IBM 株式会社，株式会社リクルートの諸先輩方や同僚の方々には，様々なアドバイスをいただきました．深く感謝いたします．

参考文献

[1]内閣府, “Society5.0”, https://www8.cao.go.jp/cstp/society5_0/society5_0-1.pdf (2018).

[2]Bojer, Casper Solheim, and Jens Peder Meldgaard. ″Kaggle forecasting competitions: An overlooked learning opportunity.″ *International Journal of Forecasting* 37.2 (2021): 587-603.

[3]Data Management Research Group. 2011. “Data management research at NEC labs.” SIGMOD Rec. 40, 3 (September 2011), 38-44.

[4]Lahdenmaki, Tapio, and Mike Leach. Relational Database Index Design and the Optimizers: DB2, Oracle, SQL Server, et al. John Wiley & Sons, 2005.

[5]Gupta, Vishal, and Gurpreet S. Lehal. ″A survey of text mining techniques and applications.″ Journal of emerging technologies in web intelligence 1.1 (2009): 60-76.

[6]Jayamalini, K., and M. Ponnavaikko. ″Research on web data mining concepts, techniques and applications.″ 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET). IEEE, 2017.

[7]Gourley, David, et al. HTTP: the definitive guide. ″ O’Reilly Media, Inc.″, 2002.

[8]一般社団法人 日本チェーンストア協会, ”2021 年度版スーパーマーケット白書”, http://www.super.or.jp/wp-content/uploads/2020/02/super_hakusho2021.jpg (2021)

[9]澁谷覚. "Web サイトとマス広告の連動による新たな広告プロモーション・モデルの可能性." *広告科学* 40 (2000): 173-180.

[10] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. 2001.

[11] Negash, Solomon, and Paul Gray. "Business intelligence." Handbook on decision support systems 2. Springer, Berlin, Heidelberg, 2008. 175-193.

[12] Chatfield, Chris. Time-series forecasting. CRC press, 2000.

[13] Lu, Jie, et al. "Recommender system application developments: a survey." Decision Support Systems 74 (2015): 12-32.

[14] Brackett, Michael, and Production Susan Earley. "The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)." (2009).

[15] Aiken, Peter, et al. "Measuring data management practice maturity: A community's self-assessment." Computer 40.4 (2007): 42-50.

[16] Fleckenstein, Mike, and Lorraine Fellows. "Overview of Data Management Frameworks." Modern Data Strategy. Springer, Cham, 2018. 55-59.

[17] Bansal, Srividya K., and Sebastian Kagemann. "Integrating big data: A semantic extract-transform-load framework." Computer 48.3 (2015): 42-50.

[18] Toyotaro Suzumura, Toshiaki Yasue, and Tamiya Onodera. 2010. Scalable performance of system S for extract-transform-load processing. In Proceedings of the 3rd Annual Haifa Experimental Systems Conference (SYSTOR '10). Association for Computing Machinery, New York, NY, USA, Article 7, 1-14.

[19] Sekiguchi, Akiyuki, et al. "Web analytics for B to B marketing in semiconductor industry." International Journal of e-Education, e-Business, e-Management and e-Learning 2.5 (2012).

[20] Ashar Javed. 2013. POSTER: A footprint of third-party tracking on mobile web. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (CCS '13). Association for Computing Machinery, New York, NY, USA, 1441-1444.

[21] Mirkovic, Jelena, et al. Internet denial of service: attack and defense mechanisms (Radia Perlman Computer Networking and Security). Prentice Hall PTR, 2004

[22] T. Tanaka et al., "Bot Detection Model using User Agent and User Behavior for WebLog Analysis," Procedia Computer Science, Vol 176, 2020.

[23] Zhang, Yang, et al. "Detecting malicious activities with user-agent-based profiles." International Journal of Network Management 25.5 (2015): 306-319.

[24] Kheir, Nizar. "Analyzing http user agent anomalies for malware detection." Data Privacy Management and Autonomous Spontaneous Security. Springer, Berlin, Heidelberg, 2012. 187-200.

[25] Kitts, Brendan, et al. "Click fraud detection with bot signatures." 2013 IEEE International Conference on Intelligence and Security Informatics. IEEE, 2013.

[26] Grill, Martin, and Martin Rehák. "Malware detection using http user-agent discrepancy identification." 2014 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2014.

[27] Loyola-González, Octavio, et al. "An approach based on contrast patterns for bot detection on Web log files." Mexican International Conference on Artificial Intelligence. Springer, Cham, 2018.

[28] Stassopoulou, Athena, and Marios D. Dikaiakos. "Web robot detection: A probabilistic reasoning approach." Computer Networks 53.3 (2009): 265-278.

[29] Ali Alhosseini, Seyed, et al. "Detect me if you can: Spam bot detection using inductive representation learning." Companion Proceedings of The 2019 World Wide Web Conference. 2019.

[30] Kouvela, Maria, Ilias Dimitriadis, and Athena Vakali. "Bot-detective: An explainable Twitter bot detection service with crowdsourcing functionalities." Proceedings of the 12th International Conference on Management of Digital EcoSystems. 2020.

[31] Mitterhofer, Stefan, et al. "Server-side bot detection in massively multiplayer online games." IEEE Security & Privacy 7.3 (2009): 29-36.

[32] Masud, Mohammad M., et al. "Flow-based identification of botnet traffic by mining multiple log files." 2008 First International Conference on Distributed Framework and Applications. IEEE, 2008

[33] Yu, Fang, Yinglian Xie, and Qifa Ke. "Sbotminer: large scale search bot detection." Proceedings of the third ACM international conference on Web search and data mining. 2010.

[34] Cai, Chiyu, Linjing Li, and Daniel Zengi. "Behavior enhanced deep bot detection in social media." 2017 IEEE

International Conference on Intelligence and Security Informatics (ISI). IEEE, 2017.

[35] Yang, Kai-Cheng, et al. "Scalable and generalizable social bot detection through data selection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.

[36] Abou Daya, Abbas, et al. "A graph-based machine learning approach for bot detection." 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). IEEE, 2019.

[37] Tao, Jianrong, et al. "Nguard: A game bot detection framework for netease mmorpgs." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[38] Kang, Ah Reum, et al. "Online game bot detection based on party-play log analysis." Computers & Mathematics with Applications 65.9 (2013): 1384-1395.

[39] Beskow, David M., and Kathleen M. Carley. "Bot conversations are different: leveraging network metrics for bot detection in twitter." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018.

[40] Pao, Hsing-Kuo, Kuan-Ta Chen, and Hong-Chung Chang. "Game bot detection via avatar trajectory analysis." *IEEE Transactions on Computational Intelligence and AI in Games* 2.3 (2010): 162-175.

[41] Efthimion, Phillip George, Scott Payne, and Nicholas Proferes. "Supervised machine learning bot detection techniques to identify social twitter bots." *SMU Data Science Review* 1.2 (2018): 5.

[42] Heidari, Maryam, H. James Jr, and Ozlem Uzuner. "An empirical study of machine learning algorithms for social media bot detection." *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021.

[43] Mohammad, Shad, et al. "Bot detection using a single post on social media." *2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*. IEEE, 2019.

[44] Heidari, Maryam, and James H. Jones. "Using bert to extract topic-independent sentiment features for social media bot detection." *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2020

[45] Beatson, Oliver, et al. "Automation on twitter: Measuring the effectiveness of approaches to bot detection." *Social Science Computer Review* (2021): 08944393211034991.

[46] Acien, Alejandro, et al. "BeCAPTCHA-Mouse: Synthetic mouse trajectories and improved bot detection." *arXiv preprint arXiv:2005.00890* (2020)

[47] Lee, Jina, et al. "In-game action sequence analysis for game bot detection on the big data analysis platform." *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems-Volume 2*. Springer, Cham, 2015.

[48] Varol, Onur, et al. "Online human-bot interactions: Detection, estimation, and characterization." *Proceedings of the international AAAI conference on web and social media*. Vol. 11. No. 1. 2017.

[49] Chu, Zi, Steven Gianvecchio, and Haining Wang. "Bot or human? A behavior-based online bot detection system." *From Database to Cyber Security*. Springer, Cham, 2018. 432-449.

[50] Ji, Yuede, et al. "A mulitiprocess mechanism of evading behavior-based bot detection approaches." *International conference on information security practice and experience*. Springer, Cham, 2014.

[51] Pham, Phu, et al. "Bot2Vec: a general approach of intra-community oriented representation learning for bot detection in different types of social networks." *Information Systems* 103 (2022): 101771.

[52] Cabri, Alberto, et al. "Online web bot detection using a sequential classification approach." *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2018

[53] Shariff, Shahnaz M., et al. "Improving the testing efficiency of selenium-based load tests." Proceedings of the 14th International Workshop on Automation of Software Test. IEEE Press, 2019.

[54] Chen, Ruifeng, and Huaikou Miao. "A selenium based approach to automatic test script generation for refactoring javascript code." 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). IEEE, 2013.

[55] de Castro, Andreza MFV, et al. "Extension of Selenium RC tool to perform automated testing with databases in web

applications." Proceedings of the 8th International Workshop on Automation of Software Test. IEEE Press, 2013.

[56] Iyama, Muneyoshi, et al. "Automatically Generating Test Scripts for GUI Testing." 2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE, 2018.

[57] Milani Fard, Amin, Mehdi Mirzaaghaei, and Ali Mesbah. "Leveraging existing tests in automated test generation for web applications." Proceedings of the 29th ACM/IEEE international conference on Automated software engineering. ACM, 2014

[58] Nagarajan, Sarvesh, Nastaran Shafiei, and Sarfraz Khurshid. "Towards Exhaustive Testing of Websites using JPF." ACM SIGSOFT Software Engineering Notes 41.6 (2017): 1-5.

[59] Mirshokraie, Shabnam, Ali Mesbah, and Karthik Pattabiraman. "PYTHIA: Generating test cases with oracles for JavaScript applications." Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering. IEEE Press, 2013.

[60] Presler-Marshall, Kai, et al. "Wait wait. No, tell me: analyzing selenium configuration effects on test flakiness." Proceedings of the 14th International Workshop on Automation of Software Test. IEEE Press, 2019.

[61] Leotta, Maurizio, et al. "Comparing the maintainability of selenium webdriver test suites employing different locators: A case study." Proceedings of the 2013 International Workshop on Joining AcadeMiA and Industry Contributions to Testing Automation, ser. JAMAICA. 2013.

[62] Neto, Nelson Mariano Leite, Patrícia Vilain, and Ronaldo dos Santos Mello. "Segen: Generation of test cases for selenium

and selendroid." Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services. ACM, 2016.

[63] Stocco, Andrea, et al. "Why creating web page objects manually if it can be done automatically?." Proceedings of the 10th International Workshop on Automation of Software Test. IEEE Press, 2015.

[64] Lim, Woosup, et al. "D-TAF: test automation framework compatible with various DBMS." Proceedings of the 14th International Workshop on Automation of Software Test. IEEE Press, 2019.

[65] Martinez, Jorge, Troy Thomas, and Tariq M. King. "Echo: A middleware architecture for domain-specific ui test automation." Proceedings of the 2014 Workshop on Joining AcadeMiA and Industry Contributions to Test Automation and Model-Based Testing. ACM, 2014.

[66] Taneja, Kunal, et al. "eXpress: guided path exploration for efficient regression test generation." Proceedings of the 2011 International Symposium on Software Testing and Analysis. ACM, 2011.

[67] Bures, Miroslav, and Martin Filipsky. "SmartDriver: Extension of selenium WebDriver to create more efficient automated tests." 2016 6th International Conference on IT Convergence and Security (ICITCS). IEEE, 2016.

[68] García, Boni, et al. "Extending WebDriver: A Cloud Approach." 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC). IEEE, 2018

[69] Kirinuki, Hiroyuki, Haruto Tanno, and Katsuyuki Natsukawa. "COLOR: Correct Locator Recommender for Broken Test Scripts using Various Clues in Web Application." 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE, 2019

[70] P.S. Fader, B.G. Hardie, and K. Loklee, "RFM and CLV:Using iso-value curves for customer base analysis," Journal of Marketing Research, vol.42, pp.415-430, 2005.

[71] Khajvand, Mahboubeh, et al. "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study." Procedia Computer Science 3 (2011): 57-63

[72] Aggelis, Vasilis, and Dimitris Christodoulakis. "Customer clustering using rfm analysis." Proceedings of the 9th WSEAS International Conference on Computers. 2005.

[73] Kuanchin Chen,Ya-Han Hu, Yi-Cheng Hsieh.: Predicting customer churn from valuable B2B customers in the logistics industry: a case study, Information Systems and e-Business Management, Volume 13, Issue 3, pp 475-494,(2015)

[74] Mehdi Bizhani, Mohammad Jafar Tarokh.: Behavioral segmentation of bank's Point-of-Sales using RF*M* approach, IEEE International Conference on Computational Photography (ICCP), vol. 00, no. , pp. 81-86, (2010)

[75] Hsin-Hung Wu, Shian-Chang Huang, Jo-Ting Wei, Shih-Yen Lin.:Using Bayesian Network and LRFM Model in a Pediatric Dental Clinic, Computer, Consumer and Control, International Symposium on, vol. 00, no. , pp. 20-23, (2012)

[76] Chu Chai Henry Chan, Ying-Rown,HwangHsin-Chieh Wu.: Marketing segmentation using the particle swarm optimization

algorithm: a case study, Journal of Ambient Intelligence and Humanized Computing, Volume 7, Issue 6, pp 855-863,(2016)
[77] Dirk Van den Poel, Michel Ballings, Dries Benoit.: RFM Variables Revisited Using Quantile Regression, 2013 IEEE 13th International Conference on Data Mining Workshops, vol. 00, no. , pp. 1163-1169, (2011)
[78] Xiaoqing Zeng, Qi Wang, Qiang Li, Jinghua Jiang.: A Multi-indicator Customer Segmentation Method Based on Consuming Behaviors Analysis, 2015 International Conference on Network and Information Systems for Computers, vol. 00, no. , pp. 289-295, (2015)
[79] Xiaoqing Zeng, Qi Wang, Qiang Li, Jinghua Jiang.: A Multi-indicator Customer Segmentation Method Based on Consuming Behaviors Analysis, 2015 International Conference on Network and Information Systems for Computers, vol. 00, no. , pp. 289-295, (2015)
[80] 日本チェーンストア協会, チェーンストア販売統計
[81] 新日本スーパーマーケット協会:2015 年版スーパーマーケット白書,2015
[82] 総務省統計局, 人口推計
[83] Petteri Nurmi, Antti Salovaara, Andreas Forsblom, Fabian Bohnert, Patrik Floréen.: PromotionRank: Ranking and Recommending Grocery Product Promotions Using Personal Shopping Lists, ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Interactive Computational Visual Analytics archive,Volume 4 Issue 1, (2014)
[84] 大澤幸生 , 臼井 優樹 , 福田 寿, 松尾 豊 , 松村 真宏 , 高山 美和 , 相馬 浩隆 , 佐橋 官:二重螺旋モデルを用いたス

ーパーの顧客行動変化の予兆発見, 情報処理学会研究報告知能と複雑系(ICS) 2002(45(2002-ICS-128)), 169-174,(2002)
[85] 飯塚久哲, 米村大介, 豊田秀樹: 顧客ランクによる行動分析, オペレーションズ・リサーチ, 2003 年 2 月号, pp. 94-99, (2003)
[86] Masakazu Takahashi, Takashi Yamada, Kazuhiko Tsuda, Takao Terano.:Towards Small-Sized Long Tail Business with the Dual-Directed Recommendation System, The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society 130(2), 317-323, (2010)
[87] Xiaojia Chen, Ying Li, Tao Hu.:Solving the supermarket shopping route planning problem based on genetic algorithm, 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), vol. 00, no. , pp. 529-533, (2015)
[88] 岸本有之, 高橋徹, 高橋雅和, 山田隆志, 津田和彦, 寺野隆雄: エージェント・シミュレーションによる店舗内顧客行動と販売促進策の分析, 情報処理学会研究報告知能と複雑系, Vol. 2009, No. 16, pp. 87-92, 2009
[89] 川田一貴, 岩宮 眞一郎: スーパーマーケットの売場における音環境に関する意識調査, 情報処理学会研究報告音楽情報科学, Vol. 2001, No. 16, pp. 79-86, (2001)
[90] 阿部誠, 近藤文代: マーケティングの科学ーPOS データの解析, 朝倉書店, 2005
[91] Olston, Christopher, and Marc Najork. Web crawling. Now Publishers Inc, 2010.
[92] Mitchell, Ryan. Web scraping with Python: Collecting more data from the modern web. ″ O'Reilly Media, Inc.″, 2018.

[93] Hostetter, Mat, et al. "Curl: a gentle slope language for the Web." World wide web journal 2.2 (1997): 121-134.
[94] Motukuru, Vamsi, Vikas Pooven Chathoth, and Vipin Anaparakkal Koottayi. "Cookie based session management." U.S. Patent No. 9,866,640. 9 Jan. 2018.
[95] Kristol, David, and Lou Montulli. HTTP state management mechanism. RFC 2965, October, 2000.
[96] Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework." International Journal of Machine Learning and Cybernetics 1.1-4 (2010): 43-52.
[97] Kleinbaum, David G., et al. Logistic regression. New York: Springer-Verlag, 2002.
[98] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.
[99] Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models." Global ecology and Biogeography 17.2 (2008): 145-151.
[100] Tax, David MJ, and Robert PW Duin. "Support vector data description." Machine learning 54.1 (2004): 45-66.
[101] 井手剛. 入門機械学習による異常検知: R による実践ガイド. コロナ社, 2015.
[102] 井手剛, and 杉山将. 異常検知と変化検知. 講談社, 2015.
[103] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence 2009 (2009).

[104] Lops, Pasquale, Marco De Gemmis, and Giovanni Semeraro. "Content-based recommender systems: State of the art and trends." Recommender systems handbook. Springer, Boston, MA, 2011. 73-105.

[105] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[106] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?-Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.

[107] Fielding, Roy T., and Gail Kaiser. "The Apache HTTP server project." IEEE Internet Computing 1.4 (1997): 88-90.

[108] Duvall, Paul M., Steve Matyas, and Andrew Glover. Continuous integration: improving software quality and reducing risk. Pearson Education, 2007.

[109] Bruns, Andreas, Andreas Kornstadt, and Dennis Wichmann. "Web application tests with selenium." IEEE software 26.5 (2009): 88-91.

[110] 北研二, 津田和彦, and 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002.

[111] 新屋良磨, 鈴木勇介, and 高田謙. 正規表現技術入門. 技術評論社, 2015.

[112] A I Marqués,V García,J S Sánchez.: On the suitability of resampling techniques for the class imbalance problem in credit scoring, Journal of the Operational Research Society, Volume 64, Issue 7, pp 1060-1070,(2013)

関連業績リスト

参考論文

・公表済み論文

[1]田中孝昌, 津田和彦. ″マルチソースアクセスログ分析によるボット検知手法の提案.″ 電気学会論文誌 C (電子・情報・システム部門誌) 141.11 (2021): 1205-1214.

[2]Takamasa Tanaka, Hidekazu Niibori, Li Shiyingxue, Shimpei Nomura, Hiroki Kawashima, and Kazuhiko Tsuda ″Bot Detection Model using User Agent and User Behavior for WebLog Analysis.″ Procedia Computer Science 176 (2020):1621-1625.

[3]Takamasa Tanaka, Hidekazu Niibori, Li Shiyingxue, Shimpei Nomura, Tadayoshi Nakao, and Kazuhiko Tsuda. ″Selenium based Testing Systems for Analytical Data Generation of Website User Behavior.″ In 2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW).IEEE,2020: 216-221.

[4]Takamasa Tanaka, Tomohiro Hamaguchi, Takumi Saigo and Kazuhiko Tsuda, ″Classifying and understanding prospective customers via heterogeneity of supermarket stores.″, Procedia Computer Science 112(2017):956-964.

その他の論文

・査読のない発表論文

[1]田中孝昌, 濱口智大, 西郷拓海, 津田和彦. “スーパーマーケットの店舗別販売傾向とRFM分析を利用した優良顧客分類”. 人工知能学会第二種研究会資料. 2017.