Web広告推薦のための閲覧カテゴリ情報を用いたユーザの潜在的興味分析方式

山口　由莉子† 　森下　民平†† 　稲垣　陽一†† 　中本　レン†† 　張　建偉†††
青井　順一†††† 　中島　伸介†††††

† 京都産業大学大学院 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
†† 株式会社きざしカンパニー　〒 103–0015　東京都中央区日本橋箱崎町 20-14 日本橋巴ビル 6F
††† 筑波技術大学 産業技術学部　〒 305–8520 茨城県つくば市 天久保 4 丁目 3-15
†††† 株式会社マイクロアド　〒 150-0045 東京都渋谷区神泉町 8-16 渋谷ファーストプレイス 5 階
††††† 京都産業大学 コンピュータ理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †i1658164@cc.kyoto-su.ac.jp, ††{mimpei,inagaki,reyn}@kizasi.jp, †††zhangjw@a.tsukuba-tech.ac.jp,
††††aoi_junichi@microad.co.jp, †††††nakajima@cse.kyoto-su.ac.jp

あらまし　企業が製品やサービスのために行う宣伝活動の一形態として，Web 広告が注目されている．ただし，現在の Web 広告推薦の主流であるキーワードマッチングをベースとした手法では，数多く存在するであろう潜在的な購買者層に対して効果的に Web 広告を推薦することは困難である．そこで，我々はユーザの潜在的興味を分析することで，より効果的な Web 広告推薦方式を実現することを目指した研究を進めている．我々はこの先行研究において，実データを用いたユーザモデルの構築を試み，学習データにおける適切なポジティブ・ネガティブ比や，分析対象となる閲覧履歴の取得期間が予測性能に及ぼす影響について検討した．しかし，先行研究での実験では学習データの特徴量としては FQDN を使用し，FQDN は異なるがページ内容（カテゴリ）が類似した Web サイトを閲覧しているユーザ間の類似度を適切に判定することができなかった．また，同一の FQDN でもページの内容（カテゴリ）が異なる Web サイトが存在した場合，その違いを捉えることができなかった．これらより FQDN によらない意味的な特徴量の採用が必要であるという知見を得ていた．そこで本稿では，FQDN だけではなく閲覧 Web ページのカテゴリ情報を用いたユーザの潜在的興味分析方式について提案を行うと共に，本方式の精度に関する評価実験を行い，カテゴリ情報を利用することの妥当性を確認した．併せて，閲覧履歴の取得期間について，最大 6 日分まで延長したデータを用いた比較実験結果についても述べる．

キーワード　Web 広告，ユーザプロファイリング，アクセスログ分析

1. はじめに

企業が行う製品やサービスの宣伝活用の一形態として，Web 広告が注目されている．Web 広告が注目されている要因としては，リアルタイムでユーザ個人に合わせた Web 広告を配信する仕組み，リアルタイムビッディング (RTB) [1]（図 1 参照）の普及が挙げられる．そしてまた，Web 広告推薦では，対象となる顧客ユーザの属性・嗜好に基づいた個別の広告を表示できるターゲティング性と，ユーザのマウス操作に合わせて能動的にアクションする等のインタラクティブ性を性質として持っており，従来の広告では実現できなかった新たな推薦が可能となっている．現在，ターゲティング性が考慮された Web 広告推薦方式としては，リスティング広告，興味関心連動型広告，リターゲティング広告等が挙げられる．これら Web 広告推薦方式では，ユーザの検索クエリや閲覧内容，及び属性等を考慮しているが，ユーザの潜在的興味を考慮した分析が行われているとは言えず，Web 広告を通じて購買者の層を広げるにはまだまだ改良の余地がある．従来の方式は既にユーザが興味を持ち，明確に認知しているキーワードの広告を掲載する，あるいは広告主サイトへのアクセス履歴があるユーザに広告を配信するものであり，広告主は潜在的興味を持つ新たな購買者，購買層を Web 広告によって獲得することが難しい．また，ユーザが興味を持っている場合や認知はしている場合でも，対象サイトにアクセスしていないという条件のみでそのユーザに対象 Web サイトの広告を提示しない事は広告主にとって機会損失であると言える．

そこで，我々はユーザの潜在的興味を分析することで，より効果的な Web 広告推薦方式を実現することを目指した研究を進めている．我々はこの先行研究 [2] において，ユーザの潜在的興味分析に基づく Web 広告推薦方式を提案すると共に，実データを用いたユーザモデルの構築を試み，学習データにおける適切なポジティブ・ネガティブ比や，分析対象となる閲覧履歴の取得期間が予測性能に及ぼす影響について検討した．この中で，先行研究にて行った評価実験では，分析対象となる閲覧

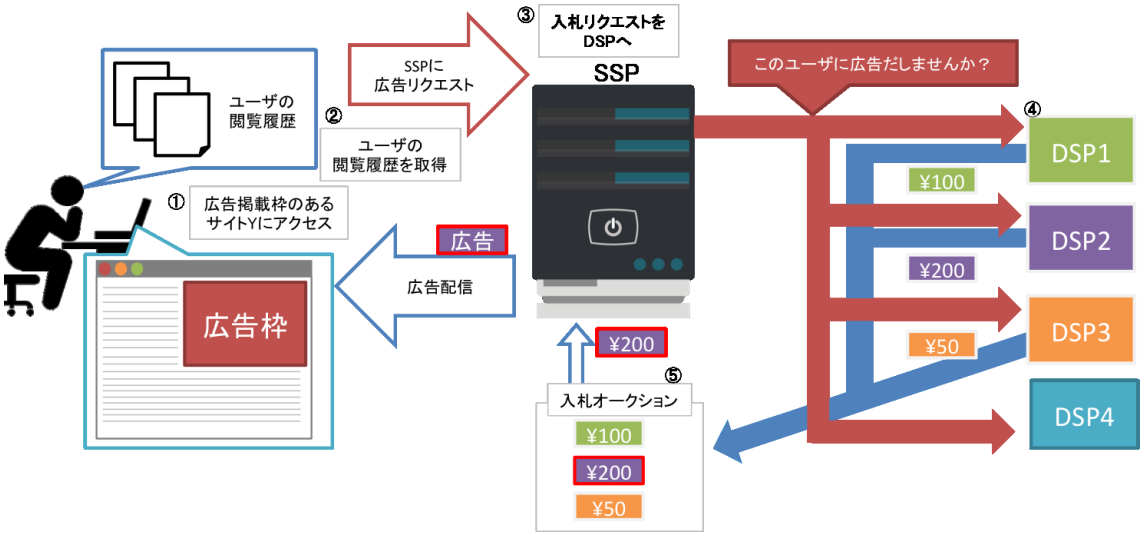

図 1　リアルタイムビッディング (RTB) 環境

履歴取得期間の最長が 24 時間と十分ではなく，適切な取得期間を評価するには至らなかった．

また学習データの特徴量としては FQDN を使用したが，FQDN は異なるがページ内容（カテゴリ）が類似した Web サイトを閲覧しているユーザ間の類似度を適切に判定することができず，また同一の FQDN でもページ内容（カテゴリ）が異なる Web サイトが存在した場合，その違いを捉えることができなかった．これらより FQDN によらない意味的な特徴量の採用が必要であるという知見を得ていた．

そこで本稿では，FQDN だけではなく閲覧 Web ページのカテゴリ情報を用いたユーザの潜在的興味分析方式について提案を行うと共に，本方式の精度に関する評価実験を行ったので報告する．なお，併せて閲覧履歴の取得期間について，最大 6 日分まで延長したデータを用いて行った比較実験結果についても報告する．

以下，2 節にて，関連研究について述べ，3 節にてこれまで先行研究において提案を行った，ユーザの潜在的興味に基づく Web 広告推薦方式について説明する．4 節では，各 Web ページのカテゴリを特徴量とする学習データを使用した場合の分類器作成の予備実験として分析対象とする閲覧履歴の最良な取得期間について述べる．その後，Web ページのカテゴリを特徴量とする学習データの抽出方法とその分類器の精度について述べる．最後に 5 節にて，まとめと今後の課題について述べる．

2. 関 連 研 究

以下に，Web 広告に関連した研究について述べ，我々の提案方式との差異を示す．

鈴木らは Web サイトのアクセスログと関連データを用いて消費者の購買行動を明らかにするため，購買行動に混合分布を当てはめて，購買サイクルを推定し購買の前後の行動の特徴を分析している [3]．また，生田目らは EC サイトのアクセスログと関連データを用いてサイト会員の日常の閲覧行動を考慮した購買予兆の発見モデルの提案 [4] をしており，また久松らはその購買予兆を発見するモデルをロジット・モデルを元に作成している [5]．以上の研究ではユーザの購買予兆を発見し広告を表示するという研究を行っているが，本研究では購買の予兆を発見するのではなく，閲覧しているユーザの潜在的な興味に基づいて広告を推薦するかどうかを決める事を目的としている．

内野らはユーザが次に見たい情報を予測し，それに関する広告配信する Web 広告配信システムを kMER およびマルコフモデルを応用した研究を行っている [6]．この研究はマルコフモデルを使っているため，閲覧履歴の時系列パターンに着目した取り組みといえる．しかしながら，多少時系列パターンが異なっていたとしても，同様の Web サイト群を閲覧しているのであれば，ユーザの興味関心は類似している，というのが本研究の立場である．すなわち本研究による提案手法の方がより幅広く類似ユーザの検索・発見することが可能であると考えている．Shuai Yuan らは RTB の概要と RTB の有用性について述べている [7] が，本研究では RTB を用いてユーザが認知していない商品を認知してもらうことによって Web 広告での宣伝活動がより活性すると考えている．Kuang-chih Lee らはユーザ，Web ページ広告をそれぞれ階層的にグループ化したものを組み合わせたときの CVR を推定している [8] が，本研究では Web サイトごとのユーザモデルによってユーザの潜在的興味を発見するという立場である．すなわち本研究による提案手法の方がより幅広く類似ユーザの検索・発見することが可能であると考えている．

3. ユーザの潜在的興味に基づく Web 広告推薦方式

本節では，3.1 節にて，RTB の処理の流れ，3.2 節にて，提案するユーザの潜在的興味に基づく Web 広告推薦方式の概要と処理について説明する．

3.1 リアルタイムビッディング (RTB) 環境

本研究で研究開発を行う Web 広告推薦方式では，リアルタイムビッディング (RTB) 環境での活用を想定している．そこ

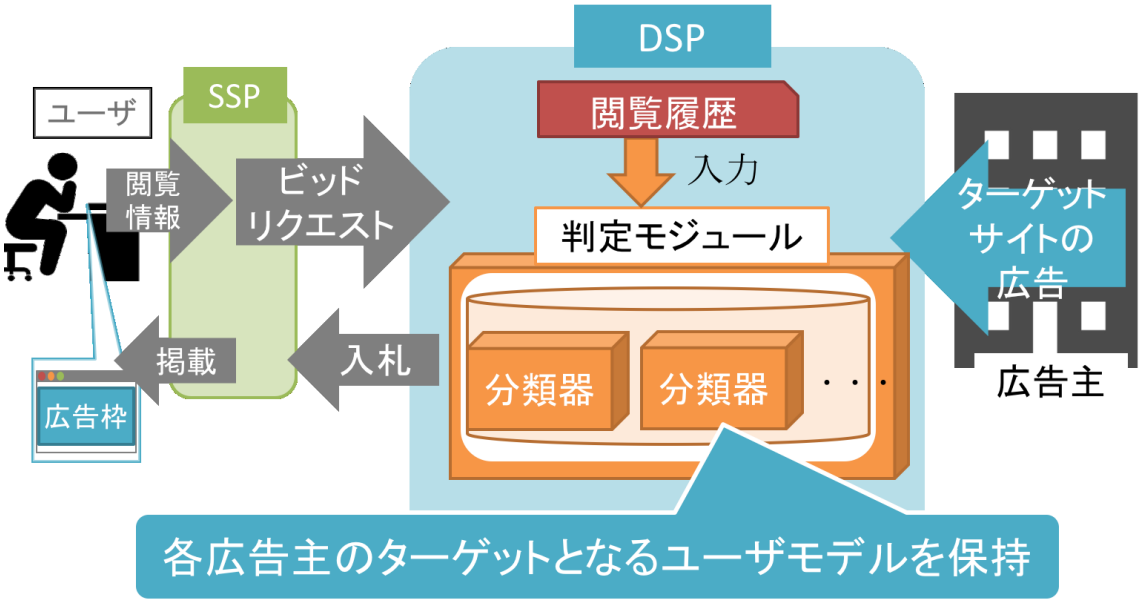

図 2　ユーザの潜在的興味に基づく Web 広告システム

でまずは，RTB について説明する（図 1 参照）．

RTB とは 1 配信 (1 インプレッション) 毎にリアルタイムで瞬時に，広告枠を買う側 (広告主や広告会社) が入札し，広告枠を供給する側 (媒体社) が入札価格の最も高い入札者に広告枠を売る仕組みのことである．SSP(サプライサイドプラットフォーム) とは，広告枠を供給する側，媒体社が使用するプラットフォームのことである．DSP(デマンドサイドプラットフォーム) とは，広告主や広告会社など，広告枠を買う側が使用するプラットフォームのことである．つまり，DSP が広告枠を買う側の都合の良い条件，(配信対象者や掲載面，配信時間など) をもとに入札し，SSP が最も高い入札価格を提示した DSP に広告枠を提供し (売り)，広告を配信するという関係である．

次に，同じく図 1 を用いて，RTB の処理の流れを説明する．まず広告枠のある Web サイト（図 1 の場合，サイト Y）にユーザがアクセスすると同時に広告タグを読み込む．すると，SSP に広告リクエストがかかる．SSP は DSP にビッドリクエストを行う．ビッドリクエストとはこのような広告枠に「こんなユーザがアクセスしましたが，広告を表示させますか」というリクエストのことである．ビッドリクエストには，アクセスしてきたユーザを識別する ID(クッキー) や IP アドレス，アクセスに用いたブラウザ（ユーザエージェント）などのユーザに関する情報や，また広告先を掲載する掲載先のドメインとコンテンツカテゴリー，広告枠の ID，広告のサイズなど，広告枠やその掲載先の情報（図 1 の場合，Y サイトの情報）などが含まれている．DSP 側は SSP から送られてきたビッドリクエストの分析を行う．例えば，ユーザは広告主が広告配信をしたいターゲットユーザであるか，広告掲載先は広告主の広告を閲覧するのに適した Web ページ，Web サイトであるかどうか，など瞬時に判断する．そして広告枠を買いたい金額を含めて，買いたい意思，ビッドレスポンスを SSP に送る．これが DSP の入札となる．図 1 の場合，DSP1,2,3,4 がそれぞれビッドリクエストの分析を行い，入札をするか，どの値段で入札するかを判断する．今回の場合は，DSP1,2,3, はそれぞれ￥100,￥200,￥50 で入札し DSP4 は広告枠を買わないと判断し，入札しなかった．SSP は各 DSP の中からの最も高い入札価格を提示した DSP に広告枠を提供し広告を配信する．そして最も高い値段で入札した DSP の広告を広告枠に入れ，広告を掲載する（インプレッション）．図 1 の場合，一番高額で入札した DSP2 が広告枠を落札し，DSP2 の広告が広告枠に掲載されるという流れである．

3.2　ユーザの潜在的興味に基づく Web 広告推薦方式

前節にて説明した従来の手法では，ユーザがすでに興味を持ち，認知しているキーワードに関連する広告を提示するものであり，広告主にとっては新しい購買者，購買層を Web 広告によって獲得することが難しい．そこで，ユーザの潜在的興味を分析することで，より効果的な Web 広告推薦を実現することが可能なユーザの潜在的興味に基づく Web 広告推薦方式について図 2 を用いて説明する．

この手法では，システムは事前にターゲットサイトにアクセスしたことのある他のユーザグループの閲覧履歴を取得する．この閲覧履歴を用いて，ターゲットサイトに訪れるユーザモデルを保持した分類器を作成する．この分類器は閲覧履歴より特定の特徴量をベクトル化し，scikit-learn [9] を用いロジスティック回帰を行い作成する．また，本研究の実験では特定の特徴量は FQDN とカテゴリを用いる．カテゴリについては 4.2 節にて説明する．つまり，閲覧履歴とターゲットサイトとの直接的な関連が小さい場合においても，ターゲットサイトに対してユーザの潜在的興味があるか推測することが可能になると考えられる．実際に広告推薦を行う場合は，アクセスしてきたユーザの閲覧履歴を取得し，事前に作成した分類器にてユーザがターゲットサイトに対して潜在的興味があるかを推定する．この方法により，今まで広告を配信していなかったユーザへの広告配信が可能となり，ターゲットサイトの広告効果を高められる．

つまり，ユーザの潜在的興味に基づく Web 広告推薦方式を用いることで新しい購買者，購買層を Web 広告によって獲得することが可能となる．以上が本研究の提案手法である．

4. 潜在的興味を持つユーザモデルの学習方法に関する実験的考察

本節では，ユーザの潜在的興味に基づく Web 広告推薦方式のユーザモデル学習方法に関して，実験に基づく検討を行ったので説明する．

本評価実験は，あるターゲットサイトを閲覧したユーザの閲覧履歴をポジティブデータとして学習し，同様の閲覧履歴を持つ他のユーザが当該ターゲットサイトを閲覧しているかどうかを調べることで，閲覧履歴である学習データの妥当性を検証するものである．本実験により，提案手法が新たな購買者・購買層を獲得できるか否かを評価することはできないが，提案手法で採用する閲覧履歴に基づく潜在的興味推定において，どのような学習データを用いることが適切か検討する上で重要である．

4.1 節ではユーザモデル構築用学習データである閲覧履歴取得期間が予測性能に及ぼす影響について，4.2 節ではユーザモデル構築用学習データの特徴量の違いが予測性能に及ぼす影響について述べる．

4.1 ユーザモデル構築用学習データである閲覧履歴取得期間が予測性能に及ぼす影響

本研究の最終的な目標である，ユーザの潜在的興味に基づく Web 広告推薦方式の実現のために，ユーザモデル構築用学習データである閲覧履歴の取得期間を適切に設定する必要がある．そこで本節では，この閲覧履歴の取得期間が予測性能に及ぼす影響に関する評価実験を行い，実験結果を踏まえて適切な閲覧履歴取得期間について検討する．

本評価実験では，“ターゲットサイトに訪れたことのあるユーザらの閲覧履歴” をポジティブデータ，“ターゲットサイトに訪れたことのないユーザらの閲覧履歴” をネガティブデータとし，双方をあわせて学習データとした．この学習データを用いて，学習することでユーザモデルを構築し，別途用意したテストデータに基づいて予測性能を評価した．本評価実験は，あるターゲットサイトを閲覧したユーザの閲覧履歴をポジティブデータとして学習し，同様の閲覧履歴を持つ他のユーザが当該ターゲットサイトを閲覧しているかどうかを調べることで，閲覧履歴である学習データの妥当性を検証するものである．本実験により，提案手法が新たな購買者・購買層を獲得できるか否かを評価することはできないが，提案手法で採用する閲覧履歴に基づく潜在的興味推定において，どのような学習データを用いることが適切かを検討する上で貴重なデータを得ることができると考えている．

先行研究 [2] では，ユーザモデル構築用学習データである閲覧履歴取得期間が予測性能に及ぼす影響について考察したが，先行研究で行った実験にて使用した学習データの取得期間が最長で 1 日であったため，本研究では学習データの最長取得期間を 6 日に延ばし，実験を行った．

前述の通り，“ターゲットサイトに訪れたことのあるユーザらの閲覧履歴” をポジティブデータ，“ターゲットサイトに訪れたことのないユーザらの閲覧履歴” をネガティブデータとし，双方をあわせて学習データとした．また，先行研究よりユーザモデル構築用学習データのポジティブ・ネガティブ比率は 1 ： 1 で問題ないという知見を得ているため，本実験では学習データのポジティブ・ネガティブ比率を 1 ： 1 で行うこととする．ターゲットサイトは以下の 7 つを使用した．

表 1　潜在的興味を持つユーザモデルの学習方法に関する実験データ

	ポジティブデータ（ターゲットサイトへアクセス済ユーザ）	ネガティブデータ（ターゲットサイトへアクセス未ユーザ）
学習データ	400 人分	400 人分
テストデータ	100 人分	10,000 人分

サイト (a):　天気情報サイト
サイト (b):　画像やイラストを投稿，閲覧するサイト
サイト (c):　2 ちゃんねるのまとめサイト
サイト (d):　オンライン小説，ケータイ小説を読む専用サイト
サイト (e):　映画情報サイト
サイト (f):　週刊誌のような情報，社会人向けサイト
サイト (g):　ビジネスニュースサイト，ビジネスマン向けサイト

閲覧履歴の取得期間は 1〜6 日の 6 パターン用意した．実験回数はそれぞれ 10 回行った．また，閲覧履歴のデータにはユーザ ID，アクセス URL，アクセス日時が含まれる．特徴量はアクセス FQDN を用いた．アクセス FQDN とはアクセス URL の FQDN である．FQDN(Fully Qualified Domain Name) とは，ホスト名，ドメイン名（サブドメイン名）を省略せずに指定した記述形式のことである．なお，学習には scikit-learn [9] を用いてロジスティック回帰を行なった．実験手順を以下に示す．

手順 1　取得した閲覧履歴を取得期間 1〜6 日の 6 パターンを用意する．ポジティブデータはターゲットサイトにアクセスした事のあるユーザの閲覧履歴，ネガティブデータはターゲットサイトにアクセスした事のないユーザの閲覧履歴を使用する．また，それぞれの学習データ，テストデータは表 1 の内容を用いた．

手順 2　特徴量を FQDN とし，6 パターンそれぞれ，scikit-learn を用い，アルゴリズムはロジスティック回帰を使用する．手順 1 で作成した学習データを学習させて分類器を構築する．

手順 3　手順 2 で作成した分類器を用いて，手順 1 で作成したテストデータをテストし，分類器の性能を求める．

上記手順により行った実験結果については，次節にて考察する．また，テストデータのポジティブデータとネガティブデータの差が大きい理由は，世の中にあるポジティブデータが極端に少なくネガティブデータが明らかに多いという実際のデータ比率に近づけるためである．

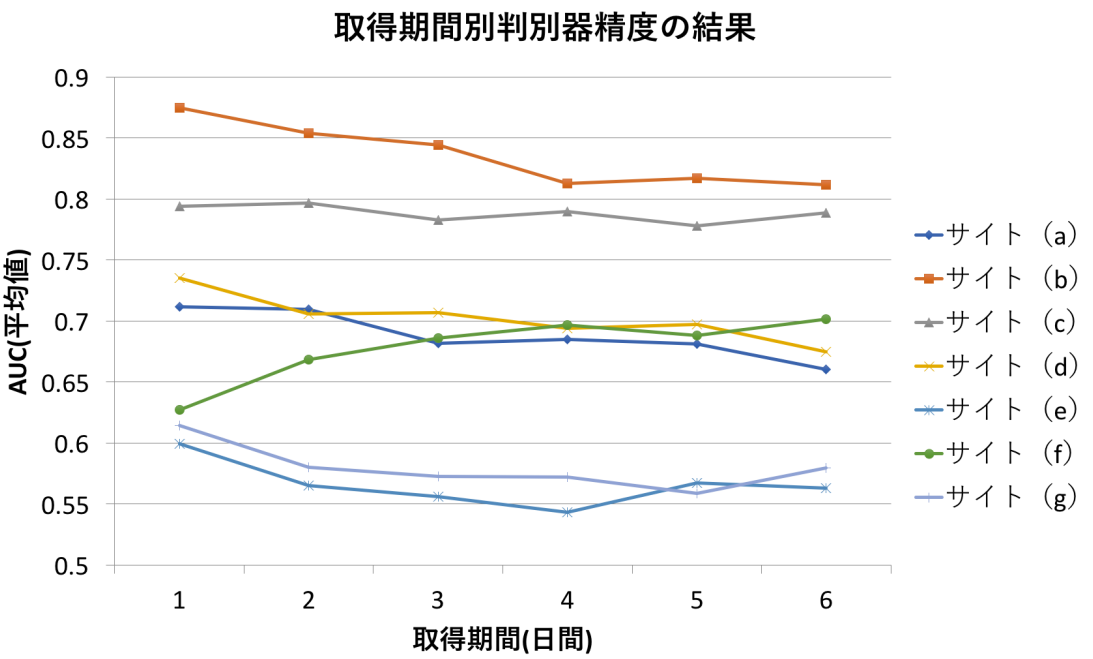

図 3　閲覧履歴取得期間が予測性能の及ぼす影響に関する実験の AUC 値の平均結果 (特徴量:FQDN)

表 2　閲覧履歴取得期間が予測性能の及ぼす影響に関する実験の AUC 値の平均結果表 (特徴量:FQDN)

取得期間	サイト (a)	サイト (b)	サイト (c)	サイト (d)	サイト (e)	サイト (f)	サイト (g)
1 日	0.7118	0.8746	0.7938	0.7352	0.5995	0.6272	0.6143
2 日	0.7096	0.8538	0.7967	0.7057	0.5650	0.6685	0.5799
3 日	0.6819	0.8442	0.7831	0.7070	0.5559	0.6859	0.5727
4 日	0.6849	0.8127	0.7895	0.6938	0.5432	0.6968	0.5721
5 日	0.6814	0.8172	0.7782	0.6972	0.5672	0.6879	0.5585
6 日	0.6602	0.8119	0.7889	0.6747	0.5632	0.7016	0.5795

なお本研究で構築する分類器の性能を判定する尺度としては，AUC を用いる．AUC を用いる理由を説明するために, 予測件数 1 万人のうち実際に閲覧したユーザが 1 人だったケースを考察する．

このとき，常に全員が閲覧していないと回答する予測モデルの正確度 (accuracy) は 99.99 %と高いが，このモデルは明らかに役に立たない．このような場合のようにクラス分布に大きな偏りがある 2 クラス分類問題では，予測モデルの性能評価指標に AUC [10] を用いるのが一般的である．予測モデルを用いた予測時に，閲覧する可能性が高いユーザほどより大きな実数値スコアが出力されるとする．あるスコアを閾値としたとき，閾値以上のユーザを閲覧したユーザ，閾値未満を閲覧していないユーザと判定すると，その閾値を選択した場合の真陽性率 (true positive rate，本当に閲覧したユーザを正しく閲覧したと判定した割合)，偽陽性率 (false positive rate，実際には閲覧しなかったユーザを誤って閲覧したと判定した割合) とを求められる．縦軸に真陽性率，横軸に偽陽性率を取り，閾値をスコアの高い方から低い方へ全ユーザが閲覧したユーザと見なされるまで動かしながらプロットしたものを ROC 曲線 (Receiver Operating Characteristic Curve) と呼び，ROC 曲線の下側面積を AUC(Area Under ROC Curve) と呼ぶ．閾値とするスコアを s，真陽性率と偽陽性率を返す関数をそれぞれ TPR，FPR とすると，AUC は定義より，式 (1) で表される．

$$AUC = \int_{\infty}^{-\infty} TPR(s)\, FPR'(s)\, ds \qquad (1)$$

ランダムな予測結果を返すモデルの AUC は 0.5 となり，必ず予測を的中させるモデルの AUC は 1.0 となる．したがって，少なくとも 0.5 以上の数値でなければ意味がなく，また 0.5 よりも明らかに大きな値であることが望ましい．

7 つのターゲットサイト (サイト (a〜g)) において，ユーザモデル構築用学習データである閲覧履歴の取得期間が予測性能に及ぼす影響について評価するために行った実験結果を図 3，表 2 に示す．図 3，表 2 に示した値は 10 回実験を行った AUC 値の平均値である．

図 3, 表 2 が示す通り，7 サイト中 1 サイト（サイト (f)）のみが最も取得期間が長かった 6 日が最も性能が高い結果となった．また，7 サイト中 5 サイトでは最も取得期間が短かった 1 日が最も性能が高い結果となった．先行研究で行なった実験は

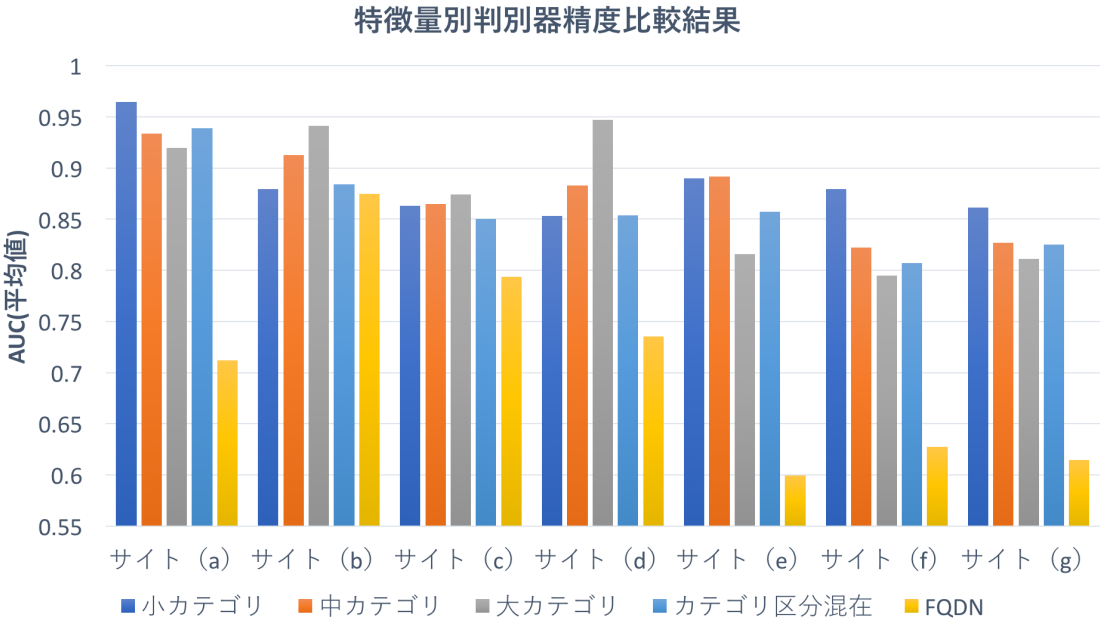

図 4　特徴量が予測性能の及ぼす影響に関する実験の AUC 値の平均結果

表 3　特徴量が予測性能の及ぼす影響に関する実験結果の AUC 値の平均

特徴量 (特徴量数)	サイト (a)	サイト (b)	サイト (c)	サイト (d)	サイト (e)	サイト (f)	サイト (g)
小カテゴリ (837 個)	**0.9646**	0.8794	0.8633	0.8530	0.8897	**0.8794**	**0.8611**
中カテゴリ (274 個)	0.9335	0.9126	0.8647	0.8831	**0.8918**	0.8225	0.8270
大カテゴリ (23 個)	0.9196	**0.9411**	**0.8740**	**0.9470**	0.8157	0.7949	0.8109
カテゴリ区分混在 (1134 個)	0.9388	0.8840	0.8504	0.8537	0.8574	0.8071	0.8254
FQDN(6,355,942 個)	0.7118	0.8746	0.7938	0.7352	0.5995	0.6272	0.6143

表 4　特徴量が予測性能の及ぼす影響に関する実験結果の AUC 値の標準偏差

特徴量 (特徴量数)	サイト (a)	サイト (b)	サイト (c)	サイト (d)	サイト (e)	サイト (f)	サイト (g)
小カテゴリ (837 個)	0.0131	0.0562	0.0403	0.0329	0.0635	0.0562	0.0452
中カテゴリ (274 個)	0.0366	0.0650	0.0498	0.0380	0.0268	0.0440	0.0389
大カテゴリ (23 個)	0.0274	0.0304	0.0376	0.0514	0.0340	0.0725	0.0503
カテゴリ区分混在 (1134 個)	0.0283	0.03130	0.0320	0.0294	0.0320	0.0205	0.0273
FQDN(6,355,942 個)	0.0429	0.0735	0.0759	0.0328	0.0554	0.0678	0.0527

比較した取得期間について，3 時間，12 時間，1 日のうち，7 サイト中 5 サイトでは最善は 1 日であり，ある Web サイトに対して興味を持つユーザモデルの学習においては，分析対象とする閲覧履歴の取得期間が長い方が効果的であるという仮説の妥当性を支持する結果であった．しかしながら本実験の結果より，すべてのターゲットサイトにおいて興味を持つユーザモデルの学習においては，分析対象とする閲覧履歴の取得期間が長い方が効果的であるとは限らず，各ターゲットサイト毎に適切な閲覧履歴取得期間が異なると考えられる．また，過去であればあるほどユーザの記憶が薄れ興味が減っているので閲覧履歴の取得期間が長い方が効果的であるとは限らないという結果になったのではないかとも考えれられる．よって今後の研究課題として過去であればあるほどユーザの記憶が薄れる事を特徴量に盛り込み，実験をすることが挙げられる．また，サイトごとに最適な取得期間が異なり，今回実験した範囲では特定の傾向を見出せなかったため，さらなる検討が必要である．

4.2 節での学習データの特徴量の違いが予測性能に及ぼす影響についての検討実験では同一のターゲットサイトを使用するため，上記の結果に伴ってターゲットサイトすべての閲覧履歴取得期間を 1 日とする。

4.2　ユーザモデル構築用学習データの特徴量の違いが予測性能に及ぼす影響

先行研究 [2] では，学習のデータの特徴量に FQDN を使用していたが，FQDN が異なるが意味的に類似するサイト間の類似性を考慮することは難しい．また，同一の FQDN でもページの内容 (カテゴリ) が異なる Web サイトが存在した時，特徴量を FQDN とした場合では，その違いを捉えることができない．例えば，Yahoo!ニュースの場合の FQDN は news.yahoo.co.jp と同一の場合でも記事内容がスポーツやエンタメと，カテゴリは多岐に渡る．そこで本実験では，各 URL(Web ページ) のカテゴリを特徴量とする学習データを抽出し，実験的考察を行った．

4.2.1 学習データの特徴量に関する実験手順と実験データ

閲覧履歴の取得期間は 4.1 節での評価実験の結果を踏まえ，1 日とした．実験回数はそれぞれ 10 回行った．なお，学習には scikit-learn を用い，アルゴリズムとしてはロジスティック回帰を使用した．実験手順は 4.1 節と同様である．ただし手順 2 における特徴量をカテゴリ (小，中，大，カテゴリ区分混合) とし，学習データを学習させて分類器を構築する．実験条件は 4.1 節と同様であるが，特徴量にカテゴリを加えた．カテゴリ分類については先行研究 [11] によって作成されたカテゴリ分類器を使用した．学習データ・予測データともに，クロールした Web ページの内容を形態素解析し，得られた語句の頻度を特徴量とした．また，カテゴリ分類器にはライブラリとしては liblinear を用い，アルゴリズムとしてはロジスティック回帰を使用している．カテゴリ区分は　大，中，小の 3 区分用意した (表 5 参照)．

表 5　大・中・小のカテゴリ区分の例

大カテゴリ	中カテゴリ	小カテゴリ
ファッション	服飾雑貨	ジュエリー
ファッション	服飾雑貨	バック
食料品	食材	果物＆野菜
食料品	レストラン	ファーストフード
不動産	不動産購入	中古戸建て

本研究で使用したユニークカテゴリ数，すなわち特徴量を大カテゴリにした場合の特徴量ベクトルの次元数は 23 個，特徴量を中カテゴリにした場合の特徴量ベクトルの次元数は 274 個，特徴量を小カテゴリにした場合の特徴量ベクトルの次元数は 837 個用意した．また，大カテゴリ，中カテゴリ，小カテゴリ全てを混合したカテゴリ区分混合の次元数は全カテゴリの合計である 1,334 個である．実験結果については，次節にて考察する．

4.2.2 学習データの特徴量に関する実験結果および考察

7 つのターゲットサイト (サイト (a〜g)) において，ユーザモデル構築用学習データの特徴量の違いが予測性能に及ぼす影響について評価するために行った実験結果を図 4，表 3，表 4 に示す．

図 4，表 3，表 4 が示す通り，7 サイト全てが特徴量が FQDN よりカテゴリの方が分類器の精度が高くなった．前節でも述べたが，特徴量は FQDN では意味的に類似するサイト間の類似性を考慮することができないが，カテゴリを特徴量にすることによって類似するサイト間の類似性を考慮することができるのではないかと考えられる．また，ターゲットサイトによって最も分類器の精度が高くなる特徴量のカテゴリ区分が違うことがわかった．ターゲットサイトによって最良のカテゴリ区分は異なるが，それぞれの最良のカテゴリ区分と FQDN との分類器の精度には有意な差が見られた．またカテゴリ区分混在特徴量は，FQDN 特徴量より精度は高かった．しかしカテゴリ区分 (大・中・小) 別特徴量と比べると，精度が最善になることはなく，ターゲットサイト (c) のように，カテゴリ特徴量のなかではもっとも精度が低くなる場合もあった．

今回の実験では，精度最善なカテゴリ区分はサイト依存で，特定の区分を常に用いれば良いわけではないことが判明した．したがって，大・中・小カテゴリ区分の自動選別，あるいは一部の大カテゴリを小カテゴリまで展開するといったカテゴリの部分展開探索を今後の課題とする．

5. ま と め

本稿では，我々が研究を進めている，ユーザの潜在的興味分析に基づく Web 広告推薦方式に関する研究において，FQDN だけではなく閲覧 Web サイトのカテゴリ情報を用いたユーザの潜在的興味分析方式について提案を行うと共に，本方式の精度に関する評価実験を行ったので報告した．なお，併せて閲覧履歴の取得期間について，最大 6 日分まで延長したデータを用いて行った比較実験結果についても報告した．

学習データの特徴量については，FQDN を使用した場合と Web ページのカテゴリを使用した場合とではカテゴリを使用した場合の方が分類器の精度が良い結果となった．ただし，学習データの特徴量にて採用するカテゴリ区分（大・中・小）については，ターゲットサイトによって最適な区分が異なる結果となった．したがって，今後は最適なカテゴリ区分の選定方式についても検討する必要がある．

最適な閲覧履歴の取得期間に関する評価では，{ 1 日，2 日，3 日，4 日，5 日，6 日} の 6 種類の取得期間に対して実験を行ったが，ダーゲットサイト 7 中 6 サイトで取得期間 1 日が最も良い結果となった．ただし，サイトごとに最適な取得期間が異なり，今回実験した範囲では特定の傾向を見出せなかったためさらなる検討が必要である．また，単純な取得期間の長さだけでなく，午前，午後，通学通勤の時間帯，週末など，ユーザの生活に沿った期間を取得期間として調べる必要があると考えている．

謝　　辞

本研究の一部は，JSPS 科研費 26330351，26870090 による．ここに記して謝意を表します．

文　　献

[1] 横山隆治，菅原健一，楳田良輝，DSP/RTB　オーディエンスターゲティング入門—ビッグデータ時代に実現する「枠」から「人」への広告革命—，インプレス R & D，2012 年.
[2] 山口由莉子，森下民平，稲垣陽一，中本レン，張建偉，青井順一，中島伸介，ユーザの潜在的興味に基づく Web 広告推薦方式の検討, 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2016) B1-2，2016 年.
[3] 鈴木元也，生田目崇，購買前後のアクセスを考慮した Web サイトの顧客行動分析，日本オペレーションズ・リサーチ学会 2012 年秋季研究発表会 (2-F-3)，2012 年.
[4] 生田目崇，朝日真弓，久松俊道，外川隆，顧客の閲覧行動を考慮した購買予兆発見モデル，日本オペレーションズ・リサーチ学会 2012 年秋季研究発表会 (2-F-2)，2012 年.
[5] 久松俊道，外川隆，朝日真弓，生田目崇，EC サイトにおける購買予兆発見モデルの提案，オペレーションズ・リサーチ：経営の科学，2013 年.
[6] 内野英治，森田博彦，下野雅芳，Web 広告動的配信システム

へのマルコフモデルと kMER の応用，22nd Fuzzy System Symposium(Sapporo,Sept.6-8,2006)6B1-1，2006 年.
[7] Shuai Yuan，Jun Wang，Xiaoxue Zhao，Real-time bidding for online advertising: measurement and analysis，Proceedings of the Seventh International Workshop on Data Mining for Online Advertising，2013 年.
[8] Kuang-chih Lee，Burkay Orten，Ali Dasdan，Wentong Li，Estimating Conversion Rate in Display Advertising from Past Performance Data，Proc. of KDD 2012，2012 年.
[9] scikit-learn: Machine Learning in Python. http://scikit-learn.org
[10] 平井 有三，はじめてのパターン認識，森北出版，2012 年.
[11] 森下民平，稲垣陽一，青井順一，オンライン広告データ分析と分析基盤， SOFTECHS Vol. 34 No. 1 P36-P40，2016 年.