DEIM Forum 2019 I2-5

Web広告推薦のためのユーザの興味分析に基づくWeb閲覧予測手法の提案

澤田　礼我† Panote Siriaraya† 森下　民平†† 稲垣　陽一†† 中本　レン††
張　建偉††† 中島　伸介††††

† 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
†† 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
††† 岩手大学 理工学部 〒020–8551 岩手県盛岡市上田 4–3–5
†††† 京都産業大学 情報理工学部 〒603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1544719,k6180,nakajima}@cc.kyoto-su.ac.jp, ††{mimpei,inagaki,reyn}@kizasi.jp,
†††zhang@iwate-u.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．我々は先行研究において，潜在的興味判別器の精度は，学習データの特徴量に 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が良いという知見を得た．また,「分析期間の長短にそれぞれメリット、デメリットがあり、これらを併用した方が良い」という仮説を立て，検証実験を行った．その結果，分析期間を長短併用とする事で，精度が有意に良くなるという知見を得た．この知見から，我々は「分析期間を長短に分けると精度が良くなるのであれば，その期間内での順序を取った方がより良い」という仮説を立てた．また，先行研究では，学習方法がロジスティック回帰であるが，それ以外の学習方法についての検証がなかった．よって，まずはロジスティック回帰以外の学習方法とロジスティック回帰について学習を行い、どの学習方法がより適切な結果を与えるかを検証する．
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．2016 年，日本の Web 広告媒体費は初めて 1 兆円を越え [1]，ますます市場の拡大が見込まれる．Web 広告が注目されている要因の一つに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) [2]（図 1 参照）の普及が挙げられる．そしてまた，Web 広告推薦は，対象となるユーザの属性・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有しており，従来では実現できなかった新たな広告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，およ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，ユーザが対象サイトに対して興味を持っている場合，および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こと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きた．我々は先行研究 [3] において，短期的興味と長期的興味併せたユーザの潜在的興味に基づく Web 広告推薦方式にを提案し，興味判定器についての評価実験を行った．その中で潜在的興味判定器の精度は，閲覧期間を長短に分けると有意に良くなるという知見を得た.「分析対象 (学習データ) となる閲覧履歴を長短に分けた事で精度が良くなったのは，学習データが時系列データに近づいた為」という仮説を立てた．本研究では，閲覧履歴について期間内での回数という形でなく，Web サイトを閲覧した順序を基に学習を行い，潜在的興味判定器を評価する．また，先行研究ではロジスティック回帰のみによる実験であった為，本研究では，まずロジスティック回帰とそれ以外のアルゴリズムの

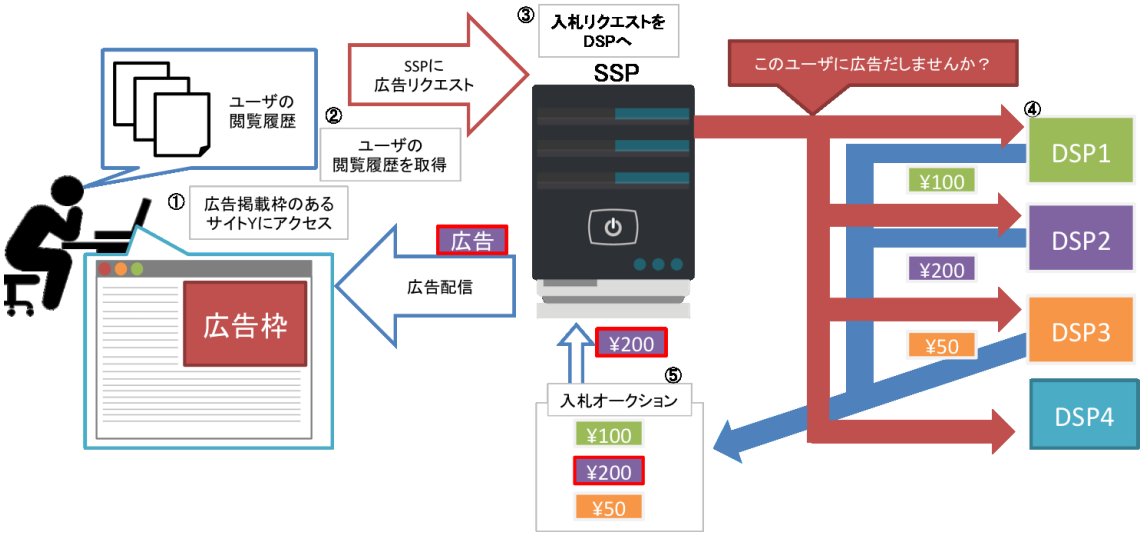

図 1　リアルタイムビッディング (RTB) 環境

比較実験を行い，その後順序データを用いた興味判定器の評価実験を行う．

以下，2 節にて，関連研究について述べ，3 節では順序データを用い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，実験の条件と結果について述べる．5 節では，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 [4]．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 [5] 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 [6]．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否かを決める事を目的としている．

Web 広告における RTB のさまざまな側面を検討するため，数多くの学術研究が発表されている．入札プロセスそのものに焦点を当て，RTB プロセスでの落札可能な価格を予測する研究 [7] や, コンバージョンされた際の利益に基づいて最適化された円滑な予算配分へのアプローチを提供する研究 [8]，リアルタイム入札のための最適戦略の特定についての研究がある [9]．また，RTB の概要と RTB の有用性について述べている研究 [10] や Web 広告の供給側から広告目録の最適な価格設定の調査が行われている [11]．さらに，コンバージョン率（特定の広告を見たユーザが行動を起こす確率）をリアルタイム入札環境でどのように見積もることができるか [12]，クリック率（CTR）がオンライン広告のマルチメディア機能からどのように予測できるか [13] についての研究などがある．また，Web 広告の推薦システムに関しては, 特定のウェブサイトに対して Web 広告の作成に役立つ複数のキーワードを自動的に抽出し, 推薦するシステムの開発が検討されている [14]．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 [15] 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がより幅広く類似ユーザの検索・発見することを可能にすると考えている．

3. 順序データを用いた Web 広告推薦方式

本節では，提案する順序データを用いた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手順について説明する．

従来の手法では，ユーザが既に興味を持ち，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，本研究で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 広告推薦を目指す．

先行研究では「分析対象となる閲覧履歴のデータは分析期間の長短にメリットとデメリットがあり，長期と短期に分割する事で興味判定器の精度が良くなる」という仮説を立て，検証実験を行った．その結果，興味判定器の精度は有意に良くなると言う結果を得た．我々はこの結果から,「分析期間を分ける事が，どの順序でサイトを閲覧したかの関係を作った」と考えた．そこで，Web 閲覧履歴の順序を考慮した潜在的興味分析を提案する．Web 閲覧履歴の順序を考慮した潜在的興味分析方式器について，図 2 を用いて説明する．この手法では，システムはターゲットサイトにアクセスした事のあるユーザの，直近の Web

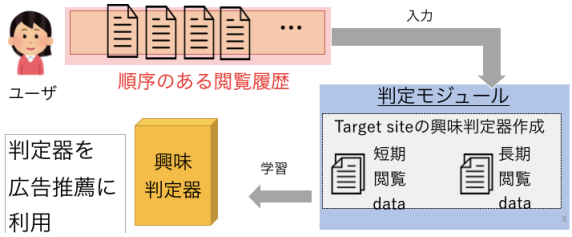

図 2　提案手法の流れ

閲覧履歴の順序を取得する．ユーザの日々揺れ動く興味の変化は，閲覧履歴の順序として現れる．この閲覧履歴の順序を取得する事で，ユーザの興味の移り変わりを保持した判定器を作成する事が出来る．閲覧履歴より，特定の特徴量をベクトル化して学習を行うが，本研究に適切なアルゴリズムが何であるかを検証する必要性がある．よって，まずはアルゴリズムについて比較検討の実験を行う．尚，比較実験では，scikit-learn [16] のアルゴリズムから従来のロジスティック回帰に加え，Random Forest，SVM を，scikit-learn 以外では XgBoost [17] について検討を行う．また，本研究の実験では特定の特徴量は先行研究 [18] の結果より，Web ページのカテゴリを用いる．カテゴリについては 4.1 節にて説明する．

広告推薦を行う場合は，ユーザの閲覧履歴の順序を取得し，事前に作成した判定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．この方法によって，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，ターゲットサイトへの広告効果を高める事が可能となる．

以上が本研究の提案手法である．次節では，本研究の為の予備実験について述べる．

4.　興味判定器のアルゴリズム別比較実験

先行研究では，興味判定器の作成にロジスティック回帰を利用していたが，これ以外のアルゴリズムでより良い精度が得られるものがないか検討する為，比較実験を行った。以下に示す表 1 がその比較実験に於いて使用したアルゴリズムとパラメータの表である．

表 1　学習時のアルゴリズムの比較実験の条件

アルゴリズム	パラメータ調整
XgBoost	'max_depth':[2,4,6],'n_estimators':[50,100,200]
SVM	'kernel':['rbf'],'C':[0.1,1,100]
	'kernel': ['linear'],'C': [0.1,1,100]
ロジスティック回帰	'C':[0.1,1,10,100]
Random Forest	'max_depth':[80,100],'max_features':[2,3]

なお本研究で構築する判定器の性能を判定する尺度としては，AUC [19] を用いる．AUC は式 (1) で表される．

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

TPR は真陽性率 (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)，FPR は偽陽性率 (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) を表す．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，AUC が 0.5 より明らかに高いアルゴリズムは精度が高いという事になる。本実験は AUC 値がロジスティック回帰よりも高くなるアルゴリズムが存在するか，存在しないかを確認する．

また，実験は 5 回行い，各アルゴリズムについて，AUC 値の平均を取る．

本実験での学習期間は，長期期間を 29 日間 (3/1 3/29)，短期期間を 1 日間 (3/30) とし，正解ラベルとなる，ターゲットサイトへのアクセス期間は，3 月 31 日に設定した．閲覧履歴のデータにはユーザ ID，アクセス URL，アクセスページカテゴリ，アクセス日が含まれる．アクセスページのカテゴリは，表 2 に示すように大・中・小のカテゴリを持つ．

表 2　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 1,134 個である．本実験では大カテゴリ，中カテゴリ，小カテゴリ全てを混合したカテゴリ区分を用いた．つまり，単一区間での特徴量の次元数は 1,134 個，短期と長期と分けた場合の特徴量は短期期間での特徴量の次元数 (1,134 次元) と長期期間での特徴量の次元数 (1,134 次元) の合計なので　2,268 個使用した．

同じターゲットサイト，同じ特徴量を用い，学習データ，テストデータをランダムに 5 回選択し実験を行った．実験手順を以下に示す．

手順 1　取得した閲覧履歴を上述した学習期間の内容に沿って短期の閲覧履歴と長期の閲覧履歴に分ける．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短期の閲覧履歴と長期の閲覧履歴，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短期の閲覧履歴と長期の閲覧履歴を使用する．学習データでは，ポジティブデータ，ネガティブデータ共に 100 人分，テストデータではポジティブデータは 100 人分，ネガティブデータは 10,000 人分のデータを使用した．また，学習データとテストデータは実験毎にランダムで取得する．

手順 2　特徴量を Web ページのカテゴリとし，scikit-learn を用い，手順 1 で作成した学習データを学習させて

判別器を構築する．アルゴリズムは XgBoost，Grid Search SVM，ロジスティック回帰，Random Forest を使用する．

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

上記手順により 4 個のターゲットサイトそれぞれの興味判定器を作成し，精度を求めた．また、実験時のターゲットサイトは、以下の通りである．

サイト A	自動車の口コミサイト
サイト B	地域別ニュースまとめサイト
サイト C	映画情報サイト
サイト D	小説投稿サイト

実験結果については，次節にて考察する．

a）実験結果

4 個のターゲットサイトに於いて，アルゴリズムの違いが予測性能に及ぼす影響を評価する為に行った実験結果を図 3 に示す．尚，結果は各アルゴリズムについて最も高い精度を得られるハイパーパラメータ (表 1) についてのみ示している．

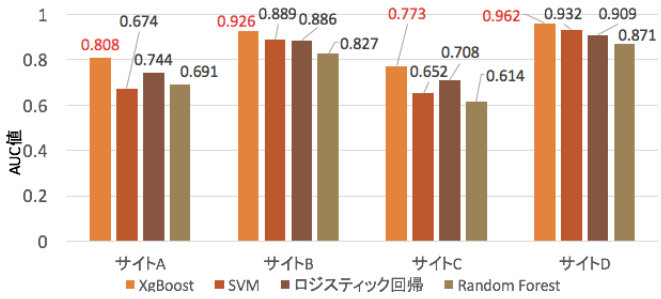

図 3　アルゴリズム別 AUC 値の比較

図 3 が示す通り，両方のサイトで AUC 値は XgBoost が最も性能が高い結果となった．この結果から順序データでの実験時には，XgBoost を利用する．

5. 順序データを用いた興味判別器の予測精度評価実験

Web 閲覧順序データによる実験は以下の手順で行う．

手順 1　取得した閲覧履歴の，学習期間の内容に沿って，各ユーザの閲覧履歴の順序を期間内の最後から 100 件まで取得する．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．学習データは，ポジティブデータ，ネガティブデータ共に 100 人分，テストデータではポジティブデータは 100 人分，ネガティブデータは 10,000 人分のデータを使用した．また，学習データとテストデータは実験毎にランダムで取得する．

手順 2　本実験では実験条件の制約上，特徴量を Web ページの大カテゴリの閲覧順序とした．手順 1 の通り取得する閲覧履歴は最大 100 件である事と，大カテゴリは 23 次元である事から，2,300 次元のデータとなる．前節の結果から XgBoost を用い，手順 1 で作成した学習データを学習させて判別器を構築する．

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

a）実験結果

4 個のターゲットサイトに於いて，アルゴリズムの違いが予測性能に及ぼす影響を評価する為に行った実験結果を図 4 に示す．

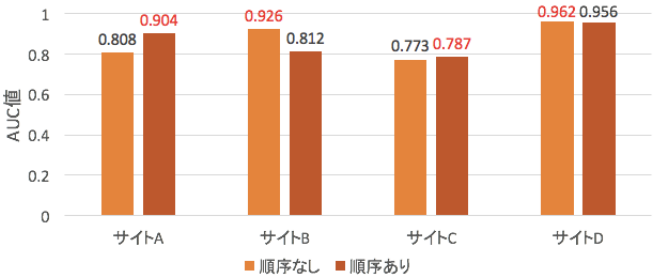

図 4　順序データによる比較

図 4 が示す通り，4 個のターゲットサイト中，2 個のサイトでは精度が向上したが，2 個のサイトでは精度が下がる結果となった．

利用したデータの形式が時系列を持つデータになっているが，XgBoost が時系列データに特化したアルゴリズムでない事が影響した可能性がある．一方，ターゲットサイトによっては精度が向上する場合もあり，どのような条件で精度が向上するか調査する為には，更に同様の実験で調査を進める必要があると考えられる．

6. 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，学習データとなるユーザの閲覧履歴の，順序を用いることによる，ユーザの興味分析方式について提案した．併せて，先行研究の手法に於ける適切なアルゴリズムの検討を行い，提案手法による興味判定器の作成を行い，予測精度の評価を行ったので報告した．改良案としては，RNN や LSTM のような，時系列データを扱う事に長けたアルゴリズムを利用する方法が考えられる．また，実験条件について，取得する閲覧履歴の件数，学習期間，訓練データの人数などは現状吟味出来ていない為，これらを変更する事で精度が向上する可能性がある．

謝　　辞

本研究の一部は，JSPS 科研費 17H01822 による．ここに記して謝意を表します．

文　　献

[1] 株式会社　電通，２０１６年日本の広告費，２０１６
[2] 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.
[3] 山口由莉子，Panote Siriaraya，森下民平，稲垣陽一，中本レン，張建偉，青井順一，中島伸介，Web 広告推薦のための長期的・短期的興味を考慮したユーザの潜在的興味分析方式 第 10 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 2018)，C5-1 2018 年.
[4] 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.
[5] 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.
[6] 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.
[7] W. C. H. Wu, M. Y. Yeh and M. S. Chen, ”Predicting winning price in real time bidding with censored data.” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1305-1314). ACM., 2015.
[8] K. C. Lee, A. Jalali and A. Dasdan, ”Real time bid optimization with smooth budget delivery in online advertising.” in Proceedings of the Seventh International Workshop on Data Mining for Online Advertising (p. 1). ACM., 2013.
[9] W. Zhang, S. Yuan and J. and Wang, ”Optimal real-time bidding for display advertising.” in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1077-1086). ACM., 2014.
[10] 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.
[11] A. Radovanovic and W. D. Heavlin, ”Risk-aware revenue maximization in display advertising.” in Proceedings of the 21st international conference on World Wide Web (pp. 91-100). ACM., 2012.
[12] K. C. Lee, B. Orten, A. Dasdan and W. Li, ”Estimating conversion rate in display advertising from past erformance data.,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 768-776) , 2012.
[13] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao and X. Fern, ”The impact of visual appearance on user response in online display advertising.” in Proceedings of the 21st International Conference on World Wide Web (pp. 457-458). ACM, 2012.
[14] S. Thomaidou and M. Vazirgiannis, ”Multiword keyword recommendation system for online advertising.” in International Conference on Advances in Social Networks Analysis and Mining (ASONAM (pp. 423- 427), IEEE, 2011.
[15] 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.
[16] scikit-learn: Machine Learning in Python. http://scikit-learn.org
[17] Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.
[18] 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.
[19] 平井 有三，はじめてのパターン認識，森北出版，2012 年.