画像広告の効率的な検索のための転置インデックス構築

茂木　哲矢† 　田頭　幸浩† 　小野　真吾† 　田島　玲†

† ヤフー株式会社　〒 107–6211 東京都港区赤坂 9-7-1 ミッドタウン・タワー
E-mail: †{tmotegi,yutagami,shiono,atajima}@yahoo-corp.jp

あらまし　オンライン広告はインターネットの経済を支える大きな柱の一つであり，ビジネスと学術の両方から大きな注目を浴びている．オンライン広告の課題の 1 つに，ユーザにとって適切な広告をどのように選択するかというものがある．テキスト広告では，どれくらい単語が重複するかによって類似度を定義し，この類似度を用いて情報検索に基づく手法で選択している．しかしながら，画像広告では類似度を用いた情報検索に基づく手法を用いることができない．本稿では，オンライン広告の中でも特に画像広告を対象とし，ユーザに適合するような画像広告を選択するために，ユーザ情報空間から画像広告空間への変換行列を提案する．『Yahoo!ディスプレイアドネットワーク』の広告配信システムログを用いて提案した変換行列の有用性を検証した．

キーワード　オンライン広告, 機械学習, 転置インデックス

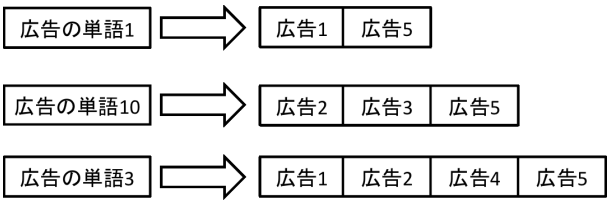
1. はじめに

オンライン広告はインターネット経済を支える大きな柱の一つである．そのため，この分野はビジネス的・学術的に大きな注目を浴びている．オンライン広告の種類としては，検索サービスにおける検索連動型広告やニュースなどのページにおけるコンテキスト広告，ポータルサイトにおけるディスプレイ広告が存在する．本稿ではコンテキスト広告のうち，クリック課金型の画像広告を扱う．クリック課金型広告とは，広告が配信されたページを閲覧しているユーザーが，広告をクリックして広告主の設定したページに移動した場合に，広告主があらかじめ入札していた金額に基づいて課金される仕組みの広告である．またオンライン広告には，テキストのみで構成されるテキスト広告と画像やテキストなどで構成される画像広告が存在する．

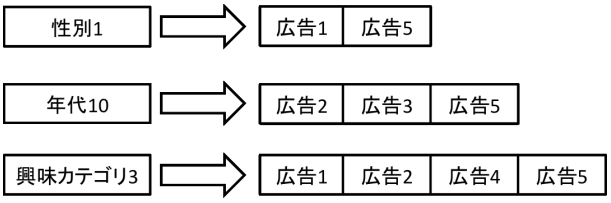
オンライン広告の課題の 1 つに，広告効果とユーザ体験の両方を満たすように，ウェブページの内容もしくはユーザ情報，あるいは両方に適合した広告を，どのように選択するかが挙げられる．本稿では，広告がウェブページ，ユーザ情報に適合しているかどうか，および広告効果を測る指標としてクリック率 (click through rate; CTR) を用い，これを最大化するような広告を選択する問題を考える．

テキスト広告の選択方法として，ユーザ情報やウェブページの内容およびテキスト広告を単語で表現し，重複する単語を用いて類似度を定義し，その類似度に基づいて広告を選択する方法が提案されている [5] ．この手法は，転置インデックスを用いた情報検索によって実現できる．しかしながら，画像広告においては画像情報しか存在しないため，画像広告に書かれている単語を検索するといった，テキスト広告と同じアプローチを単純に導入することができない．また，画像広告に対してその画像を表すような単語を付与して，画像広告を単語で表現してから情報検索に基づいたアプローチを用いることも考えられるが，ユーザ情報やウェブページに含まれる単語と広告に含まれる単語に隔たりがある場合には有効な手法ではない [14]．また，ユーザ情報やウェブページの内容に適合する画像広告を選択するために，あるユーザ情報やウェブページの内容と類似度が高くなるように，単語を画像広告に対して付与することは難易度の高いタスクである．

これらの問題を解決するために，本稿では広告リクエスト時に得られるユーザ情報を画像広告空間に変換する手法を提案する．従来手法を用いて構築した転置インデックスを図 1a および画像広告選択のための提案手法を用いて構築した転置インデックスの例を図 1b に示す．従来手法による転置インデックスは，広告に含まれる単語を用いて広告をインデックスするが，提案手法による転置インデックスでは，ユーザ情報のみを用いて広告をインデックスする．



(a) 従来手法による転置インデックス



(b) 提案手法による転置インデックス

図 1: 従来手法および提案手法による転置インデックス

これによりユーザ情報と画像広告の間に類似度を定めることが可能となり，テキスト広告と同じように情報検索に基づいたアプローチで画像広告を選択することができるようになる．ユーザ情報から画像広告空間への変換は行列の形で表現され，この行列は過去のクリックデータを用いて推定する．この変換

行列を転置インデックスとして用いることで，既存の広告検索システムに大きな変更を加える事なく，ユーザ情報を用いた画像広告の選択が実現できる．本稿では，クリック課金型広告の広告配信システムログを用いて，提案した広告選択手法の有効性を検証した．

本稿の構成は以下のとおりである．2 章では関連研究について述べる．続く 3 章では広告検索について述べる．4 章では実際の広告配信ログを用いて，提案手法に対する評価の結果を紹介する．以上を踏まえて， 5 章では本稿を結び，将来の展望について述べる．

2. 背　　景

2.1 広告配信システム概要

ユーザがウェブページに訪問した際に，ウェブページ上に掲載される広告は，ユーザ，広告およびウェブページの情報を用いて，その組み合わせにおける予測 CTR が最も高い広告を表示している．予測 CTR を算出する際には，ユーザ，広告およびウェブページから抽出された様々な素性を用いている．配信対象ウェブページやユーザが異なれば，予測 CTR も異なると仮定しているため，配信時にリアルタイムで CTR を予測する必要があるが，配信対象となる候補の広告の量は膨大である．そのため，すべての候補の広告に対してリアルタイムで CTR 予測を行うことは，計算コストの制約の観点から難しい．これを解決するために，予測 CTR の計算を行う前に，配信候補の広告からマッチングスコアを計算し，そのマッチングスコアの大きい方から top-k の広告を選択する方法 [3], [8] をとっている，残った top-k の広告に対して，学習しておいた CTR 予測モデルを用いて，ユーザと広告とページの組み合わせに対する予測 CTR を算出する．このように実際の広告配信時には，配信候補の広告を減らしてから，予測 CTR 計算を行う 2 段階の方法をとっている [1], [14]．このように，広告選択を行ってから，CTR 予測モデルを用いてスコアの計算を行うシステムの様子を図 2 に示す．

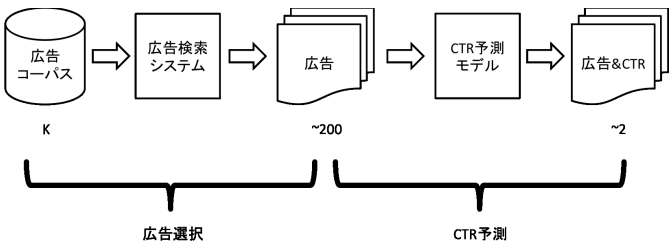


図 2: 広告システム概要

2.2 関 連 研 究

2.2.1 広告選択に関する研究

広告を選択する課題に関連する研究として，ウェブページに適合する広告を選択する手法を提案したものが挙げられる．これは一般的な広告よりもウェブページに適合した広告のほうが，より良いユーザ体験を提供するという仮定に基づいている．このような手法では，広告とウェブページをベクトル空間モデルで表現し，広告を選択する問題を情報検索に基づく手法で解決する．Chakrabarti ら [5] や Karimzadehgan ら [10] は，HTML タグやテキスト広告が表示されている位置の情報を用いて，ウェブページやテキスト広告に含まれる単語に関するパラメータを学習する手法を提案している．これらの手法では，転置インデックスを使うことで，ウェブページに適合するテキスト広告を効率的に選択できる．

2.2.2 画 像 検 索

画像広告の選択に関連した研究として，画像検索が挙げられる．Kulis ら [11] は，任意のカーネル特徴空間上における LSH (KLSH) を提案している．KLSH を類似画像検索に用いることで，高速かつ精度の高い検索結果となることを確認している．

また，機械学習を用いてハッシュ関数の生成と特徴ベクトルの圧縮を行う手法が提案されている [13], [15], [17]．これらの手法では低次元特徴空間にデータを圧縮した時のハミング距離が元の特徴空間の距離を保存するような方法で圧縮を行う．また，圧縮後のコードが短い場合には，ハッシュ値として用いることもできる．例えば Spectral Hashing [17] は，圧縮したバイナリコードを求めるための制約付き最適化問題を，グラフ分割の手法を用いることで解いている．

2.2.1 節および 2.2.2 節で挙げた手法は，クエリとなる情報と選択および検索する対象となる情報が同一のベクトル空間で表すことができる．しかしながら，本稿の対象であるユーザ情報と画像広告は，同一のベクトル空間モデルで表すことができないため，これらのような単語の共起を情報に基づく手法やハッシュ値に基づく手法をそのまま用いることができない

2.2.3 ウェブページ，クエリ，広告のベクトル空間の差を埋める研究

ウェブページやテキスト広告をベクトル空間モデルで表現した際に，それぞれのベクトルで使われる語が一致しない課題があげられる．ウェブページに適合する広告を選択する際に用いるスコアを，意味の階層構造を利用して求める手法 [4] が提案されている．また，トピックモデルを用いて，クリックされたウェブページと広告のペアから，ウェブページやテキスト広告に含まれる単語のトピックを EM アルゴリズムで推定する手法 [12] や，検索クエリ内の元の単語を “concept space” に写像する行列を推定する手法 [18]，ウェブページ内の単語やユーザ情報をテキスト広告のベクトル空間に写像する行列をクリックログから推定する手法 [14] が提案されている．更に，デモグラフィック情報や行動履歴などのユーザ情報からテキスト広告に含まれる単語のベクトル空間へ写像し，広告選択を行う手法 [9] が提案されている．また，クエリに適合するような広告のベクトル表現を，CTR 予測モデルから算出される予測値を再現するように回帰で学習している [1].

2.2.4 CTR 予測

広告選択の周辺の研究として，選択した広告の CTR を予測する課題が存在する．クリック課金型広告であるコンテキスト広告や検索連動型広告は，CTR を正確に予測することで，広告効果やユーザ体験およびウェブサイトの広告収益を最大化できる．広告の CTR 予測は，過去のクリックデータを利用した，

統計的モデリングで実現される．Azimi ら [2] は画像から抽出した画像素性を用いて，CTR を予測する手法を提案している．また，Cheng ら [7] は画像素性だけでなく音声素性を合わせたマルチメディア素性を用いて，新規の画像広告の CTR 予測を行うことを提案している．

3. 提 案 手 法

ユーザがウェブページに訪問した際に，ウェブページ上に掲載される広告は，ユーザ，広告およびウェブページの情報を用いて，予測 CTR が最も高い広告を表示している．しかしながら，配信対象となる候補の広告は膨大であるため，リアルタイムで CTR 予測を行うことは，計算コストの制約の観点から困難である．図 2 に示したように，検索システムを用いて配信候補の広告を削減することで，残った少数の候補広告に対して，あらかじめ推定済みの予測モデルに基づいて，リアルタイムに予測 CTR を算出する．

このように，予測モデルを使って CTR を計算する前に広告を絞り込む場合には，最終的に予測 CTR が高くなるであろう広告を CTR 予測の候補に残す必要がある．このため，top-k の広告を選択する際に，マッチングスコアに基づいて予測 CTR の高い広告を選択することができるように，転置インデックスを構築する必要がある．

テキスト広告では，ユーザ情報やウェブページの内容およびテキスト広告を単語で表現することができる．そのため，重複する単語による類似度に基づいた広告選択や CTR を最大化するようにウェブページやテキスト広告の単語のパラメータを学習した手法に基づいた広告選択 [5] を情報検索のアプローチを用いて行うことが可能である．しかしながら，画像広告においては単語情報が存在しないため，このような単語の重複から計算される類似度を用いたアプローチを導入することができない．

本稿では前段の広告選択において，予測 CTR が高くなるであろう画像広告の候補を効率的に検索するため，ユーザ情報から広告画像ベクトル空間への変換行列を提案する．提案手法は，あるユーザに対して CTR が高い画像広告を選択するために，変換行列をクリックログから推定する．推定した変換行列を広告検索システムの転置インデックスとして用いることで，テキスト広告と同じように情報検索のアプローチを利用して画像広告を選択することが可能となる．

3.1 マッチングスコア

ユーザがウェブページに訪れた際に，得られるユーザ情報を $\boldsymbol{q} = (q_1, \ldots, q_{D_q})$，すべての画像広告を $\boldsymbol{a} = (a_1, \ldots, a_{D_a})$，ユーザ情報から画像広告空間への変換行列を $\boldsymbol{W} = [w_{ij}]_{D_q \times D_a}$ と定義する．このとき，top-k の広告を選択する指標として用いるマッチングスコアは式 1 の形式で書ける．

$$score(\boldsymbol{q}, \boldsymbol{a}) = \boldsymbol{q}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{a} = \sum_{i=1}^{D_q} \sum_{j=1}^{D_a} w_{ij} q_i a_j \qquad (1)$$

ここで，D_q はユーザ情報ベクトル空間の次元数を表し，D_a は画像広告ベクトル空間の次元数を表すことにする．

画像広告ベクトル $\boldsymbol{a}$ は，変換行列 $\boldsymbol{W}$ と掛けあわせたベクトル $\boldsymbol{a}' = \boldsymbol{W}\boldsymbol{a}$ として保持しておき，マッチングスコア計算は式 2 を用いる．

$$score(\boldsymbol{q}, \boldsymbol{a}) = \boldsymbol{q}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{a} = \boldsymbol{q}^{\mathrm{T}} \boldsymbol{a}' \qquad (2)$$

変換行列と画像広告ベクトルの積である $\boldsymbol{a}'$ を転置インデックスとして用いることで，テキスト広告と同じように情報検索のアプローチによって，画像広告を選択することが可能になる．もし $\boldsymbol{a}$ が ID のようなカテゴリカルデータの場合は，ある 1 要素が 1 となる one-hot 表現され，$\boldsymbol{W}$ の対応する 1 列がインデックスになる．

図 1a および図 1b に示すように，従来手法による転置インデックスは，広告に含まれる単語を用いて，広告をポスティングリストに追加する．それに対して，提案手法による転置インデックスでは，デモグラフィック情報や行動履歴から推定したカテゴリ情報などのユーザ情報のみを用いて広告をポスティングリストに追加する点が大きく異る．

3.2 変換行列の学習

過去のクリックログを用いて，式 1 の w_{ij} を学習する．クエリ $\boldsymbol{q}$ と広告 $\boldsymbol{a}$ のマッチングスコアが，CTR を最大化するように w_{ij} を求める．広告配信システムログに存在する i と j の組み合わせの集合を P と定義し，q_i と a_j が同じログに存在するとき 1 になる素性を x_{ij} と置くと，マッチングスコアは式 3 の線形モデルとなる．

$$\begin{aligned} score(\boldsymbol{q}, \boldsymbol{a}) &= \sum_{(i,j) \in P} w_{ij} q_i a_j \\ &= \sum_{(i,j) \in P} w_{ij} x_{ij} \\ &= \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} \end{aligned} \qquad (3)$$

変換行列のパラメータの予測モデルは式 4 で表現されるロジスティック回帰モデルを用いた．

$$p(c|q, a) = \frac{1}{1 + \exp(-c(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}(q, a)))} \qquad (4)$$

なお， $c \in \{+1, -1\}$ は，画像広告をクリックしたかどうかを表す変数であり，$c = +1$ の場合はクリックしたことを表し，$c = -1$ の場合はクリックしなかったことを表す．また，$p(c = +1|q, a)$ は広告 a，ユーザ情報 q が与えられた時に，クリックする確率を表す．$\boldsymbol{x}(q, a)$ はその 2 つから抽出された素性ベクトルを表し，$\boldsymbol{w}$ はその素性に対応する重みベクトルを表現する．

本稿では，転置インデックスとして推定したパラメータを用いるため，過学習を防ぐための正則化項 $\Omega(\boldsymbol{w})$ を加え，式 5 の最適化問題を解き $\hat{\boldsymbol{w}}$ を求めた．

$$\begin{aligned} \hat{\boldsymbol{w}} = &\ \mathrm{argmin}_{\boldsymbol{w}} C\Omega(\boldsymbol{w}) \\ &+ \sum_{i=1}^{N} \log(1 + \exp(-c_i(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i(q, a)))) \end{aligned} \qquad (5)$$

なお，$C > 0$ は正則化パラメータである．C を変化させて学習データで学習を行い，バリデーションデータでの評価値が高い

$\hat{\boldsymbol{w}}$ を用いてテストデータの評価を行った．また，スパースな転置インデックスを作成するために，L_1 正則化 $\Omega(\boldsymbol{w}) = |\boldsymbol{w}|$ を用いた．

本稿では，転置インデックスを推定する際に，Agarwal ら [1] のように広告ごとにモデルを推定するのではなく，広告全体に対するモデルを推定し，そのパラメータを転置インデックスとして用いた．これによって，広告間の CTR の大小を考慮することが可能となっている．

評価では，素性ベクトル $\boldsymbol{x}(q, a)$ の次元数を hashing trick [16] を用いて制限した．ハッシュ値として 24 ビット整数を用い，素数ベクトルの次元をおよそ 1,600 万とした．

4. 実　　験

本稿では，クリックログを用いてユーザに対して，CTR が最も高くなるような広告を検索するための変換行列を推定した．しかしながら，実際の広告システムでは検索を行いマッチングスコアが最も高い広告をユーザに表示するのではなく，後段の処理の CTR 予測モデルで候補となる広告を評価した上で，ユーザに適合する広告を選択している．提案手法による変換行列を広告検索の転置インデックスとして用いることで，広告システムの後段の処理の予測 CTR が高くなる広告を選択できるか否かを，実際に配信された広告のシステムログを用いたシミュレーションによって評価した．

4.1 データセット

評価には『Yahoo!ディスプレイアドネットワーク (YDN)』(注1) のある 10 のウェブサイトで，実際に配信された広告のシステムログを 6 週間分用いた．このデータの前半 4 週間を学習データ，続く 1 週間をバリデーションデータ，最後の 1 週間をテストデータとして扱った．

データの各サンプルは配信された広告 1 つに対応しており，クリックされたか否かがラベル付けされている．本稿では，クリックされたサンプルを正例，クリックされなかったサンプルを負例とした．それぞれのデータは，同じ規則でフィルタリングとサンプリングを行った．Chapelle らの方法 [6] に倣い，負例からサブサンプリングを行った．最終的に得られた学習データ，バリデーションデータ，テストデータのサンプル数を表 1 にまとめた．

表 1: データセットのサンプル数

	学習データ	バリデーションデータ	テストデータ
正例数	4,248,743	1,033,727	1,193,258
負例数	8,700,249	2,223,281	2,435,395
合計	12,948,992	3,257,008	3,628,653

4.2 提案手法の変換行列を用いた広告選択の実験

本稿では，配信システムログから CTR が高い広告を検索できるように転置インデックスを構築した．実際の配信システムでは，ユーザに適合する広告を選ぶ際に，2 段階の処理で広告を選択している．ユーザに表示される広告は，後段の CTR 予測において算出される予測 CTR の値が最大となる広告が選択される．このため前段の広告検索の処理では，予測 CTR が最大となるような広告を選択する必要がある．提案手法によって CTR が高い広告を検索できるような変換行列を用いて広告を選択する際に，後段の処理において予測 CTR が大きくなる広告を選ぶことができるか否かを，実際に配信された広告のシステムログを用いてシミュレーションを行った．本節では，実験に用いた素性やベースラインおよび評価指標について述べる．

4.2.1 素　　性

用いた素性は種類に応じてグループ分けを行った．それぞれの素性グループの詳細は表 2 にまとめた．広告主に紐づく素性は，広告の階層構造を利用した．広告の階層構造は，広告主，キャンペーン，広告グループ，広告の順に粒度が細かくなり，それぞれの ID を素性として用いた．ユーザに紐づく素性としてはアクセスしたデバイス種別に加えて，登録しているユーザであれば，性別，年代，地域情報，ウェブ上の行動履歴から推定した興味カテゴリなどを用いた．性別は男性，女性，不明の 3 クラスに分割し，同様に年代は 13 グループに分割した．登録していないユーザの場合には，推定した性別，年代，地域情報を用いた．その他の素性としては，過去に配信した際に得られた実績 CTR，ウェブページと広告の類似度と，2 種類のユーザーと広告の関連度を用いた．また，データに含まれる各 ID 素性のユニーク数は表 3 にまとめた．

表 2: 素性グループ

素性グループ	素性詳細
広告主	広告 ID，広告グループ ID，キャンペーン ID，広告主 ID
ユーザ	性別，年代，地域，デバイス種別，興味カテゴリなど
その他	実績 CTR，ウェブページと広告の類似度，ユーザと広告の関連度 (2 種類)

表 3: ID 素性のユニーク数

	学習データ	バリデーションデータ	テストデータ
広告 ID	730,500	172,335	163,231
広告グループ ID	139,236	225,822	219,474
キャンペーン ID	35,868	26,247	28,062
広告主 ID	9,783	8,464	7,993

4.2.2 比 較 手 法

シミュレーションのベースラインの手法として，ユーザ情報を考慮せずに広告主素性グループのみを用いて，CTR が高い広告を検索できるように構築した most popular モデルについて述べる．

表 2 に示す広告主素性グループの広告 ID と 広告主 ID を用いて，CTR を最大化するような広告選択のモデルをベースラインとして提案する．これは，要素の 1 つのみが 1 となるような one-hot なベクトルを $\boldsymbol{q}$ としたときの式 1 と等価であり，

(注1)：http://promotionalads.yahoo.co.jp/service/ydn/index.html

広告 ID と 広告主 ID の線形モデルになる．提案手法と同様に式 5 を解くことで得られたパラメータ $\hat{\boldsymbol{w}}$ を検索インデックスとして用いることで，CTR を最大化する広告選択を実現する．most popular モデルは，ユーザ情報を考慮していないため，すべてのユーザに対して，訓練データ中で CTR の高い広告を広告候補して選択するモデルになっている．

提案手法について表 4 にまとめる．

表 4: 手法一覧

手法	素性
ad	広告 ID とユーザ素性グループの組み合わせ
adg	広告グループ ID とユーザ素性グループの組み合わせ
camp	キャンペーン ID とユーザ素性グループの組み合わせ
adv	広告主 ID とユーザ素性グループの組み合わせ
baseline	広告 ID と広告主 ID

4.2.3 評　　価

提案手法は CTR が高い広告を選択するように変換行列を推定している．提案手法で構築した変換行列を用いてユーザに対して CTR が高くなる広告を top-k 検索する際に，検索システムの後段の CTR 予測システムが算出する予測 CTR が最大となる広告を見つけることができるか否かを確認する．検索システムの後段の処理である CTR 予測は，実際の画像広告配信システムで用いられている画像広告 CTR 予測モデルを用いて値を算出した．CTR 予測モデルは，表 2 に示した素性を用いた．

予測 CTR が高い広告を引き当てられたか否かの評価として，検索対象のすべての広告の中で最大となる予測 CTR と，提案手法の変換行列に基づいて計算したマッチングスコアを用いて，top-k の広告候補を選択した際に最大となる予測 CTR の比を評価指標とした．

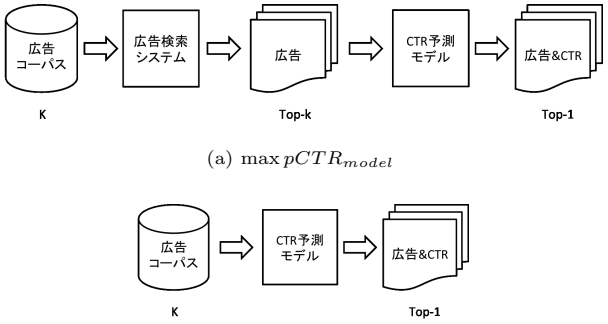
$$pCTR\text{ 比 }(model) = \frac{\max pCTR_{model}}{\max pCTR_{corpus}} \times 100\%$$

$\max pCTR_{model}$ と $\max pCTR_{corpus}$ の関係を図 3a，3b に示す．$\max pCTR_{model}$ は提案手法に基づいて，全ての広告候補から top-k 検索した結果に対して，予測 CTR を計算し最大のものを選んだ．$\max pCTR_{corpus}$ は全ての広告候補に対して，予測 CTR を計算し最大のものを選んだ．pCTR 比は 100 に近いほど，広告検索で検索対象の広告集合の中から，ユーザに適合する広告を見つけることができたことを示す．

k の値は実配信システムの top-k 検索の値を参考に最大 200 件の広告が得られるように k を [1, 10, 50, 100, 200] と選んだ．また，画像広告ベクトル空間に変換するユーザ情報は，テストデータから 100 人分のユーザ情報をサンプリングした．提案手法の変換行列を用いた際の top-k 検索結果に対応する予測 CTR 比を図 4 に示す．

10 のすべてのウェブサイトで，k が小さい場合には，adv を用いた場合の pCTR 比が大きいが，k を大きくするにつれて，ad を用いた場合の pCTR 比が，adv の pCTR 比よりも改善するという傾向が見られた．

これは adv のように，広告の階層構造の上位階層に位置する広告主 ID で学習すると，広告主 ID が同じであるような新しい広告に対してもマッチングスコアを計算することができるため，広告選択を行うことができることを示している．また，k を大きくしても， adv では，精緻なマッチングスコアを計算することができないため，pCTR 比が改善しないと考えられる．



(a) $\max pCTR_{model}$

(b) $\max pCTR_{corpus}$

図 3: $\max pCTR_{model}$ および $\max pCTR_{corpus}$

しかしながら， 階層構造の最下位に位置する広告主 ID で学習した ad では，学習データに含まれない新しい広告のマッチングスコアを正しく評価できないこともあり，k が小さい場合には探索範囲が狭く予測 CTR が高い広告を選択することができないが，探索範囲が十分広い場合には，adv よりも予測 CTR が高い広告を選択することが可能であることが確認できた．

また，ベースライン手法は，ユーザ情報を考慮せずに広告主素性グループでクリックを学習した変換行列になっており，訓練データ中の CTR が高い広告のマッチングスコアが高くなるモデルである．k が小さい場合でも，pCTR 比が大きい値になっているが，k を増やして探索範囲を増やしても，pCTR 比に変化が見られない．このことから，ユーザ情報を考慮しない変換行列による広告選択では，提案手法と比較して，予測 CTR の高い広告を選択することができないといえる．

5. おわりに

本稿ではオンライン広告のうち，クリック課金型の画像広告に注目し，特にユーザ情報に適合した広告を選択するために，ユーザ情報から画像広告空間に変換する手法を提案した．ユーザ情報から画像広告空間への変換は行列の形で表現され，過去のクリックデータを用いて推定される．この変換行列を転置インデックスとして用いることで，既存の広告検索システムに大きな変更を加えることなく，情報検索に基づくアプローチで CTR が最大となるであろう画像広告を選択することができる．

提案手法を『Yahoo!ディスプレイアドネットワーク』の広告配信システムログを用いて検証を行った結果予測 CTR が高い広告を選択できることを確認した．

提案手法は CTR を最大化するような変換行列を求め，実験的に予測 CTR が大きい広告を選択することができることを確認したが，既存のシステムに導入し，実際の広告選択に用いた場合の検証を行うことが，今後の課題として挙げられる．また，複数の広告階層構造とユーザ情報の組み合わせ素性を考慮する

ことができるようにするために，ユーザ情報から画像広告空間への変換行列を推定する際に，分散学習を採用することが挙げられる．また，予測 CTR の計算式と広告検索のマッチングスコアの計算式を近似するような手法 [1] も提案されており，これらについても考慮する必要がある．

文　献

[1] Deepak Agarwal and Maxim Gurevich. Fast top-k retrieval for model based recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pp. 483–492, 2012.

[2] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. The impact of visual appearance on user response in online display advertising. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pp. 457–458, 2012.

[3] Andrei Z. Broder, David Carmel, Michael Herscovici, Aya Soffer, and Jason Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pp. 426–434, 2003.

[4] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pp. 559–566, 2007.

[5] Deepayan Chakrabarti, Deepak Agarwal, and Vanja Josifovski. Contextual advertising by combining relevance with click feedback. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pp. 417–426, 2008.

[6] Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol.*, 2013.

[7] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 777–785, 2012.

[8] Marcus Fontoura, Vanja Josifovski, Jinhui Liu, Srihari Venkatesan, Xiangfei Zhu, and Jason Y. Zien. Evaluation strategies for top-k queries over memory-resident inverted indexes. *PVLDB*, Vol. 4, No. 12, pp. 1213–1224, 2011.

[9] Amruta Joshi, Abraham Bagherjeiran, and Adwait Ratnaparkhi. User demographic and behavioral targeting for content match advertising. In *Proceedings of the fifth international workshop on Data mining and audience intelligence for advertising*, ADKDD '11, 2011.

[10] Maryam Karimzadehgan, Wei Li, Ruofei Zhang, and Jianchang Mao. A stochastic learning-to-rank algorithm and its application to contextual advertising. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 377–386, 2011.

[11] Brian Kulis and Kristen Grauman. Kernelized locality-sensitive hashing for scalable image search. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pp. 2130–2137, 2009.

[12] Adwait Ratnaparkhi. A hidden class page-ad probability model for contextual advertising. In *Workshop on Targeting and Ranking for Online Advertising at the 17th International World Wide Web Conference*, 2008.

[13] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *Int. J. Approx. Reasoning*, Vol. 50, No. 7, pp. 969–978, July 2009.

[14] Yukihiro Tagami, Toru Hotta, Yusuke Tanaka, Shingo Ono, Koji Tsukamoto, and Akira Tajima. Filling context-ad vocabulary gaps with click logs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1955–1964, 2014.

[15] Antonio Torralba, Robert Fergus, and Yair Weiss. Small codes and large image databases for recognition. In *CVPR*. IEEE Computer Society, 2008.

[16] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1113–1120, 2009.

[17] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 1753–1760, 2008.

[18] Wen-tau Yih and Ning Jiang. Similarity models for ad relevance measures. In *MLOAD - NIPS 2010 Workshop on online advertising*, 2010.

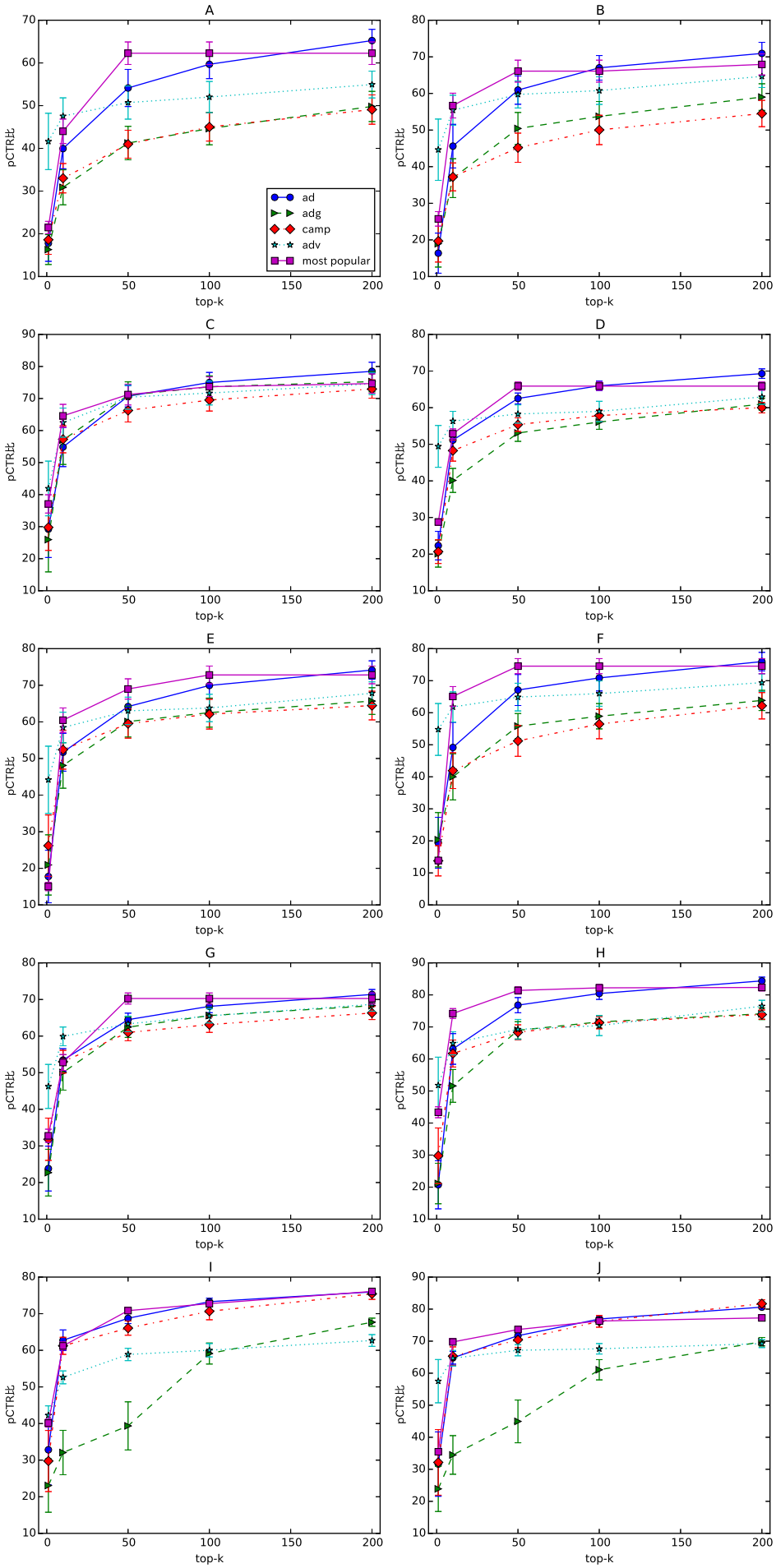


図 4: top-k 検索に対する予測 CTR の比