# クリックフィードバックによる広告検索精度向上

堀田　徹† 田頭　幸浩† 小野　真吾†
山本　浩司† 塚本　浩司† 田島　玲†

† ヤフー株式会社　〒 107–6211 東京都港区赤坂 9–7–1
E-mail: †{thotta,yutagami,shiono,koyamamo,kotsukam,atajima}@yahoo-corp.jp

**あらまし**　オンライン広告におけるコンテキスト広告では，閲覧ページの内容やユーザーの行動履歴といったコンテキスト情報から，それぞれの広告のクリック率 (click-through rate; CTR) を予測している．この CTR の予測値をもとに期待収益を計算し，それを最大化するようにランキングし表示させている．CTR は過去の広告クリックログを用いて機械学習モデルを作成し，モデルのスコアによって予測するのが一般的である．しかし，機械学習モデルは一般に精度が高い一方で，配信時に大規模なアイテム集合に対して CTR を予測すると計算コストが高い．多くの広告配信システムにおいては，ウェブページを素早く表示するために数十ミリ秒で上位の結果を検索する必要がある．精度とレイテンシの両方を追求するため，コサイン類似度などの単純なスコアで少数の広告を情報検索のシステムを用いて取得し，より精度の高い機械学習モデルでリランキングする手法が用いられる．本稿では，その 1 段階目の情報検索のシステムにおいて，標準的な情報検索システムで扱うことができる関数形のモデルを機械学習によって構築し，このモデルのスコアを従来のコサイン類似度の代わりに用いることにより，効率的な広告検索を行うアプローチをとり，評価を行った結果を報告する．

**キーワード**　オンライン広告，CTR 予測，機械学習，情報検索

## 1. 導　　入

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る．そのため，この分野はビジネス的に，また学術的にも大きな注目を浴びている [3], [14]．オンライン広告の例としては，検索サイトにおける検索連動型広告，ポータルサイトにおけるディスプレイ広告，ニュースやブログ記事のページにおけるコンテキスト広告が挙げられる．

本稿ではオンライン広告のうち，広告自体がテキストからなる，クリック課金型のコンテキスト広告を扱う．クリック課金型の広告では，ユーザーがその広告をクリックした場合にのみ，広告主に課金が行われる．コンテキスト広告ではユーザーが閲覧中のページや行動履歴などから潜在的な興味を抽出し，その興味に連動した，クリックされやすいであろう広告を配信する．これにより，広告主は広告に興味を持っているであろうユーザーに対する効果的な広告配信を行うことができる．

広告主はある広告の掲載を依頼する時に，課金されてもよいと考える額を入札額として提示し，その入札額をもとに実際の課金額が決定される．

コンテキスト広告には，ユーザーや広告主の他に広告掲載ページを提供するパブリッシャーと呼ばれる事業者が存在する．パブリッシャーは広告掲載で収益を得るため，限られた広告掲載の機会の中で可能な限り収益を最大化することを目的とする．そのためには，それぞれの広告を配信した際における収益の期待値を知る必要があるが，これは広告主が設定した入札額と広告のクリック率 (click-through rate; CTR) で決まる．これら二つのうち広告主が設定した入札額は既知であるが，同じ広告であっても閲覧中のページやユーザーにより CTR は異なるため，配信時に予測が必要である．たとえば，株価や為替情報のページを閲覧中であればそれに関連した証券会社の広告がクリックされやすく，自動車に興味を持つユーザーであれば自動車の広告がクリックされやすいであろうと考えられる．

現在の広告配信システムでは，CTR は過去の配信ログを用いて機械学習により作成された予測モデルのスコアによって求めている．予測精度の観点からはリクエストごとにすべての広告の CTR を予測することが望ましい．しかし，実際の配信システムでは 1 回のリクエストに対して数十ミリ秒で応答することが求められており，すべての広告に対して機械学習モデルを適用するのは計算コストが高く，現実的ではない．そのため，情報検索のシステムを用いてコサイン類似度などの単純なスコアで少数の広告を取得し，それらをより精度の高い機械学習モデルでリランキングする手法が一般的である．概念図を図 1 に示す．しかしながら，この手法では検索システムとランキングロジックとの間でスコアが異なるため，最適な広告配信ができているとは必ずしも言えない．

本稿ではこの問題へのアプローチとして，図 1 の太枠で囲った広告検索システムにおいて，機械学習によって構築したモデルを使う効率的な広告検索手法 [5] を適用し，評価を行った結果を報告する．この手法では，標準的な情報検索システムで扱うことができる関数形のモデルを構築することにより，従来の検索システムの効率性を維持したまま精度の高い広告検索を実現することができる．

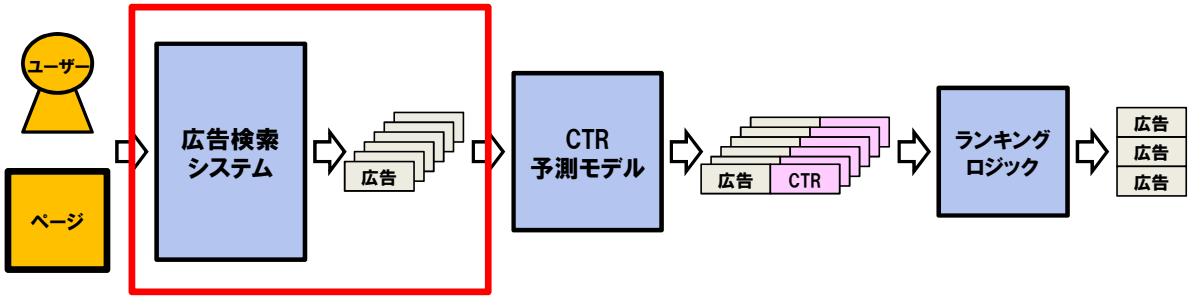


図 1　コンテキスト広告の概念図．ユーザーとページから素性を抽出してクエリとし，検索システムを用いて候補となる広告を取得する．その後候補の広告に対して CTR を予測し，ランキングロジックにて最終的に配信する広告と順序を決定する．本稿では太枠で囲んだ広告検索システムの部分において，機械学習を用いた最適化を行う．

本稿の構成は下記の通りである．2 章では関連する先行研究について紹介する．続く 3 章で転置インデックスを用いた検索システムについて説明する．4 章では検索システムで用いられる従来のコサイン類似度などのスコアを，機械学習モデルのスコア関数に置き換えるアプローチについて解説する．5 章では実際の広告配信ログを用いたオフラインシミュレーションの結果を示し，6 章で全体のまとめをする．

## 2. 関 連 研 究

この章では先行研究として，広告検索と CTR 予測に関する研究を挙げる．

Wang らの研究 [13] では，ユーザービリティやクリック率を向上させるために，ユーザーの興味に関連した広告を表示することが有効であると確認されている．関連度の高い広告を検索する技術としては，ウェブ検索で用いられる情報検索の手法が効果的とされている．

Ribeiro-Neto ら [11] はコンテキスト広告の広告引き当てにおいて，ページと広告から抽出した単語をもとに引き当てを行ういくつかの手法を評価した．この研究では，広告とページはそれぞれ多次元ベクトルとして表現され，それら 2 つのベクトルのコサイン類似度を利用して広告引き当てを行う手法について検証している．この中で，広告のタイトルや本文，課金フレーズといったさまざまな広告セクションを利用し，精度の高い手法を探求している．

ページと広告が同じベクトル空間上にマッピングされる手法では，ページと広告の持つ単語群の違いから少なからずミスマッチが生じる．つまり，それぞれの次元が各単語を表現するベクトル空間モデルでは，同意語などを適切に扱うことができない．この問題に対応するために，ページから抽出される単語群を，そのページに似ているページの単語群を利用して拡張することにより広告検索精度を向上させる手法が提案されている [11]．また，ラベルづけされた教師データを用いて分類器を作成し，その分類器を用いてページと広告にクラス情報を付与し，広告検索に利用する手法も提案されている [2]．

Chakrabarti ら [5] は，従来のコサイン類似度の代わりに，広告の配信ログからロジスティック回帰モデルによって学習したページや広告の単語の重みパラメータを利用して，そのコンテキストにおいて CTR が高い広告を取得する手法を提案した．本稿では，この手法をもとに評価を行った結果を報告する．

ここまで広告検索に関する研究を挙げてきたが，もう一つの研究の方向性として CTR 予測がある．CTR 予測では，過去の広告クリックログを用いた機械学習モデルが利用される．一般的には線形モデルが用いられ，その中でも CTR が確率であることから確率値を出力とするロジスティック回帰モデルが用いられることが多い [6], [7], [12]．

## 3. 検索システム

この章ではキーワードをクエリとしてドキュメントを取得する一般的な検索システムについて説明を行う．コンテキスト広告における広告検索では，大まかに閲覧ページとユーザーの 2 種類の情報を用いて広告を取得するため，広告をドキュメントとし，閲覧ページとユーザーから抽出した情報をキーワードとして用いて検索を行う．

まずベクトル空間モデルについて述べ，次に転置インデックスを用いた検索システムについて説明する．最後に，クエリに依存しないドキュメントごとの静的クオリティスコアについて述べる．

### 3.1 ベクトル空間モデル

ベクトル空間モデルにおいては，クエリとドキュメントの両方が同じ多次元空間上のベクトルとして表現される．$m$ を単語の種類数とすると，クエリ $q$ を表すクエリベクトルは $\boldsymbol{q} = (q_1, q_2, \ldots, q_m)$ と表現され，同様にドキュメント $d$ を表すドキュメントベクトルは $\boldsymbol{d} = (d_1, d_2, \ldots, d_m)$ と表現される．$q_i$ はクエリ $q$ における単語 $i$ の重みを表す．$d_i$ も同様である．

クエリ $q$ に対するドキュメント $d$ のスコアは，二つのベクトルの内積 $\boldsymbol{q}^T\boldsymbol{d} = \sum_i q_i d_i$ で定義される．このクエリとドキュメントの内積スコアは，3.2 節で紹介する転置インデックスを用いた top-$k$ retrieval システムにおいて，高速に計算することができる．

### 3.2 転置インデックスを用いた検索システム

検索システムにおける標準的なアプローチでは，まず最初に検索アイテムであるドキュメントコーパスにおいて転置イ

ンデックスを構築する．これは，ポスティングリストと呼ばれる，コーパス中での各単語の出現位置を表すリストの集合 $(L_1, L_2, \ldots, L_m)$ を作成することと同義である．リスト $L_i$ は，単語 $i$ の重みが 0 でない広告を列挙することで作成される．ほとんどの広告において，ある単語 $i$ の重みは 0 なので，そのまばらさを利用してポスティングリストは，ドキュメントの ID とそのドキュメント内の単語の重みからなる $< doc\_id, d_i >$ の二つ組を要素とする集合で記述可能である．今，クエリが与えられたとすると，そのクエリに含まれる単語に対応するリストのみを考慮すればよく，これにより検索空間を削減し高速なシステムを構築することができる．

top-$k$ retrieval では，転置インデックスを用いることで，与えられたクエリに対してスコア上位 $k$ 件のドキュメントを高速に取得することができる．これはシステム内でのスコアの評価途中で，既に上位 $k$ 件に入ることがないと分かっているドキュメントに対する評価を省くことで効率化を行うもので，さまざまな手法が提案されている [4], [8].

### 3.3 ドキュメントごとの静的クオリティスコア

各ドキュメントには，クエリに依存しない静的なクオリティスコアがあると考えられ，例としてウェブ検索における PageRank [10] が挙げられる．一般的な情報検索のシステムでは，このような静的なクオリティスコアを，クエリに応じて変化するドキュメントの動的なスコアに加算したものを最終的なスコアとして扱うことができる [9].

## 4. 機械学習モデルによる広告検索

この章では，標準的な情報検索システムで扱うことができる関数形のモデルを機械学習によって構築し，このモデルのスコアを従来のコサイン類似度の代わりに用いて，効率的な広告検索を行うアプローチについて説明する．なお，以下では CTR の予測値が高い広告を取得することを目的とする．期待収益が高い広告を取得することを目的とする場合もほぼ同様である．

### 4.1 機械学習による CTR 予測モデル

2 章で述べたように，CTR 予測モデルとしては線形モデルが用いられることが多い．ここでは以下のような線形モデルで CTR の予測値が表現されるとする．

$$CTR = \sum_i w_i f_i(p, u, a) \qquad (1)$$

ここで，$f_i(p, u, a)$ はページ $p$，ユーザー $u$，広告 $a$ の 3 つから抽出された $i$ 番目の素性を，$w_i$ はその素性に対応する重みを表現している．重みベクトル $\boldsymbol{w} = (w_1, w_2, \ldots)$ は，データから機械学習で推定される．

### 4.2 機械学習による広告検索モデル

式 (1) で表現されるモデルの予測値が高い広告を検索時に取得することがここでの目的である．広告検索時に用いるモデルを，以下の関数形で表現する．

$$score_{train} = \sum_i w_{d,i} q_i a_i + \sum_j w_{s,j} a_j + \sum_k w_{b,k} q_k \qquad (2)$$

なお，$\boldsymbol{q} = (q_1, q_2, \ldots)$ はクエリベクトルを，$\boldsymbol{a} = (a_1, a_2, \ldots)$ は広告ベクトルを表す．それぞれ，ページとユーザーから抽出した素性ベクトルと，広告のみの素性ベクトルを表すものとする．

式 (2) の第一項はクエリと広告の関連度を表す項であり，各素性に応じて $\boldsymbol{w}_d = (w_{d,1}, w_{d,2}, \ldots)$ で重みづけをした形となっている．$q'_i = w_{d,i} q_i$ とした値をクエリとして用いるか，もしくは $a'_i = w_{s,i} a_i$ とした値をドキュメントベクトルとして転置インデックスを構築することで，従来の検索システムで検索を行うことが可能である．

第二項は閲覧ページやユーザーなどのコンテキストに依存しない，その広告自身のクリックされやすさを表す静的なクオリティスコアである．これは広告ベクトル $\boldsymbol{a} = (a_1, a_2, \ldots)$ と対応する重みベクトル $\boldsymbol{w}_s = (w_{s,1}, w_{s,2}, \ldots)$ によって計算される．広告検索時には，ドキュメントの静的クオリティスコアとしてこの値を用いる．

第三項はページやユーザーによって，広告の種類に関わらず CTR が変動することを表す項である．これはクエリベクトル $\boldsymbol{q} = (q_1, q_2, \ldots)$ と対応する重みベクトル $\boldsymbol{w}_b = (w_{b,1}, w_{b,2}, \ldots)$ によって計算される．広告検索時には，既にページとユーザーは固定されているため，この項は広告間の CTR 予測値の大小関係には影響しない．しかしながら，学習データはページとユーザーによって CTR に偏りがあると考えられ，単純に式 (2) の第一項と第二項のみからなるモデルを学習すると，二つの項でその偏りを表現するため，精度が悪化することが予想される．この項は学習時には偏りを表現するために用いるが，広告間の CTR 予測値の大小関係には影響しないため，広告検索時のスコア計算には用いない．つまり，実際の広告検索時には下記のようなスコアが計算される．

$$score_{ad} = \sum_i w_{d,i} q_i a_i + \sum_j w_{s,j} a_j \qquad (3)$$

式 (2) の重みパラメータベクトル $\boldsymbol{w}_d, \boldsymbol{w}_s, \boldsymbol{w}_b$ は機械学習によってデータから推定される．推定には大まかに二つの手法が考えられる．一つ目は，広告配信ログを用いてクリックされたか否かを目的変数として直接学習を行う手法である [5]．二つ目は，あるページ，ユーザー，広告に対して既存の機械学習モデルで CTR を予測し，その予測値を広告検索時のモデルの目的変数として学習を行う手法である [1]．本稿では一つ目のアプローチを取り，この手法を用いた広告検索システムの概念図を，図 2 に示す．

図 2 の広告検索システムでは，事前に広告から素性ベクトルを作成し，転置インデックスと静的クオリティスコアを作成しておく．ユーザーとページの組み合わせが検索リクエストとして与えられれば，それらから同様の素性ベクトルを作成し，機械学習によって推定された重みパラメータ $\boldsymbol{w}_d$ を用いてクエリの素性ベクトルを変換する．この変換されたクエリの素性ベクトルと広告の素性ベクトルの内積スコアと広告の静的クオリティスコアを合計すると，式 (3) が計算できる．

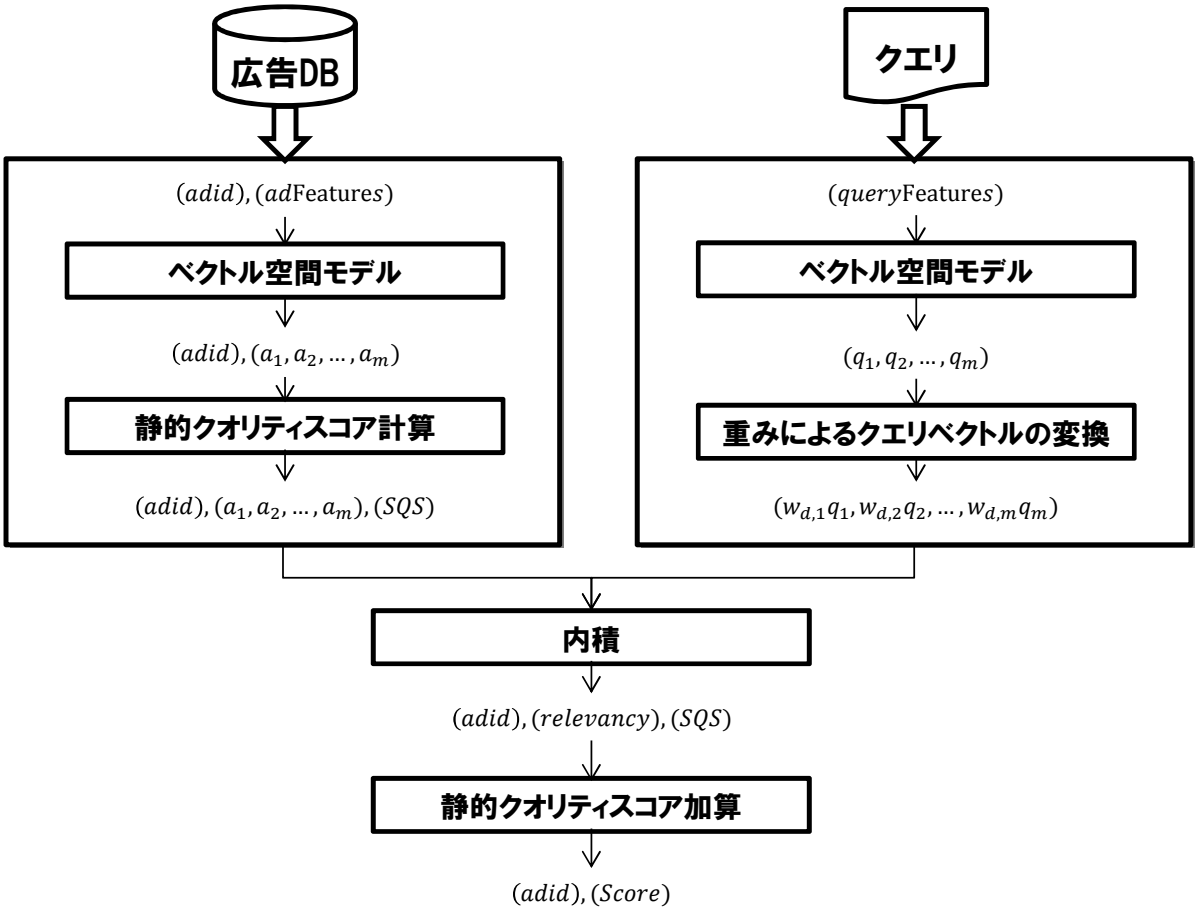


図 2　機械学習モデルを用いた広告検索システムの概念図．図中の SQS は静的クオリティスコア (Static Quality Score) を意味する．

## 5. 実　　験

この章では，4.2 節で述べた機械学習モデルによる広告検索と，従来のコサイン類似度を用いた検索の比較を行う．

### 5.1　広告検索手法

以下の三つの手法を比較する．

- コサイン類似度モデル
- 機械学習モデル（静的クオリティスコアなし）
- 機械学習モデル

コサイン類似度モデルは，従来のコサイン類似度での広告検索方式である．この手法をベースラインとし，その他の手法の評価を行う．2 つ目のモデルは，式 (3) の第一項だけでスコアリングする手法で，クエリと広告の関連度だけで検索する手法である．3 つ目の機械学習モデルは，式 (3) でスコアリングする広告検索手法である．機械学習モデルのうち、静的クオリティスコアなしのモデルとありのモデルを比較することで，広告自身の魅力を表す静的クオリティスコアが，CTR にどれだけ影響を与えるかを見る．

### 5.2　実 験 設 定

実験では，実際の広告配信システムの中にある広告集合の中から，広告主の予算状況などによるフィルタリングやサンプリングを施すことで約 30 万件を選択し，広告配信のシミュレーションを行った．実際の配信システムでは広告主の予算は動的に変化し，予算がなくなればその広告主の広告は配信されないが，本実験においては，予算の変化については考慮せず評価を行った．

また，クエリとしては，ウェブページの語句情報のみを用いる．それぞれの広告検索手法の比較をするため，いくつかのウェブページを固定して評価を行う．今回は，11 のウェブサイトを実験対象とし，ある 1 日の広告配信回数が多かった上位 100 ページをそれぞれのウェブサイトごとに選択して用いた．クエリが与えられ広告検索を行った後，取得した全ての広告に対して既存のモデルで CTR を予測し，上位 12 本の広告を最終的な配信対象とみなした．広告検索時には最大 1,000 本の広告をリクエストした．評価基準としては，CTR の高い広告を検索システムから取得することを目的としているため，配信対象となった 12 本の CTR 予測値を平均して用いた．

式 (2) の重みパラメータを学習するためのデータとしては，広告配信ログを約 3 か月分サンプリングし，用いた．

### 5.3　実 験 結 果

実験の結果を図 3 に示す．アルファベットの A から K は各ウェブサイトを表しており，グラフの縦軸はコサイン類似度モデルを用いた場合と比較した，CTR 予測値の変化率である．

図 3 の黒色のグラフは静的クオリティスコアありの機械学

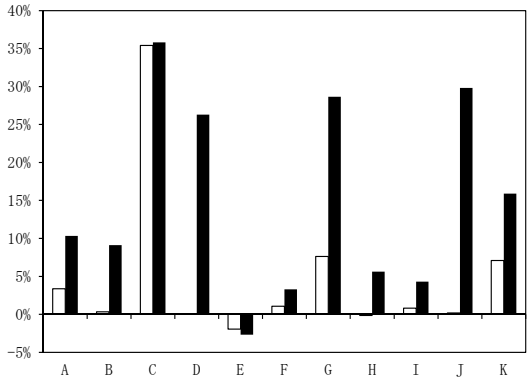


図 3　ベースラインと比較した CTR 予測値の変化率．白色が静的クオリティスコアなしの機械学習モデル，黒色が静的クオリティスコアありの機械学習モデルの変化率を表している．

習モデルにおける合計 CTR の変化率であり，E 以外のすべてのドメインで CTR が向上しているのがわかる．11 サイト中 6 サイトで 10%以上の向上をしており，多くのウェブサイトで大幅に向上しているのがわかる．全ウェブサイトでみると，約 15%の向上であった．

一方，静的クオリティスコアなしの機械学習モデル（白色）では，11 サイト中 9 サイトがコサイン類似度モデルを上回ったが，その向上率は小さく全ウェブサイトでみると 5%の向上であった．このことから，静的クオリティスコアで表現される広告自身の持つ魅力が，CTR に大きな影響を与えるといえる．

## 6. まとめと今後の課題

本稿では，一般的なコンテキスト広告の配信システムにおける，機械学習モデルを用いた広告検索を行う手法について述べた．また実験では，この手法を用いることで従来のコサイン類似度で広告検索を行った場合と比較して，実際に CTR の予測値が高い広告を取得できることを確認した．

今回はページと広告の語句情報のみを用いた実験を行ったが，ユーザー情報をクエリとした検索の実験も考えられる．これは今後の課題である．

## 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, pages 483–492, New York, NY, USA, 2012. ACM.

[2] 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, pages 559–566, New York, NY, USA, 2007. ACM.

[3] Andrei Broder and Vanja Josifovski. Introduction to computational advertising. `http://www.stanford.edu/class/msande239/`. Accessed: 14/12/2012.

[4] 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, pages 426–434, New York, NY, USA, 2003. ACM.

[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, pages 417–426, New York, NY, USA, 2008. ACM.

[6] Haibin Cheng and Erick Cantú-Paz. Personalized click prediction in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 351–360, New York, NY, USA, 2010. ACM.

[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, pages 777–785, New York, NY, USA, 2012. ACM.

[8] Marcus Fontoura, Vanja Josifovski, Jinhui Liu, Srihari Venkatesan, Xiangfei Zhu, and Jason Zien. Evaluation strategies for top-k queries over memory-resident inverted indexes. In *Proceedings of the 37th International Conference on Very Large Data Bases*, volume 4, pages 1213–1224, 2011.

[9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[10] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[11] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 496–503, New York, NY, USA, 2005. ACM.

[12] Rómer Rosales, Haibin Cheng, and Eren Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 293–302, New York, NY, USA, 2012. ACM.

[13] Chingning Wang, Ping Zhang, Risook Choi, and Michael D'Eredita. Understanding consumers attitude toward advertising. In *In: Eighth Americas Conference on Information Systems*, pages 1143–1148, 2002.

[14] Shuai Yuan, Ahmad Zainal Abidin, Marc Sloan, and Jun Wang. Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users. *CoRR*, 2012.