トピックを用いたコンテンツ連動型広告の検索

山本 浩司　　野口 正樹　　小野 真吾　　塚本 浩司
ヤフー株式会社
{koyamamo, manoguch, shiono, kotsukam}@yahoo-corp.jp

1　はじめに

Web ページのコンテンツの内容に関連した広告を表示するコンテンツ連動型広告は，オンライン広告の中でも主要な位置を占めている．ここで広告を選択する際の基準として，コンテンツ内の単語と，広告文の中の単語の類似度が挙げられる．しかし，この基準の問題点として，もしコンテンツに適した広告があったとしても，コンテンツと広告とで用いられる語彙が異なっている場合には，適切な広告を表示の候補にできないという点が挙げられる．そのため，たとえば広告に出現した語の同義語や類義語がコンテンツに存在しても，広告とコンテンツの内容の一致を考慮できないことになる．また，異なる意味で用いられる多義語がコンテンツと広告文の両方に出現することにより，関連がない広告にもかかわらず表示される候補になってしまう問題もある．本稿では，上記のような語彙のミスマッチにより適切な広告を表示候補に出来ないという問題を解決するため，ページのコンテンツと広告文の双方の単語をトピックに変換し，変換したトピックの空間上で広告を検索する手法を提案する．実際に表示された広告の中でクリックされたものを正解とした場合と，人手による評価を正解とした場合の評価を行った結果，単語によるコサイン類似度によるスコアとトピックによるスコアの線形和による検索を行う場合に，コサイン類似度のみで検索する場合よりも精度が向上することを示す．

2　トピックによる広告引当て

2.1　広告配信システム

本稿で対象とするコンテンツ連動型広告は，Web ページに表示されてクリックされた場合に広告主への課金が発生する,「クリック課金型」であることを想定している．クリック課金型の広告を配信する際の収益の期待値は，広告主の設定した入札額とクリック率 (CTR; Click-Through Rate) によって決定される．同じ広告であっても表示される Web ページやユーザによって CTR は一定ではないと考えられる．ここで CTR を正確に予測することができれば，期待収益を最大化する広告の候補を選択することが可能になるため，CTR の予測手法が多く提案されている [3][6] ．配信候補となる広告は数が多く，すべての広告に対して CTR の予測を行うことは現実的ではないため，まず検索技術を用いて候補となる広告の数を絞り，残った比較的少数の候補に対して CTR 予測を行うという 2 段階の処理が行われることが一般的である．本稿では前者の，広告を検索する処理を対象として述べる．

2.2　単語のトピックへの変換

広告文の中の単語を検索対象とする場合には，単語によって検索インデクスを作るが，本稿では語彙のミスマッチにより適切な広告を表示候補に出来ないという問題の解決のため，広告とコンテンツの双方のトピックへの変換を行う．1 つの広告文を構成する単語のベクトルからトピックへの変換を行い，単語ではなくトピックでインデクスを作成する．同様に，検索時のクエリとして Web ページから抽出した単語ベクトルをそのままクエリとするのではなく，単語ベクトルをトピックに変換してからクエリとして検索を行う．コンテンツや広告文をトピックの空間に変換する手法としては，コンテンツや広告文の単語ベクトルが潜在的なクラスタに属していると見なして各クラスタへの所属確率を学習する Ratnaparkhi の手法 [5] を用いる．ある広告とコンテンツの単語ベクトルをそれぞれ

$$\mathbf{ad} = (a_1, \ldots, a_n)$$
$$\mathbf{page} = (b_1, \ldots, b_m)$$

と表す．$a_1, \ldots, a_n$ は単語ベクトルの要素となる広告文内の各単語，$b_1, \ldots, b_m$ は同様にコンテンツ内の各単語を表す．

Ratnaparkhi[5] の学習手法は，クリックされた広告の単語ベクトルとそのときにその広告が表示されていたコンテンツの単語ベクトルのペアを学習データとして用い，EM アルゴリズムを用いて潜在パラメータで

ある各トピックでの単語の生起確率を推定する．これにより，広告とコンテンツの単語ベクトルが与えられたときに，これらがあるトピック c_i に所属している確率 $p(c_i|\mathbf{ad})$ および $p(c_i|\mathbf{page})$ がそれぞれ求められる．なお，トピックの総数 k は学習時に事前に決めておく必要がある．これらの値は，以下のようにそれぞれ広告 **ad** やコンテンツ **page** を構成する単語ベクトルの各要素の単語が各トピックで生起する確率の積に比例する．

$$p(c|\mathbf{ad}) \propto \prod_{j}^{n} q(a_j|c)$$

$$p(c|\mathbf{page}) \propto \prod_{j}^{m} q(b_j|c)$$

n, m はそれぞれ広告 **ad** とコンテンツ **page** に含まれる単語数である．各トピックでの単語の生起確率 $q(a_j|c)$ は学習で推定されるパラメータであるが，学習データに出現しない単語についてはそのパラメータが 0 となり，結果として積の値が 0 となってしまうのを防ぐために，元の手法に加え，ディリクレスムージングを行って各パラメータが 0 になることを回避する．

2.3 トピックによる広告検索

検索エンジンを用いて広告を検索するために，広告が表現するトピックでインデクスを構築する．インデクス作成のため，各広告に対し，$p(c_1|\mathbf{ad}), \ldots, p(c_k|\mathbf{ad})$ を求める．これらの値をスコア，それに対応するトピック番号 k を term の ID とし，インデクスを作成する．検索にかかるレイテンシを悪化させないために，各広告に対し，$p(c_i|\mathbf{ad})$ を上位 K 個のクラスのみをインデクスするようにする．

クエリとしては，広告を掲載しようとするコンテンツから抽出した単語ベクトルから，同様に $p(c_1|\mathbf{page}), \ldots, p(c_k|\mathbf{page})$, を求め，値の高い上位 K 個のトピックの値をクエリとして広告検索エンジンに問い合わせる．クエリを受け取った広告検索エンジンは，広告とコンテンツの各トピックへの所属確率の積の和

$$score_{topic} = \sum_{c_i \in topK} p(c_i|\mathbf{ad}) p(c_i|\mathbf{page}) \qquad (1)$$

を求める．

関連研究 [7] では，トピックモデルとして広く知られている LDA[1] を検索に用いることが提案されているが，LDA 単独では検索に用いる表現としては粗いという報告がなされており，従来の検索による手法と組み合わせて用いる手法が提案されている．本稿でも，単語による検索スコアとの線形和を用いる．単語による検索のスコアは以下のように，クエリの単語ベクトルと，インデクスの単語ベクトルのコサイン類似度とする．

$$score_{cos} = \frac{\sum_{j \in query} a_j * b_j}{\sqrt{\sum_{j \in query} {a_j}^2} \sqrt{\sum_{j \in query} {b_j}^2}} \qquad (2)$$

この値と，(1) 式のトピックのスコアの，パラメータ α による線形和を最終的なスコアとする．

$$score = \alpha * score_{cos} + (1 - \alpha) * score_{topic} \qquad (3)$$

広告検索エンジンはこの値が高い順に広告を最大で要求された本数分返却する．Ratnaparkhi[5] の手法では，トピックモデルのパラメータを用いてクリック予測の精度を向上させるアプローチを取っている． 本稿ではクリック予測の前段階としての，広告の引当ての適切さの改善を目的とする点で異なっている．

3 実験

本稿で述べた手法を用いて，ページに対して適切な広告の引当てができるかの評価を行う．引当ての適切さの定義としては，広告の CTR の高さや，ページとの関連度の高さが考えられる．そこで，実際に表示を行った際にクリックされたものを正解とした場合と，人手でページと広告の関連度を判定したデータを正解とした場合の 2 通りの評価を行う．評価指標として AUC(Area Under the ROC Curve)[4] を用いる．

3.1 クリック予測

広告の CTR の高さを引当ての適切さと考えた場合，クリックされやすい広告に対して高いスコアを与えるようなモデルを用いれば，クリックされやすい広告を引当てることができるといえる．そこで，実際のクリックログを用いてクリックの予測性能についての評価を行う．学習に用いるデータとして，Yahoo!ディスプレイアドネットワークの広告を掲載する，あるドメインの Web サイトの実際のクリックログを 86 日分，696,296 サンプルを用いた．クリックログには，表示されたページのコンテンツの単語ベクトル，そのページがリクエストされたときに配信された広告文に含まれる単語ベクトル，およびその広告に対するクリックの有無が記録されている．クリックログを用いた AUC 計算においては，このデータを期間によって 5 分割し，そのうち 1 つをテストデータ，残りの 4 つを学習データとする 5-fold クロスバリデーションを行う．学習データとして用いるときには，そのうちのクリックの発生したページと広告の単語ベクトルのみを抽出して用いる．このときのサンプル数は 176,957 である．学習のイテレーションを 15 回，広告検索時に用いる

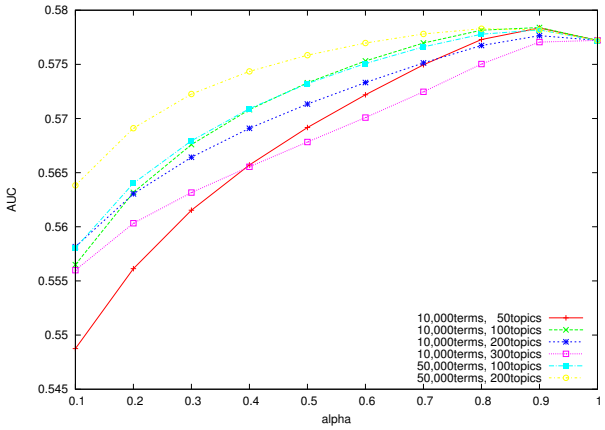

図 1: クリックログを用いた AUC

トピックの数 K は 10 とし，トピックの数は 50, 100, 200 および 300 とし，ページ側, 広告側ともに学習に用いる単語数を 10,000 語, 50,000 語の各場合について試行した．テストデータの各サンプルのページの単語ベクトルと，広告の単語ベクトルを用いて，(3) 式のスコアを計算する．クリックの発生したサンプルを正例，クリックの発生しなかったサンプルを負例とし，求めたスコアを用いて AUC を計算する．α については，[0,1] の区間で 0.1 刻みで変化させた値を用いる．

α を変化させた 5-fold での AUC の平均値を図 1 に示す．また，$\alpha = 0$ (トピックによるスコアのみを用いた検索) および $\alpha = 1$ (単語のみによる検索)，AUC が最大になることが多い $\alpha = 0.8$ および 0.9 のときの AUC を表 1, 表 2 に示す．単語数が 10,000 の場合には，トピック数が 50, 100, 200 のとき，いずれも α が 0.9 の場合に，AUC の平均値が最大となっている．ただし，トピック数が 300 の場合には，単語による検索のみを用いたとき AUC が最大となった．単語数が 50,000 の場合には，α が 0.8 または 0.9 の場合に AUC が最大となった．単語数を変化させても AUC の最大値は大きく変化しなかった．トピックによる引当てのみでは AUC が低く，単独でのクリックの予測性能が単語ベクトルのコサイン類似度より低かったが，トピックによる引当てを組み合わせることによって，単語ベクトルのコサイン類似度のみで引当てを行う場合 ($\alpha = 1.0$) よりも，AUC が向上した．これは，単語のみで検索するよりも，トピックによるスコアを考慮した検索を行うことで，単語だけでは検索できないが内容としては関連のある広告を検索できることを示唆している．

トピック数 /α	0	0.8	0.9	1.0
50	0.5066	0.5773	**0.5784**	0.5772
100	0.5083	0.5782	**0.5784**	0.5772
200	0.5048	0.5767	**0.5777**	0.5772
300	0.5002	0.5750	0.5771	**0.5772**

表 1: クリックログを用いた AUC (単語数:10,000)

トピック数 /α	0	0.8	0.9	1.0
100	0.5078	0.5778	**0.5782**	0.5772
200	0.5094	0.5783	**0.5782**	0.5772

表 2: クリックログを用いた AUC (単語数:50,000)

3.2 人手による評価を用いた実験

引当ての適切さを評価する別の指標として，Web ページと広告の関連性を人手で評価したデータを利用し，それを正解とみなした AUC を求める．スコアの計算に用いるモデルは，前節のクリック予測に用いたものと同様の手法で学習したモデルを用いる．評価データの収集に際しては，学習データと同じドメインの Web サイトから，モデルの学習データの期間に含まれない Web ページをランダムに 51 件取得した．これらの Web ページのそれぞれに対し，トピックのみによる検索 ($\alpha = 0$) と，単語のみによる検索 ($\alpha = 1.0$) の 2 通りの手法によって広告の引当てのみを行った．それぞれのコンテンツから検索した広告に対し，Web ページとの関連性を人手によって “good”, “fair”, “bad” の 3 段階で評価した．1 つの Web ページに対し，トピックによる検索 ($\alpha = 0$) と単語による検索 ($\alpha = 1.0$) それぞれ最大で 10 本の広告を取得する．評価の際には，被験者に Web ページのスクリーンショットとともに，その Web ページから 2 つの手法で検索された広告最大各 10 本ずつをランダムに混ぜた順序で同時に提示する．各広告がどちらの手法で検索されたものかは明示しない．3 人の評価者による，のべ 2,420 件の評価が収集された．

この評価で使われたページと広告の組に対し，(1) 式のスコアを計算する．評価データが 3 段階評価のため，各評価のペア $x_i, x_j (i < j)$ ごとに評価の値を比較して正例または負例を新たに作り，その新たなデータによって AUC を求める．ペアのどちらの評価が良いかに応じて，x_i の評価のほうが良ければ正例, x_j の評価のほうが良ければ負例とする．評価が同じであれば，そのペアについてはデータとして用いない．

ペアから正例または負例が作られる場合，そのスコ

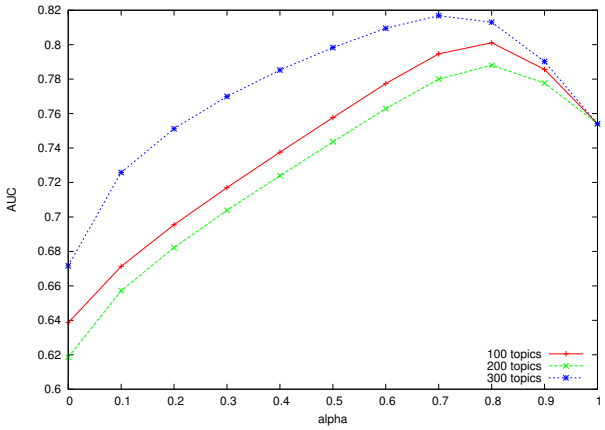

図 2: 人手による評価を用いた AUC (単語数:10,000)

アの部分は，両者のスコアの差分 (x_i のスコア - x_j のスコア) とする．このデータを差分を取った後のスコア順に並べて AUC を求める．クリックログを用いた場合と同様に，α を 0.1 刻みで変化させて AUC を算出する．

単語数を 10,000，トピック数は 100, 200, 300 の場合の AUC を図 2 に示す．数値は 5-fold の AUC の平均値を評価者毎に求め，各評価者の評価したページ数に応じて，全評価者の平均を取ったものである．概ね $\alpha = 0.7$ または 0.8 の領域において AUC が最大となり，人手による評価を用いた場合でも，コサイン類似度のみを用いる場合よりも，トピックのスコアを考慮した場合のほうが AUC が向上した．これにより，トピックのスコアを考慮することが，クリックの予測だけでなくコンテンツと広告の関連度の評価についても有効であることがわかった．

3.3 考察

2 つの実験での評価指標には傾向の異なりが見られ，例えば単語数が 10,000，トピック数を 300 としたモデルでは，クリックを正解とした場合には AUC が向上しないが，人手による正解データを使った場合には AUC が向上した．そして AUC が最大となる α の値も人手による評価を正解にした場合のほうがやや小さい値となった．この結果は，トピックモデルがクリックログを用いて学習しているものの，クリックログに対する予測を行う場合よりも，人手の評価に対して精度を改善する影響が大きいことを示している．

また，単語のみで語彙の一致がないと検索できない場合よりも精度が向上していることから，語彙の一致はないが関連性のある広告についても検索できているものがあることが窺える．単語による一致がなくても関連性のある広告を検索できた例として，ページと広告，ページの単語から求まる $p(c|\mathbf{page})$ が最大になるトピックの確率上位 5 単語を表 3 に示す．

ページ	広告	$p(c\|\mathbf{page})$ が最大となったトピックの確率上位 5 語
車のページ	アルミホイールの広告	免許, 車検, 新車, メンテナンス, カスタマイズ
入試のページ	予備校の広告	学校, 子育て, 国際, 中学校, 幼稚園

表 3: 単語による一致がなくても関連性のある広告を検索できた例

4 まとめ

本稿ではトピックを用いたコンテンツ連動型広告の引当て手法を提案した．単語による検索とトピックによる検索を組み合わせることで，両者を単独で使う場合よりも広告検索の精度が向上することを示した．今後の課題として，本稿の 2 つの実験で最適な α の値が異なる原因についての検証や，実際の配信における効果の検証が挙げられる．

参考文献

[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan: Latent Dirichlet Allocation, *Journal of Machine Learning Research 3*, pp. 993–1022, 2003.

[2] Adrien Bougouin, Florian Boudin and Beatrice Daille: TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction, In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (IJCNLP2013), pp. 543–551, 2013.

[3] 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* (WWW2008), pp. 417–426, 2008.

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

[5] 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.

[6] Yukihiro Tagami, Shingo Ono, Koji Yamamoto, Koji Tsukamoto and Akira Tajima: CTR Prediction for Contextual Advertising: Learning-to-Rank Approach, In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising* (ADKDD2013), 2013.

[7] Xing Wei, W. Bruce Croft: LDA-Based Document Models for Ad-hoc Retrieval, In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR2006), pp. 178–185, 2006.