HTML 文書からのリスティング広告の自動生成*

幾島 克洋 †1　藤田 篤 †2　佐藤 理史 †2　横川 睦 †3　岩本 宜式 †3　片岡 亮 †3
†1 名古屋大学工学部電気電子・情報工学科　†2 名古屋大学大学院工学研究科
†3 株式会社リクルート インターネットマーケティング局

1　はじめに

インターネットの普及率向上，機能向上に伴い，製品やサービスの広告手段としてインターネットおよび検索エンジンを利用する**検索エンジンマーケティング**(Search Engine Marketing; SEM) が盛んになってきた．特に，クリック回数に応じて費用が決定する検索連動型広告やコンテンツ連動型広告が急速に普及しており，ある試算 [1] では，国内の売上高は 2007 年に約 1,300 億円まで拡大したとされている．

検索連動型広告（**リスティング広告**）は，検索エンジンが検索クエリに応じて検索結果画面に表示させる広告である（図 1）．広告がクリックされることによって初めて課金が発生するため，広告掲載の費用対効果に優れており，SEM 市場における売上高の大半を占めている．リスティング広告は，掲載順位が上位である（画面上部に表示される）ほどクリックされやすい [5]．このため，Google AdWords [1]や Overture[2]などのリスティング広告サービス（スポンサードサーチ）では，広告の入札価格だけでなく広告表示回数に対するクリック回数 (Click Through Rate; CTR) が高いほど表示順位を高くする戦略を取っている．

クリック課金という仕組みは，ロングテールと呼ばれる，ニッチで多様なコンテンツを活かす機会をもたらした．しかしながら，コンテンツの規模が大きくなるほど，個々のコンテンツに応じた広告を人手で作成することは困難になる．テンプレートにキーワードを埋め込むなどの方法がとられることもあるが，コンテンツ固有の情報を的確に反映できないため，ミスリードやアクセスの機会損失を誘発する恐れがある．

このような背景から，人的コストの軽減，CTR の向上を目的として，我々は，HTML 文書からリスティング広告を自動生成するシステムを作成した．本稿では，大規模なコンテンツを有するサイトを対象としたリスティング広告の生成システムについて述べ，CTR の測定実験を通じて得られた知見について報告する．

図 1: リスティング広告の例 (Google AdWords)

2　リスティング広告と CTR

本研究では，以下の特徴を持つ Google AdWords を基盤としてシステムの開発を進めている．

- 1 つの HTML 文書に対して複数の広告を入稿できる
- 入稿した広告に対する審査が自動化されており，即時性に優れている

以下，本稿では，AdWords 向けのリスティング広告を単に広告と略記する．

2.1　広告の構成要素

1 件の広告は次の 5 つの情報で構成される．

k: 検索キーワード，t: 見出し，d: 説明文，
u_l: クリック時のリンク先 HTML 文書の URL，
u_p: 広告内に表示する URL

AdWords ではリスティング広告に関する種々の制約が設けられている[3]．例を下に示す．

- t は全角 12 文字（半角 25 文字）以内
- d は全角 17 文字（半角 35 文字）を 2 行以内
- 標準的な日本語を誤字，脱字なく用いること
- 同じ表現を繰り返し使用しないこと
- 疑問符を適切に使うこと
- t 中に感嘆符を使わないこと
- d 中に感嘆符を使うとしても 1 つに留めること
- 大文字アルファベットを過度に含まないこと．ただし略語は構わない

*Automatic Generation of Ads from HTML Documents.
Katsuhiro Ikushima†1, Atsushi Fujita†2, Satoshi Sato†2, Mutsumi Yokokawa†3, Yoshinori Iwamoto†3, Ryo Kataoka†3
†1Department of Electrical and Electronic Engineering and Information Engineering, School of Engineering, Nagoya University
†2Graduate School of Engineering, Nagoya University
†3Internet Marketing Office, RECRUIT CO., LTD.

[1] https://adwords.google.co.jp/
[2] http://www.overture.co.jp/ja_JP/srch/index.php
[3] https://adwords.google.co.jp/select/guidelines.html

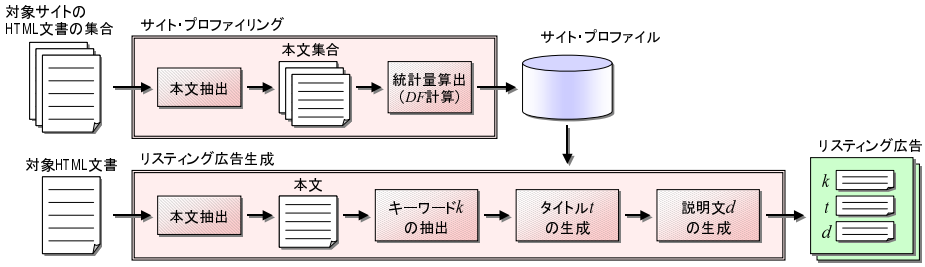

図 2: リスティング広告自動生成システム概要

2.2 先行研究における知見

文献 [2] では，広告と通常の検索結果に対するユーザの振る舞いに関する次のような傾向を報告している．

傾向 1. ユーザによっては，広告は通常の検索結果よりも信頼性に欠けるとみなす．

傾向 2. ユーザが広告に関連性がないと判断する主な観点は広告の見出し，次いで説明文である．

傾向 3. ユーザが広告に関連性があると判断する際は，通常の検索結果よりも説明文が重要視される．

我々自身もこれまでの予備調査を経て，広告の言語的特徴と CTR について次のような傾向を観察している．

傾向 4. t および d に k を含めると CTR が上昇する．また，k そのものでなくても，類義語，上位概念語を d に含めることで CTR を高くできる（傾向 2 に関連）．

傾向 5. 個々のコンテンツに関連のある数値を d に含めると CTR が上昇する（傾向 3 に関連）．

傾向 6. 広告中の文章の読みやすさも CTR に影響する．

ただし，いくつかの傾向は，サイトの持つ情報の種類に特化している可能性があることを断っておく．

3 リスティング広告生成システム

我々が作成したシステム（図 2）は，1 つの HTML 文書に対して複数の $\langle k, t, d\rangle$ を生成する．この節では，サイト全体の性質をとらえるためのサイト・プロファイリングと，広告生成の手順について述べる．

3.1 サイト・プロファイリング

大規模なコンテンツを持つサイトでは，そのサイトの名前やトップページへのリンクのアンカーテキストのように，複数の HTML 文書で共通の表現が用いられやすい．このような表現を広告の要素として用いることは適切でないので，あらかじめサイト中の文書集合から表現の重みを算出しておくことにした．この過程をサイト・プロファイリング（図 2 上段），結果の統計量をサイト・プロファイルと呼ぶ．サイト・プロファイリングは次の 2 つのステップからなる．

ステップ 1. 本文抽出: 各 HTML 文書から本文を抽出する．これには，解放型検索エンジン基盤 TSUBAKI において文書のインデキシングに用いられている標準フォーマット変換ツール[4] を改変して用いている．

ステップ 2. 統計量算出: 各文を形態素解析し，個々の内容語 w の出現文書数 $DF(w)$ を求める．

3.2 リスティング広告生成

サイト・プロファイルを用いて，入力された HTML 文書に対する複数の $\langle k, t, d\rangle$ を生成する（図 2 下段）．

ステップ 1. 本文抽出: サイト・プロファイリングと同様にして HTML 文書から本文を抽出する．

ステップ 2. キーワード k の抽出: HTML 文書中に出現する各名詞句 np のスコア $S_K(np)$ を次式によって算出し，スコアの上位 n_K 個を抽出する．

$$S_K(np) \quad = \quad \sum_{w \in cw(np)} S_W(w) \sum_{p \in Pos} f(np, p)\lambda(p).$$

ここで，$f(np, p)$ は文書中の位置 p における np の出現回数，$\lambda(p)$ は位置ごとの重みである．位置 Pos としては，本文，HTML 文書の META-keyword タグ，TITLE タグを区別する．$cw(np)$ は np を構成する内容語の集合であり，その各要素 w に対するスコア $S_W(w)$ は次式で与える．

$$S_W(w) \quad = \quad tf(w) \log_e(N/DF(w) + 1).$$

ここで，$tf(w)$ は文書中の w の出現回数，N はサイト・プロファイリングに用いた文書数を表す．

ステップ 3. タイトル t の生成: t に k そのものを含めることで CTR の向上が見込める（2.2 節傾向 4）．そこで，k を含む名詞句，動詞句を本文から抽出して t として用いる．この際，簡単な言い換え（名詞句化，動詞句化）も施し，複数の候補を可能な限り生成する．t の各候補に対して，スコア $S_T(t)$ を次式で算出し，スコアの上位 n_T 個を出力する．

$$S_T(t) \quad = \quad \sum_{w \in cw(t)} S_W(w).$$

[4] http://nlp.kuee.kyoto-u.ac.jp/~skeiji/html2sf.tgz

表 1: 文節間の関係のスコア S_R

関係	例〈係り元，係り先〉	S_r
動詞の格関係 動詞が名詞を修飾する関係 動詞がサ変名詞化したもの	〈アイデアを，募集する〉 〈募集した，アイデア〉 〈アイデアの，募集〉	3
形容詞・形容動詞を含む関係 名詞間関係（並列以外）	〈新しい，アイデア〉 〈アイデアが，新しい〉 〈騒音の，レベル〉	2
並列	〈鉄と，銅〉	1
副詞を含む関係	〈すでに，乗り込んだ〉	0

ステップ 4. 説明文 d の生成: 文献 [4] のアルゴリズムを参考に，次の手順で本文から d を生成する．なお，現状では k は参照していない．

1. 本文を係り受け関係にある文節対 $\langle b_1, b_2\rangle$ の集合に変換する．
2. 各文節対のスコア $S_C(\langle b_1, b_2\rangle)$ を，文節間の関係のスコア $S_R(\langle b_1, b_2\rangle)$（表 1）と各文節のスコア $S_B(b)$ に基づいて算出する．

$$S_C(\langle b_1, b_2\rangle) = S_R(\langle b_1, b_2\rangle)\big(S_B(b_1) + S_B(b_2)\big).$$
$$S_B(b) = \max_{w \in cw(b)} S_W(w).$$

3. スコアが最も高い係り受け文節対を取り出す．これをコアと呼ぶ．
4. スコアが高い係り受け文節対から順に，次の 5 つの要件を満たす限り，順次コアに接続する．
   - 係り受け文節の一方をコアと共有する
   - 仮に接続しても指定されたバイト数を超えない
   - 係り受け距離が閾値 len を超えない
   - 接続する文節が代名詞を含まない
   - 係り元の品詞が動詞，名詞–非自立，名詞–副詞可能のいずれでもない
5. 表現を整形する．
   - 不要な読点，括弧記号，代名詞などを削除
   - d の末尾の助詞を削除
   - d の末尾の動詞の活用形を修正
6. コアの選び方を第 2 位，第 3 位の係り受け文節対とすることで，d の候補を複数生成する．各 d のスコア $S_D(d)$ を次式で算出し，スコアの上位 n_D 個を出力する．

$$S_D(d) = \sum_{\langle b_1, b_2\rangle \in edge(d)} S_C(\langle b_1, b_2\rangle).$$

$edge(d)$ は d 中の係り受け文節対の集合を指す．

4 リスティング広告の CTR 測定実験

サイト運営者にとっては，低いコストで多くの顧客を獲得することが最も重要な課題である．個々の HTML 文書を通じての顧客獲得数がその文書への訪問者数に比例すること，集客戦略としてリスティング広告を用いることを前提とすると，上記の課題は次の 2 つの問題に置き換えられる．

- いかに低いコストで広告を掲載するか
- いかに各広告を通じての訪問者数を増やすか

1 つ目の問題は，クリックに対する単価が低くなるように対立候補が少ない（ただしそれなりに検索される）キーワードを選定すること，2 つ目の問題は，CTR を上げることが解決策となる．今回は，CTR の測定実験を通じて 2 つ目の問題に対するシステムの性能を調べた．

4.1 リスティング広告の自動生成

株式会社リクルートの事業サイトのうち，今回は，大量のコンテンツを持っており，かつ CTR 改善のニーズが高い『とらばーゆ』[5]を対象とした．

まず,『とらばーゆ』から得た 42,954 文書からサイト・プロファイルを作成した．そして，サイト中で比較的浅い階層の広い範囲から取り出した 540 文書に対して，3 節の手順で広告を生成した．広告生成時の各パラメタを，λ(Meta-description) $= 1$，λ(TITLE タグ中) $= 1$，λ(本文) $= 0$，$len = 10$，$n_K = 3$，各 k に対して $n_T = 3$，$n_D = 2$ とした結果，2,854 件の広告が得られた．自動生成した広告の例を示す．

(1) k: 法律事務，t: 法律事務の転職・求人情報，
d: 海外からの特許出願をサポートする

(2) k: 臨床検査技師，t: 臨床検査技師の女性，
d: 睡眠呼吸障害を専門に扱う

CTR の比較対象として，自動生成した広告の各 $\langle u_l, k\rangle$ に対して 1 件ずつ，例 (3) のテンプレートを用いて広告（以下，既存の広告）を生成した．

(3) t: k の求人，d: k の仕事なら

なお，自動生成，既存の両広告ともに，d の 2 行目は「とらばーゆ。お宝求人掲載中。」という定型文とした．

4.2 CTR の測定結果と考察

自動生成した広告 2,854 件と既存の広告 1,368 件の合計 4,222 件を，2008 年 1 月 9 日から 2008 年 1 月 22 日までの 14 日間，Google AdWords のスポンサードサーチに掲載した．なお，掲載期間中に広告の入札価格は変化させていない．結果，表 2 に示すように，約 26% (1,093 / 4,222) の広告に対して CTR が測定できた．ただし，表示回数が少ない場合は，CTR の信頼性が保証されないので，主に表示回数が 100 以上の広告を分析の対象として傾向を調査した．

[5]女性のための転職サイトとらばーゆ，http://toranet.yahoo.co.jp/

表 2: 表示回数 (imp) と広告数

	入札	$imp \geq 1$	$imp \geq 100$
既存の広告	1,368	344	147
自動生成した広告	2,854	749	341
合計	4,222	1,093	488

表 3: 各広告集合の CTR

	広告数	Ave_{micro}	Ave_{macro}
既存の広告 ($imp \geq 1$)	344	1.28%	0.92%
自動生成した広告 ($imp \geq 1$)	749	0.49%	0.66%
既存の広告 ($imp \geq 100$)	147	1.28%	1.24%
自動生成した広告 ($imp \geq 100$)	341	0.46%	0.64%

表 4: $\langle u_l, k \rangle$ ごとに見た CTR の向上可能性

最低表示回数	目標達成率
1,000	63% (24/ 38)
500	54% (26/ 48)
100	44% (63/142)

4.2.1 全体の統計的評価

まず，複数の広告の CTR をマイクロ平均 (Ave_{micro})，マクロ平均 (Ave_{macro}) で総合的に定量化して，既存の広告と自動生成した広告を比較する．

$$Ave_{micro} = \frac{\sum_{td \in TD} click(td)}{\sum_{td \in TD} imp(td)}.$$

$$Ave_{macro} = \frac{1}{|TD|} \sum_{td \in TD} \frac{click(td)}{imp(td)}.$$

ここで，$imp(td)$ は広告 td の表示回数，$click(td)$ は広告 td のクリック回数を表す．TD は広告の集合である．各広告集合の CTR の平均を表 3 に示す．

自動生成した広告はすべての平均で既存の広告に敗れた．特に，表示回数の多い $\langle u_l, k \rangle$ の中に CTR が高いものがあった影響でマイクロ平均の差が大きい．一方，個々の広告を平等視するマクロ平均の差は小さくなる．このことから，自動生成した広告の中には既存の広告よりも高い CTR を持つ広告もあったと考えられる．次節で詳しく見る．

4.2.2 文書・キーワードごとの比較

表示回数が 100 回以上の広告（表 2 最右列）のうち，既存の広告と自動生成した広告が同じ $\langle u_l, k \rangle$ に対応するものをグループ化したところ 142 グループが得られた．ここで，CTR が低い広告は掲載順位が下がり，自然淘汰される（不要なコストも発生しない）ことを考慮すると，個々の $\langle u_l, k \rangle$ について 1 件でも既存の広告よりも高い CTR を持つことを目標達成としても差し支えない．広告の最低表示回数の閾値を変えて，目標の達成率を計算したところ，表 4 のようになった．データは十分多いとは言えないが，約 50%の $\langle u_l, k \rangle$ に対して既存の広告よりも高い CTR を持つ広告を自動生成できたということは良好な結果と言える．

一方，既存の広告よりも CTR が低かった広告には，大きく次の 2 つの問題が含まれていた．

- t または d に非言語的な表現が含まれる
- d に k と関連のない表現が含まれる

1 つ目の問題は文整形のアルゴリズムを改良することで改善できる．2 つ目の問題は，d を k とは独立に生成していることに起因する．たとえば，k としては，具体的な職種ではなく雇用形態を表す表現（「アルバイト」，「パート」など）や，サイト全体に現れる表現が採用される場合がある．しかし，このような表現は DF が大きいため S_W の値も低く，d を生成する際の要素として選択されにくくなってしまう（3 節のステップ 4）．この問題を解決するには，d の生成時に，k または k の関連語を重要視するような改良が必要である．

5 おわりに

本稿では，HTML 文書から自動的にリスティング広告を生成する手法を提案した．商業サイトの中には，HTML 文書横断的にサイト全体にのみ関連する表現が用いられやすいことをふまえ，サイト全体における各内容語の重要度を事前に計算しておき，広告の生成に利用した．広告の各要素は，本文中の表現を抽出，再構成して生成した．実装したシステムで自動生成した広告の CTR を測定する実験を通じて，自動生成した広告の中には，テンプレートで生成された既存の広告よりも高い CTR を得るような広告もある程度あることを確認した．リスティング広告サービスの仕組みを考えると今回実験対象としたサイトのように多様なコンテンツを含むサイトにおける広告自動生成の効果は大きいと考えられる．

今後は，実証実験を継続しつつ，より高い CTR を得られる広告の自動生成を目指し，広告中の非言語的な表現の除去・整形処理の実装，ユーザのクリックを促すような言語的因子 [3, 5] の解明に取り組みたい．

参考文献

[1] アウンコンサルティング株式会社. 第 3 回（2008 年）国内 P4P 市場規模予測. http://www.auncon.co.jp/ir/pdf/20080108-1.pdf, 2008.

[2] B. J. Jansen and M. Resnick. Examining searcher perceptions of and interactions with sponsored results. In *Workshop on Sponsored Search Auctions, the 6th ACM Conference on Electronic Commerce* (*EC*), 2005.

[3] 西原陽子, 砂山渡, 谷内田正彦. 聴講者の興味をひく研究発表タイトルの作成支援. 言語処理学会第 13 回年次大会発表論文集, pp. 448–451, 2007.

[4] 岡満美子, 小山剛弘, 上田良寛. 句表現要約の句合成手法. 情報処理学会研究報告, NL-129-15, pp. 101–108, 1999.

[5] M. Riachardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International World Wide Web Conference* (*WWW*), pp. 521–529, 2007.