D-026

ジャストインタイムウェブ広告におけるタクソノミ自動生成手法
Automatic Taxonomy Generating Methods for Just-in-Time Web Advertising

櫻庭 敦之 * 成田 龍太 * 全 眞嬉 * 徳山 豪 *
Atsuyuki Sakuraba Ryota Narita Jinhee Chun Takeshi Tokuyama

1 はじめに

近年のインターネットの急速な普及により, ウェブ広告は急速な成長を見せている. ニュースサイトのような動的なウェブサイトにおいて関連性の高い広告を表示するシステムが現在よく使用されているが, そのようなシステムの設計においては, 広告の関連性の精度とシステムの効率にはトレードオフがある. 本論文ではウェブサイトの内容と関連性の高い広告をジャストインタイムで提供する効率的な手法を提案する. この問題に対する既存研究では, 人間の手で広告を分類したタクソノミをデータセットとして用いており, タクソノミ生成に手間がかかり, 最新の状態で広告を提供することが困難である.

本論文では, 提供される広告を自動的に木構造に分類し, タクソノミを自動生成するシステムを提案し, その精度と有用性を調べる.

2 関連研究との比較

タクソノミを用いたウェブページと文章広告とのマッチング手法は Broder らにより提案された [2]. この手法は, あらかじめ文章広告を大規模な階層的タクソノミに分類しておき, ウェブページとの関連性の高いものを算出する方法によってウェブページと文章広告文脈的な類似度を計算するものである. Anagnostpoulos らはウェブページを表示するときにリアルタイムで文章広告を表示するために, ウェブページの要約部分とタクソノミを使い, 高い精度を保ちながら高速に計算する手法 [1] を提案した.

これらの研究で扱われる文章広告は, 文字数に制限がある場合が多く, 短い文で効率よく情報を伝えようとするため情報量が限られている. このため, 人間はある単語から関連する事柄を連想できるため, 広告同士の類似度を正しく評価できるが, コンピュータは異なる単語を同一の意味でとらえることができず, 正確な評価は困難である.

情報量が少ない文書を高い精度でカテゴリ分けするには, 文書に含まれる情報をもとに, 関連する情報を追加する手法が使われる. Nuntiyagul らは, カテゴリにキーワードをつけたものをタクソノミとして, 数学の問題文に現れる単語と比較して基礎計算, 図形などのカテゴリ分けをする手法 [5] を提案した. また, Gabrilovich らは, Open Directory Project[6] のようなオンライン上のディレクトリサービスなどを用いて文書にキーワードを追加してからカテゴリ分けする手法を提案した [4]. この手法では単語の多義性や単語間の同義性, 関連性といったこれまでのカテゴリ分け手法での問題を, オンライン上で得たキーワードで解決する手法が提案された.

しかし, これらのタクソノミおよびディレクトリサービスとして精度のよいものを作るには人間の手に頼るのが一番であるが, これを構築または更新するためには多くの時間がかかり非効率的である. 実際に, Open Directory Project では世界中のボランティアがカテゴリ分けの作業をしており, コストをかけずに大規模なディレクトリを最新状態に維持している.

そこで, 本研究では文章広告からタクソノミを自動生成する手法を提案し, 人間の手で行われていた分類作業を自動し, 効率的に大規模なタクソノミを構築する手法を提案する.

3 タクソノミ自動生成システムの設計

3.1 システムの概要

本システムは以下の入出力を持つ.

- 入力
  キーワード付きの文章広告群
- 出力
  文章広告を階層的に分類したタクソノミ

出力されるタクソノミは, 階層構造を保持するために xml ファイルとして出力する. 図 1 にツリー構造で表示した出力例を示す.

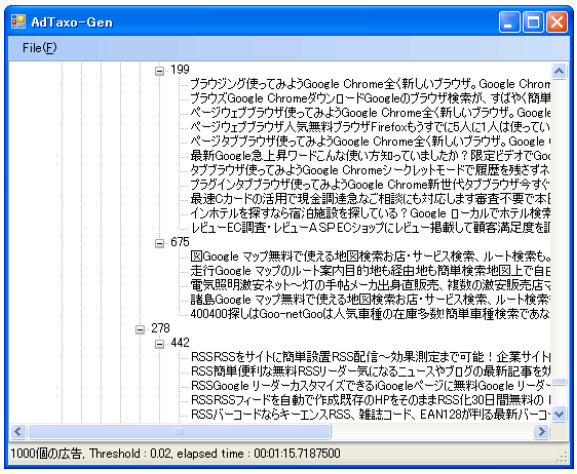

図 1: 出力例

本システムでは, 文章広告同士の関連性を算出するためにベクトル空間モデルを用いる. ベクトル空間モデルは, 出現する単語に基づいて文書または文章をベクトルで表現し, その向き

*東北大学大学院情報科学研究科

によって内容を判断するものであり, 情報検索分野で広く利用されている [3].

本研究では tf-idf(term frequency-inverse document frequency) 法によってベクトルに含まれる単語に重みを付ける. tf-idf 法は, 特定の単語がその単語が出現する文書が少なく, かつ特定の文書に多数出現するとき, その文書においてその単語は重要であるとする考えかたで重要度を示す方法である. ベクトル表現された広告同士の類似度はベクトルのコサインの値で 0 から 1 で表す.

3.2 広告の前処理

本システムでは, オンラインから無作為に抽出した文章広告を入力として用いる. この文章広告は, リンクが張られているウェブサイトに含まれる名詞の tf-idf 値を計算し, これを重要度順キーワードとして加える. そもそも, 文章広告は文章自体が短く, 情報量が少ない場合や広告の本来意図しているものと違う文面に場合があるため, キーワードによって適切に情報を追加することが有効であると考える.

3.3 欲張り法を用いたタクソノミ自動生成手法

このアルゴリズムは欲張り法を用いており, 広告ペアの類似度が, 設定された閾値以上であるならば同じカテゴリに分類するという方針をとる. これをループさせることで, 階層構造のタクソノミが下の階層から生成される.

各ループでは, 最上階層にあるすべての広告およびカテゴリのペアについてその類似度を計算した後, 類似度の高いものから降順でソートし, 類似度と閾値の比較を行う. また, 生成されるタクソノミの上位の階層には幅広いジャンルのカテゴリが含まれるべきであるので,1 回のループごとに閾値を半減させるようにした. これをボトムアップで実行し, 最後に一つのカテゴリにまとめる.

tf-idf 値を用いたカテゴライズでは, 広告のみを用いた場合, 単語の類似性を考慮した分類は困難である. そのため, 本研究では類似性を考慮した分類を行うために広告 A と広告 B が同じカテゴリに分類されており, なおかつ広告 C も広告 B との類似度が高ければ同じカテゴリに分類する, というように和集合をとるようにして分類をする. また, 前述のキーワードを加えて情報量を増やすことで類似語を含ませることを意図する.

欲張り法を用いた場合, 実際には類似度の低い広告同士が同じカテゴリに分類されて懸念があげられるが, これは類似度の高いペアから閾値との比較を行い, 閾値の設定次第で回避できると考える.

タクソノミ生成の流れを図 2 に示す.

4 結果と今後の課題

本研究では, 文章広告を階層型のカテゴリに分類し, タクソノミを自動生成する欲張り法を用いた手法と, キーワードの追加, 類似度による広告ペアのソーティングにより分類の精度を高める手法を提案し, 提案手法を用いて構築したシステムでタクソノミを自動生成し, 生成されたタクソノミの検証を行った.

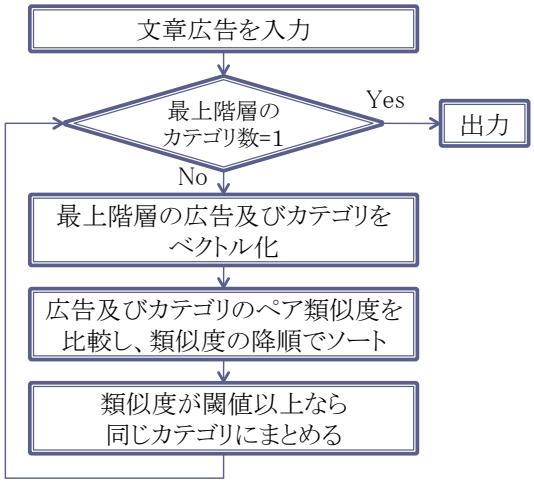

図 2: タクソノミ生成の流れ

生成されたタクソノミより, 各カテゴリが類似度の高い広告からなることが確認でき, 高精度なタクソノミの自動生成に成功した. 欲張り法を用いた際の懸念として挙げられた類似度の低い広告の混入もほとんど起こらなかったことも確認できた. なお, 実験結果の詳細は講演当日に発表する.

今後の改良点として挙げられるのは, 類似度が高く, 一つのカテゴリに分類されるのが望ましい広告が複数のカテゴリに分散してしまったことである. これは, 初回のループ時の類似度に依存していると考えられ, 適正な閾値の設定を更なる実験から求めてるとともに, アルゴリズムの改良をする予定である.

参考文献

[1] Aris Anagnostpoulos, Andrei Z. Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. “Just-in-time Contextual Advertising,” *Proceedings of the ACM 16th, Conference on Information and Knowledge Management(CIKM’07).* 331-340

[2] Andrei Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. “Robust classification of rare queries using web knowledge,” *Proceedings of the 30th ACM International Conference on Research and Development in Information Retrieval* 2007

[3] 大谷紀子, “情報検索におけるベクトル空間モデルの応用” 武蔵野工業大学環境情報学部研究論文 3-6

[4] Evgeniy Gabrilovich and Shaul Markovitch. “Feature Generation for Text Categorization Using World Knowledge,” *IJCAI’05, pages 1048–1053, 2005.*

[5] Atom Nuntiyagul, Nick Cercone, and Kanlaya Naruedomkul. “Recovering “Lack of words” in Text Categorization for Item Banks,” *Proceedings of the 29th Annual International Computer Software and Applications Conference(COMPSAC’05)*

[6] Open Directory Project : http://www.dmoz.org/