3T-01

Paragraph Vector を利用したインターネット広告の高度化

青柳 志穂里，橋本幸二郎，土屋 健，三代沢正，広瀬啓雄，(諏訪東京理科大学)
澤野弘明(愛知工業大学)，小柳恵一 (早稲田大学)

Research on Improvement of Information Platform for Internet Advertising Using Paragraph Vector Shihori Aoyagi, Kojiro Hashimoto, Takeshi Tsuchiya, Tadashi Miyosawa, Hiroo Hirose (Tokyo University of Science, Suwa), Hiroaki Sawano (Aichi Institute of Technology), Keiichi Koyanagi (Waseda University)

1. はじめに

インターネット広告を限られた時間や予算のもとで効率的に配信するためには，その広告に興味を持ったり広告主の顧客になりそうなユーザに対して，優先的に広告を配信することが好ましい．そのため，ウェブ上の行動をもとにユーザをモデリングし，広告を表示するユーザを適切に絞り込むことは，非常に重要なタスクである．

近年，自然言語処理の分野では，ベクトル空間上での単語の分散表現が注目を浴びている．[文献(1)]従来のインターネット広告では，ユーザ一人の行動履歴に基づき，Cookie の履歴から以前閲覧してたものを表示させるものが基本であった．また，他の研究との比較という観点に対しても，広告推薦の世界ではデータ数の処理負荷の関係で，単純な直近の統計的な処理しか実装されておらず，自然言語解析の手法をサーバ履歴の解析に適用するといった研究が発表された程度であるため，リーチしたい広告，直近の行動に自然言語解析の手法を持ち込む方法は採用されていなかった．そのため，ユーザがその場で欲している広告を表示することができないという課題があった．そこで，本研究ではコンテンツのベクトル化に含まれる単語とその順序性に着目した Paragraph Vector(以降 PV)[文献(2)]を適用することで，上記課題の解決を目指し，ユーザ一人一人の趣味嗜好に適応した広告を表示させ，インターネット広告のクリック率を向上させるための対応を考える．

2. Paragraph Vector

PV とは，単語ではなく各ユーザが入力した文書の意味を，ベクトルで表現するものである．ユーザをパラグラフもしくは文書，ユーザ行動を単語と見なし，このベクトルモデルをユーザの行動列に対して適用する．[文献(1)]文書全体を 1 つのベクトルで表現することで，ベクトル化された文書は，意味的に関連が近い文書はベクトルが近くなり，2 つのベクトルの差は 2 つの文章の関係を表すという特徴を示す．その後，文書中に出現する単語から，前後に出現する単語の予測モデルを構築し，入力単語から文脈を示すベクトルを取得する．そして，各単語を表すベクトルを入力すると，単語の確率を出力とするニューラルネットワークを構築する．また，文書の長さに左右されることなく特徴を内包することが可能となるため，文書全体の文脈を保持することが可能となる．そして，意味的に関連が近い文書は，ベクトルが近くなることから問題はベクトルの類似に帰着すると考えられる．また，このモデルの構築には，ビッグデータに相当する一定量のデータが必要となる．

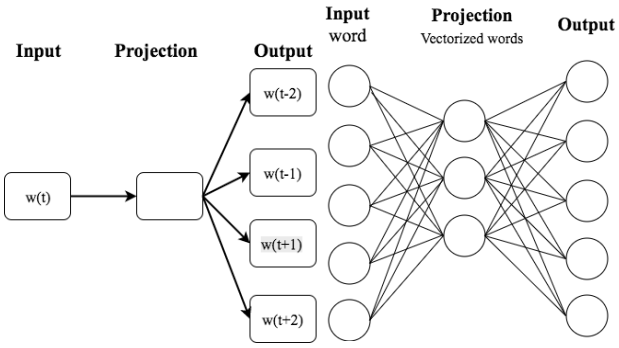

図 2 ニューラルネットワークのイメージ図

3. 提案手法

課題解決のための提案手法として，PV を用いた解析を行う．web ページを 1 パラグラフとして解析したのち，取得した各コンテンツ内の単語の順序性を学習させ，出現語の予測モデルを構築する．まず PV の優位性を確かめるため，スクレイピングにより収集された SNS を中心とした 3 ポップ内のリンク先をデータの対象とし，各単語のドキュメント内，全ドキュメントでの頻度をベースにして導出する tf,idf + LDA に基づく解析を行い性能の比較をする．その後，さまざまなサービスを利用した不特定多数のユーザの 1 週間分(数万件)の行動履歴を用いて，PV による解析を行う．ここでのユーザ行動履歴とは，ユー

ザが Web 上でどのようなサイトやコンテンツを参照したかという行動履歴のことを指す．ユーザ ID と閲覧したリンク先の URL をもとに，スクレイピングによりリンク先のコンテンツを収集し、データベースに保存する．その後，コンテンツをデータベースからユーザごと取り出し文書を単語に分解したのち、PV の機能を提供する Python ライブラリの実装である doc2vec を利用してモデルの構築を行う．

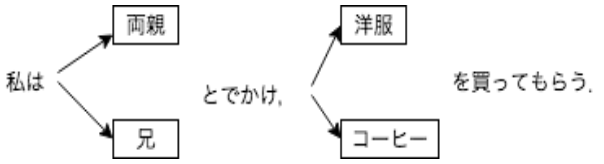

図 2 出現単語予測の例

ユーザ ID	リンク先 URL
1-TU2KcVUxjO	https://www.suwa.tus.ac.jp

スクレイピングにより収集

学校法人東京理科大学諏訪東京理科大学受験生の方保護者の方企業一般の方在学生卒業生の方大学案内教育学部大学院東京理科大学との連携入試情報キャンパスライフ就職進学地域貢献地域連携図書館交通アクセス……

図 3 ユーザ行動の履歴化の例

5. 評価と考察

提案手法の評価として，スコープ内の関係性と類似トピックの語順の相関性の観点から評価を行う．対象とする情報は，スクレイピングにより収集された SNS を中心とした 3 ポップ内のリンク先のデータ，また不特定多数のユーザの 1 週間分の行動履歴のうち，6 日分をモデル化に適用し，残り 1 日分を評価対象とし，ベクトル化を行った．類似語の導出は現在のデータ数においても tf,idf + LDA と遜色ない性能を有していた．また，データ数が多くなれば，さらに性能が上がると期待できる．

6. まとめと今後の課題

本研究では，PV を利用して収集したユーザの行動履歴をベクトル化に適用した ．このとき，行動履歴がさまざまなサービスを利用した不特定多数のユーザのものであった場合に，スコープ内の関係性と類似トピックの語順の相関性の観点から評価を行った．情報システムとしての性能の向上には，対象データが多い方が有利であり，検索クエリや web ページのコンテンツなどのさまざまな情報を用いることで，表現の質が向上することが期待される．また，実践環境に近い形でも運用できるのか検討していく必要がある．

文　献

(1) 田頭 幸浩, “オンライン広告におけるウェブ閲覧系列の分散表現の獲得”, 2016 年 6 月

(2) Quoc Le, Tomas Mikolov,“Distributed Representations of Sentences and Documents”,Proc. of Int. Conf. on Machine Learning, Beijing, China, 2014 vol. 32.