テキストと画像を入力とする学習者モデルの構築手法の提案

長谷川 忍[*1]，Wan Hua[*1]，太田 光一[*1]
*1 北陸先端科学技術大学院大学

A Proposal for Building Learner Models using Text and Images

Shinobu Hasegawa[*1], Wan Hua[*1], Koichi Ota
*1 Japan Advanced Institute of Science and Technology

The purpose of this research is to propose a method for constructing learner models from keywords and images of interest to individual learners in Web-based Learning by combining deep learning techniques. This makes it possible to represent learners' characteristics in self-directed learning in a vector representation. As a preliminary step, this article describes a method for generating common multimodal vector representations from text and images and confirms the characteristics of subjects' text-based and image-based preferences through preliminary experiments.

キーワード: Web-based Learning，学習者モデル，テキスト，画像

1. はじめに

Web-based Learning は膨大なリソースを持つ Web 空間において学習者が学習に役立つリソースを自由に選択し，主体的かつ網羅的に学習できる学習環境である(1). しかしながら, 学習項目や順序があらかじめ示されているテキスト教材とは異なり，Web 空間そのものが Open-ended であり，学習者の自由度が高いため，学習者が持つ興味や関心を適切に捉えることが難しい(2). さらに，Web リソースはテキストだけでなく画像や映像など様々なメディアを通じて情報が提供されるため, 学習過程を記録していたとしても個々の嗜好を反映した学習者モデルを構築することは困難である.

本研究では，テキストおよび画像に対する深層学習手法を組み合わせて，Web-based Learning 等において個々の学習者が興味を持っているキーワードや画像から学習者モデルを構築する手法を提案する．これにより，主体的学習過程における学習者の特性をベクトル表現で表すことが可能となる．本稿ではその前段階として，テキストおよび画像から共通するマルチモーダルなベクトル表現を生成する手法について述べるとともに，予備実験を通じて，被験者のテキストベースと画像ベースの嗜好性の特徴について検討する.

2. 関連研究

Ji らは，潜在因子モデル（LFM），重み付き行列分解（WMF），畳み込みニューラルネットワーク（CNN）を組み合わせたラベル無しデータセットにおける最適な特徴表現モデルを提案し，個人向け写真推薦システムを提案している(3). Savchenko らは，シーン理解，物体検出，顔認識に基づいて，ユーザの嗜好予測エンジンを開発しており，今後の課題として，テキスト認識技術が嗜好予測の信頼性を高めることを提案している(4). Díez らは, 個々のユーザがアップロードした写真からパーソナライズされた情報を抽出し，ユーザの主な嗜好を理解し，画像の同義性が将来的にユーザの嗜好と一致することに言及している(5). Wang らは，ウェブ上のテキストデータの数値表現を検討し，画像とテキストを組み合わせたマルチモーダルなフレームワークを提案している(6). Yao らは，テキストと視覚的特徴を利用して画像のモデリングと分類のタスクを行っている(7). Chen らは，ローカルとグローバルな画像プールから関連画像を推薦するソーシャルメディアプラットフォームの新しいアプリケーションを紹介している(8). このように, テキスト情報と画像情報を組み合わ

せ，近年急速に発展している深層学習技術を活用することで，画像分類や画像推薦を行う様々な研究が提案されている．しかしながら，これらの議論において，個々の対象者の嗜好を表現できるような，いわゆる「モデル化」を対象としたものはほとんど存在していない．

3. マルチモーダル嗜好性モデル

3.1 Text Preference Vector (TPV)

3.1.1 キーワードベクトルの生成

本研究では，学習者が選択または学習したキーワード群から学習者の嗜好性を表現するキーワードベクトルを生成することを目指す．具体的には，学習者が選択した各単語に対して，膨大な文書をコーパスとして収集し，word2vec や doc2vec のアルゴリズムを並列実装したオープンソースの自然言語処理ライブラリである Gensim[9]を用い，Wikipedia2014 +Gigaword トークンを元に事前学習したベクトルサイズ 50 の glove-wiki-gigaword-50 と Google News のデータセットの一部を元に事前学習したベクトルサイズ 300 の word2vec-google-news-300 によるベクトル化について比較した．埋め込み出力はベクトルサイズが大きい後者の方が忠実度や厳密性が高くなるが，時間的コストや計算の複雑さを考慮して，本研究では glove-wiki-gigaword-50 を選択することとした．

3.1.2 TPV の生成

各学習者の選択テキストに基づく嗜好性のモデルは，n 個のキーワードに対して，i 番目のキーワードに対するキーワードベクトルを $Embed(P_i)$，キーワードの出現頻度の総和が 1 となるように正規化したものを WT_i とするとき，式(1)により求められる．

$$TPV = \sum_{i=1}^{n} Embed(P_i) * WT_i \qquad (1)$$

これにより，glove-wiki-gigaword-50 が持つ 50 次元の中で各学習者のテキスト嗜好性をモデルとして表現することができる．

3.2 Image Preference Vector (IPV)

3.2.1 画像分類モデルの構築

各学習者が選択した画像から学習者モデルを構築するために，本研究ではまず画像分類モデルを構築した．データセットとして，32*32 ピクセル，10 クラス各 6,000 枚の画像を含む Cifar-10 と 100 の一般クラスと 20 の抽象クラスを持ち，各一般クラスに 600 枚の画像を含む Cifar-100 データセットに対して，それぞれ入力層と隠れ層に ReLU 活性化関数，出力層に SoftMax を配し，25 エポック，32 バッチサイズで CNN (Convolutional Neural Network)をトレーニングした．その結果，Cifar-10 データセットに対しては検証精度 79%であったが，Cifar-100 データセットでは検証精度は 44%に留まった．これらの結果から，本研究では Cifar-10 による画像分類モデルを利用することとした．

3.2.2 IPV の生成

Cifar-10 には飛行機や自動車，鳥などの 10 個のテキストラベルが存在し，新たな画像を入力することで，各ラベルに画像が分類され，全ラベルに対する総和が 1 となる確率が出力される．そこで，本研究では，各学習者の選択画像に基づく嗜好性のモデルを式(2)の形で求める．

$$IPV = \sum_{i=1}^{m} \left(\sum_{j=1}^{10} Embed(L_{ij}) * S_j \right) * WI_i \qquad (2)$$

具体的には，各画像 i に対して，Cifar-10 の 10 個のキーワードラベルそれぞれについて，3.1.1 節の手法で求めたキーワードベクトル $Embed(L_{ij})$と，3.2.1 節の画像分類モデルの出力である S_j を乗じて，その総和を取ったものに，画像の選択頻度の総和が 1 となるように正規化した WI_i を乗じて m 個の画像について総和を取ったものである． これにより，選択画像から抽出されたベクトル情報を，3.1 節のテキストに対するベクトル情報と同様な 50 次元で表現することができる．

3.3 ユークリッド距離

本稿で提案する TPV と IPV はいずれも 50 次元で構成されたベクトル表現となっているため，嗜好の類似性や違いはこれらのベクトルの距離で求めることがで

きる．1 例として，「動物」「ゲーム」「スポーツ」がそれぞれ 0.7，0.2，0.1 の重みで表される TPV を持つ学習者を想定した場合，図 1 に示す 2 種類の画像(32 × 32 ピクセルに変換したもの)に対する IPV ベクトルはそれぞれ図 2 のようになる．TPV に対する 2 種類の画像の IPV に対するユークリッド距離はそれぞれ IPV(dog) = 3.23 と IPV(AI) = 4.42 となり，テキスト情報から得られた嗜好ベクトルにより，画像に対する嗜好を推定することが可能となる．また，右図の AI の概念図は元々の Cifar-10 データセットのクラスには含まれていないクラスであるが，3.2.1 節の画像分類モデルの出力により Cifar-10 の各クラスへの分類確率が出力されるため，同様に取り扱うことができる．

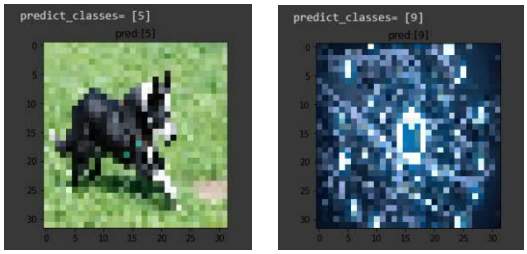

図 1. 入力画像(左：dog，右：AI)

```
[ 0.079452   -0.18018584 -0.56749153 -0.14972119  0.52327824  0.48888397
 -0.9144386  -0.35711497  1.0629215  -0.6874811  -0.00434914  0.3973705
  0.58873284  0.1434996   0.33573982 -0.11327112  0.3529217   0.97037464
 -1.3235319  -0.63535976 -0.47318769 -0.18532903  0.5084187   0.5450897
  0.5029429  -1.5349712  -0.996383    0.36899236  0.2576209  -0.5483999
  1.4407467   0.18684179 -0.43067747  0.70803463  0.05142963  0.2685303
  0.04932332 -0.22032823 -0.02764383 -0.66956335 -0.24391721  0.02221985
 -0.7180467   0.78386796  1.113106   -0.6324798   0.08800104 -1.0976168
  0.65203065  0.06439348]
```

```
[ 0.40721685 -0.30968162  0.8061926   0.20937891  0.2895352   0.5639808
 -0.98706806 -0.06757575  0.8009534  -0.58428097  0.17275801 -0.4664548
 -0.1561554   0.2376391   0.2733396  -0.2641144  -0.27900422  1.4549868
 -0.6896564  -1.2935286   0.14471814 -0.22190216 -0.33108807  0.43710762
  0.21466614 -1.1447489  -0.38195974  0.8836399   0.9627742  -0.34597522
  1.5204972  -0.28706053 -0.29911384  0.7369812   0.80522346  0.21978685
  0.10214268 -0.4735214   0.22229815  0.32472828 -0.39949733 -0.06418443
 -0.04040132 -0.71077466  0.7472847  -0.26451674 -0.04642011 -0.65324557
  0.7269966  -0.4535722 ]
```

図 2. ICV(上：dog，下：AI)

4. 予備実験

4.1 データセット

予備実験の画像データセットとして，本稿では図 3 に示す 3 種類 10 枚ずつの画像を google イメージ検索により収集し，全ての画像に対して 3.2 節に従って ICV を算出した．最初の 10 枚（グループ : Belong）は Cifar-10 データセットのクラスと一致した画像であり，次の 10 枚（グループ：Ambiguous）は Cifar-10 と同じクラスに分類されるものの，複数の対象物が含まれていたり，実際と異なるキャラクターなど曖昧性の高いものである．最後の 10 枚（グループ : No Related）は，Cifar-10 のクラスとはまったく関係のないものを選定した．

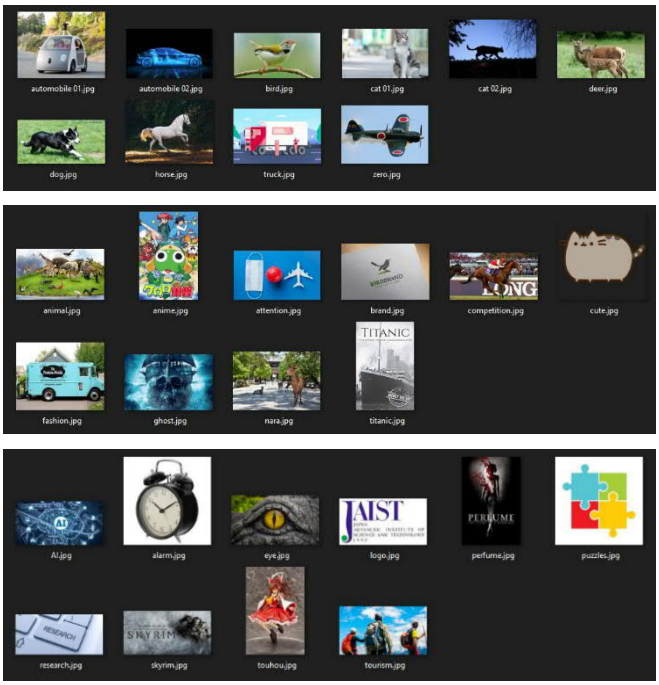

図 3. データセット

4.2 TPV と IPV による画像嗜好予測

4 名の大学院生を被験者として，画像と関連する 30 のキーワードから 3 つ選択し，総和が 1 となるような嗜好度を設定してもらい，テキストによる仮想的な嗜好ベクトルを生成した．次に，図 3 のデータセットのそれぞれのグループから 1 枚ずつ選択した 3 つの画像に対して，総和が 1 となるような嗜好度を設定してもらった．被験者につき 10 セットの嗜好度設定を行った結果，全部で 40 組のデータを得ることができた．

被験者による設定と TPV/IPV による嗜好予測の結果を比較したところ，3 つの順序が全て一致したケースが 6 組，一部の順序が一致したケースが 20 組，まったく一致しなかったケースが 14 組となった．

TPV と IPV の一致がうまく反映されなかった例として，ある被験者は「飛行機」「船」「タイタニック」をキーワードとして選択したにも関わらず，「飛行機」や「船」に関連する画像を優先して選択しなかったことが挙げられる．キーワードは対象の概念そのものを表すが，画像には（複数の）対象物だけでなく色合いや

スタイル，潜在的なメッセージなど多様なものが含まれるため，その点に関するギャップをいかに反映するかは今後の課題であると考えられる．

5. おわりに

本研究では，テキスト情報から生成される学習者の嗜好性ベクトル(TPV)と画像情報から生成される学習者の嗜好性ベクトル(IPV)の表現手法についての提案を通じて，Open-ended な Web-based Learning における学習者モデリングの手法について検討した．従来のテキストおよび画像のベクトル化手法が，画像分類や推薦課題に対して行われてきたことと比較して，個々の学習者の嗜好性の表現に利用できる可能性を示した．

本稿では，キーワードのベクトル化には単語埋め込み(GloVe)のみを適用したが，段落や文章全体を埋め込む手法の適用などへの拡張は今後の課題の一つである．また，画像に対しても，分類モデルに基づくベクトル化の手法を提案したが，オブジェクト認識などのより発展した手法を組み合わせることも可能であろう．今後はこれらの手法をさらに検討しながら，Web-based Learning の学習履歴を対象とした学習者モデリングに向けて検討を進めていきたい．

参 考 文 献

(1) Hasegawa, S., Kashihara, A., and Toyoda, J.: “A local Indexing for Learning Resources on WWW” Systems and Computers in Japan, 34(3), pp.1-9, (2003)

(2) 柏原昭博，坂本雅直，長谷川忍，豊田順一：”ハイパー空間における主体的学習プロセスのリフレクション支援” 人工知能学会誌, 18(5), pp.245-256, (2003)

(3) Ji, Z., Tang, J., Wu, G.: “Personalized Recommendation of Photography Based on Deep Learning”. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, WH., Vrochidis, S. (eds) MultiMedia Modeling. MMM 2019. Lecture Notes in Computer Science, vol 11295. Springer, Cham, (2019) https://doi.org/10.1007/978-3-030-05710-7_18

(4) Savchenko, A.V., Demochkin, K.V., Grechikhin, I.S.: “Preference prediction based on a photo gallery analysis with scene recognition and object detection”. Pattern Recognition 121, (2022) https://doi.org/10.1016/j.patcog.2021.108248

(5) Díez, J. Pérez-Núñez, P., Luaces, O., Remeseiro, B., Bahamonde, A.: “Towards explainable personalized recommendations by learning from users’ photos” Information Science, 520, pp.416-430, (2020)

(6) Wang, D., Mao, K., and Ng, G.: "Convolutional neural networks and multimodal fusion for text aided image classification" 20th International Conference on Information Fusion, pp. 1-7, (2017), doi:10.23919/ICIF.2017.8009768.

(7) Yao, Y., Yang, W., Huang, P., Wang, Q., Cai, Y., Tang, Z. : “Exploiting textual and visual features for image categorization”, Pattern Recognition Letters, 117, pp.140-145, (2019).

(8) Chen, T., Chen, Y., Guo, H., Luo, J.: "You Type a Few Words and We Do the Rest: Image Recommendation for Social Multimedia Posts" IEEE International Conference on Big Data, pp.2124-2133, (2018) doi:10.1109/BigData.2018.8622513.

(9) Radim Řehůřek: GENSIM: topic modeling for humans, https://radimrehurek.com/gensim/ (2022/4/12 アクセス)