5P-04

地域情報と類似度を用いた書籍推薦システムの開発

萱谷 勇太† 黒田 久泰‡

愛媛大学工学部情報工学科† 愛媛大学大学院理工学研究科‡

1. はじめに

近年，情報処理の技術の進歩はめざましく，SNS などを利用して，気軽に情報を発信できる環境が整っている．これに伴い，インターネット上の大量の情報は個人が処理できる範囲を超えており，ユーザー自身が取捨選択をしなければいけない．そのため，個人ごとに異なる趣味嗜好に合わせて収集する技術が今望まれていると考えられる．青空文庫[1]というサイトでは，多くの作品があるが，実際，個人の興味がある作品を探すのは困難である．本研究では，公開されている青空文庫の作品から地域情報などを抜き出すことで，ユーザーの興味が湧きそうな作品を推薦する．さらに，その作品と類似度の高い作品を提案するシステムを開発する．

2. システムの概要

本研究で開発したシステムの概要は以下のとおりである．

まず，ユーザーに地域情報による推薦を行う．その後，ユーザーが評価を行い，最高の評価をつけたものと類似度の高い書籍を推薦する．書籍の評価として，1～5 の数値で評価をしてもらい，5 を付けた場合を最高の評価とする．

3. 研究の概要

3.1 書籍データの取得

青空文庫では，収録されている作品を共有 Web サービスである GitHub に公開しており，毎日更新されているため，最新データ一式を容易に入手できる．

3.2 前処理

自然言語処理において，前処理を行うことは，精度を高める上で非常に重要な作業である．

本研究では，以下の前処理を行った．

- 書籍のいらない部分を除去
- ストップワードの除去

書籍のいらない部分というのは，テキストファイル中の注意書き・説明書きの部分のことである．

また，今回は，品詞が動詞・形容詞，名詞以外のものと，ストップワード辞書[2]に含まれているものをストップワードとして，除去を行った．

3.3 地域情報の抽出

本研究では，地域情報として，「都道府県」と「市区町村」を抽出した．

3.4 類似度の計算

3.4.1 形態素解析

日本語において，単語の境界は曖昧である．そのため，単語ごとに切り出す形態素解析が必要である．本研究では MeCab[3]を用いた．

3.4.2 Harris の分布仮説

「意味の似ている単語が，同じような使われ方をする」という傾向が Harris によって分布仮説として知られている[4]．

例えば，「飼っている○と散歩する」という文があった時，「イヌ」「ネコ」「チワワ」などが入る確率が高いため，これら 3 つの単語は意味的に似ているということである．この，分布仮説を元に単語ベクトルを表現する．

3.4.3 word2vec

word2vec[5]とは単語をベクトル表現するためのアルゴリズムである． 既存の手法と違い，word2vec では文脈情報を考慮するため，単語に対して，精度の高いベクトル表現を得ることができる．

以下に word2vec のアルゴリズムを示す．

まず，入力単語を表す One-hot ベクトルを入力層からの出力として，重み$\boldsymbol{w_1}$と掛け合わせる．

$$\boldsymbol{W_{xi}} = \begin{pmatrix} w_{1,1} & \cdots & w_{S,1} \\ \vdots & \ddots & \vdots \\ w_{1,N} & \cdots & w_{S,N} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} w_{i,1} \\ \vdots \\ w_{i,i} \\ \vdots \\ w_{i,N} \end{pmatrix} = \boldsymbol{w_i}$$

これにより中間層への入力として，単語ベクトルが算出される．次に入力層より出力された単語ベクトルから総和をとる．これにより，文脈ベクトルが算出される．

$$c = w_1 + \cdots + w_i$$

そして，中間層の出力(文脈ベクトル)に重み

Development of recommendation system using regional information and similarity

†Yuta KAYATANI
Computer Science, Faculty of Engineering, Ehime University
‡Hisayasu KURODA
Graduate School of Science and Engineering, Ehime University

$w_2(=w^t)$を掛け合わせる．これは文脈ベクトルと，単語ベクトルの転置行列の内積のことである．

$$W^t c = \begin{pmatrix} w_{1,1} & \cdots & w_{S,1} \\ \vdots & \ddots & \vdots \\ w_{1,N} & \cdots & w_{S,N} \end{pmatrix} \begin{pmatrix} v_{1+\cdots+i,1} \\ \vdots \\ v_{1+\cdots+i,2} \\ \vdots \\ v_{1+\cdots+i,N} \end{pmatrix} = \begin{pmatrix} \boldsymbol{w_1^t \cdot c} \\ \vdots \\ \boldsymbol{w_i^t \cdot c} \\ \vdots \\ \boldsymbol{w_S^t \cdot c} \end{pmatrix}$$

そしてこの値に対して Softmax 関数を適用することで，確率値p_iに変換する．この確率値p_iと正解値p'_iとの誤差を計算し，重みw_1，重みw_2を誤差逆伝搬法により更新する．最終的な重みw_1が単語ベクトルとなる．

3.4.4 Paragraph Vector

Paragraph Vector[6]とは文書をベクトルで表す手法である．

3.4.3 節で述べた word2vec の入力に文書 ID を付与することで，文書に対するベクトル表現を得ることができる．

本研究では，書籍の類似度を計算するアルゴリズムとして Paragraph Vector を使用した．Paragraph Vector で作成した学習モデルを使用し，「夏目漱石」の「吾輩は猫である」と類似度の高い書籍を 10 個算出した結果を以下の表 1 に示す．

3.5 Web システムとしての実装

本研究では，HTML,CSS,JavaScript,PHP を用いた Web システムを構築する．

Web システムを使用するうえで，類似度の算出を毎回行うのは非常に大きなオーバーヘッドとなってしまう．そのため，類似している書籍のデータをあらかじめ計算してテキストファイルに書き込んでおき，無駄な処理を行わないようにした．

図 2 は「愛媛」という文字を含む書籍のタイトルを表示した結果である．テーブルの列名をクリックすることで，列に対してソートをすることができる．

4. おわりに

地域情報と類似度を用いた書籍の推薦システムを開発した．これにより，興味のある書籍をユーザーが見つけることができるようになった．今後は，地域情報として「都道府県」と「市区町村」だけでなく，地域に関係する言葉（甲子園など）を抽出し，推薦を行うという機能なども考えられる．

表 1 「夏目漱石」の「吾輩は猫である」と類似度の高い書籍

夏目漱石	琴のそら音	0.7558
内田魯庵	二葉亭余談	0.7148
夏目漱石	倫敦消息	0.6984
高村光雲	幕末維新懐古談 17 猫と鼠のはなし	0.6855
夏目漱石	坊っちゃん	0.6823
夏目漱石	彼岸過迄	0.6689
夏目漱石	野分	0.6552
宮沢賢治	フランドン農学校の豚	0.6414
夏目漱石	文芸の哲学的基礎	0.6409
幸田露伴	鷲鳥	0.6400

図 2 Web システムの検索画面

参考文献

[1]青空文庫，入手先<http://www.aozora.gr.jp/>(参照 2018-01-08)
[2]田中克己:Slothlib Wiki FrontPage，入手先<http://www.dl.kuis.kyoto-u.ac.jp/slothlib/>(参照 2018-01-08)
[3]MeCab Yet Another Part-of-Speech and Morphonological Analyzer，入手先<http://taku910.github.io/mecab/>(参照 2018-01-08)
[4]Zellig S. Harris,"Distributional structure",Word,Vol.10,pp.146-162,1954.
[5]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In Proceedings of Workshop at ICLR, 2013.
[6]Quoc Le,Tomas Mikolov "Distributed Representations of Sentences and Documents",Proceedings of The 31st ICML,pp.1188-1196,2014