DEIM Forum 2020 P2-26(day2 p99)

個人の特性を反映した文章の類似度判定による小説推薦

丸山　正人†　　竹川　高志††

† 工学院大学工学研究科情報学専攻　〒 163–8677 東京都新宿区西新宿 1–24–2
†† 工学院大学情報学部システム数理学科　〒 163–8677 東京都新宿区西新宿 1–24–2
E-mail: †em19022@ns.kogakuin.ac.jp, ††takekawa@cc.kogakuin.ac.jp

あらまし　同じ小説の組み合わせでもその類似度は個人によって異なると考えられる．そこで本研究では，小説に出現するような単語の特徴量を得るモデルについて，モデル毎に学習を行う対象を任意の条件で抽出した小説のみに絞るという学習差を作ることによって，個人毎に異なる小説の類似度の推定を目指す．類似度の推定に用いる単語の特徴量を得る手法は，Word2vec で得られる単語ベクトルを用いる手法と，VBGMM で単語がクラスタに属する確率を用いる手法の 2 つを提案する．異なった学習を行ったモデル毎で同じ小説の組み合わせの類似度の比較を行うことにより，異なった学習を行ったモデルでも特徴量の差が小さい組み合わせと大きな組み合わせがあることを確認され，特定のモデルで特徴量の差が大きくなる時，小説の組み合わせがそのモデルが学習を行う条件について大きな差があることを確認された．

キーワード　推薦システム，小説推薦，分散表現，Word2vec，Variational Bayesian Gaussian Mixture Model，Linear Discriminant Analysis

1 はじめに

社会に発信される小説の数は日々増え続け，お金を払い読むことができる作品，お金を払わずとも Web サイトなどから読むことのできる作品両者が増えている．また，小説へのアクセス方法では，書店や図書館などから直接小説を手に取る方法だけでなく，EC(Electronic commerce) サイトから小説を手にする方法，電子書籍などの普及により，読み手がアクセスできる小説の量は増加している．この膨大な量の中から目的に合った小説を探すことは困難となり，この状態を解決するには推薦システムによる補助が解決策の 1 つと考えられる．

しかし推薦システムで小説を扱うには，小説について特徴量を求める必要があり，その特徴量の求める方法が推薦システムにとって重要な問題である．また，小説に対して感じる感想は個人によって異なり，その事に起因して，同じ小説の組み合わせでも個人によって似ている似ていないの判断が異なることがあり，全体で統一した小説の特徴量を求めることは問題があると考えられる．

そこで本研究では小説推薦システムを作成するために，仮想の個人を想定し，その個人が考える小説に対する類似度を求めることを目標とする．提案手法は，単語のベクトル化を行う Word2vec を用い，小説本文全てを使用した特徴量を求め，個人毎に好んだ小説のみを学習させるという学習差を作ることによって，個人毎に異なる小説に対する類似度を求める手法である．また，個人毎に異なる特徴量を得る手法は，Word2vec で異なる学習をさせる手法と，クラスタリング手法である Variational Bayesian Gaussian Mixture Model(以下 VBGMM) で異なる学習をさせる手法の 2 種類について行い，比較する．異なる学習によって得られる効果の検証は，異なった学習を行ったモデルそれぞれで同じ小説の組み合わせの類似度を求め，グラフにプロットを行い，モデル毎にその類似度の差の確認を行う．

結果は，異なる学習を行ったモデルそれぞれで同じ小説の類似度を求めると，学習を行う基準で特徴が大きく異なる作品が，類似度が下がることが確認でき，そのモデルのみで異なると判定される作品の組み合わせ，どのモデルでも同じぐらい異なると判定される作品の組み合わせがあることが確認された．また，Word2vec で異なる学習を行ったモデルと VBGMM で異なった学習を行ったモデルの比較を行うと，VBGMM の方がモデル毎の類似度の差が大きくなることが確認された．

本稿の構成は以下の通りである．2 章では関連研究について述べる．3 章では提案手法を述べ，4 章では結果と考察を述べる．5 章では結論と今後の課題を述べる．

2 関連研究

これまでに小説の推薦システムは読み手が特定の感情になるような本を抽出し推薦を行うようなシステムについての研究が行われている. [1] [2] [3] [4] また，推薦のために作品情報の自動抽出を行い，推薦文の自動生成を行う研究 [5]，著者の情報を利用し，関連した経歴を持つ著者の作品を推薦するシステムについての研究 [6] も存在する．しかし，読み手が小説に期待するもの，小説の選び方は多様に存在すると考えられ，既存の手法では個人毎に異なると考えられる小説に対しての類似度を反映した推薦が困難であると考えられる．

本研究では，個人が考える小説の特徴を個人毎にそれぞれ Word2vec などのモデルに学習させ，そのモデル上での小説の特徴を用い，個人毎に持つ小説の類似度を用いた小説の推薦システムの作成を目指す．

3 提案手法

本研究では，Mikolov らによって研究された Word2vec [7] を用いる．Word2vec は，同じ文脈に出現する単語は似た意味を持つ傾向があるという分布仮説 [8] に基づいて，単語を多次元ベクトル化した，単語の分散表現を得る手法であり，この手法によって得られる単語のベクトルは，似た意味を持つような単語が近く分布するようになります．これにより小説の本文に登場する単語を数値化し，計算機で扱うことが可能となる．

また，Word2vec によって得られた単語のベクトルについて，小説に出現する単語のベクトルを 1 つの群とみなすことによって，2 つの小説が持つ群の分離度を求め，それを 2 つの小説がどのくらい似てないと感じるか表す度合いとする．分離度の求め方は，Linear Discriminant Analysis(以下 LDA) [9]，群内分散，群間分散を用いる．LDA は教師ありのクラス分類手法であり，クラス分類の際に同じ群に属するデータの分散 (群内分散) を小さく，群同士の分散 (群間分散) を大きくするように次元削減を行う．その次元削減によって得られたデータについて，群間分散を群内分散で割った値をその 2 つの作品の分離度とする．

得られた類似度について，異なる学習を行ったモデルが異なる類似度を得たかを検証するため，異なる小説の選び方をする仮想の個人を想定し，それぞれでモデルの学習を行い，同じ小説の組み合わせで異なる類似度が得られているかを確認する．

個人毎に異なる小説の特徴を得る手法は章 3.1，3.2 で検証を行う．小説の作品データについては，青空文庫を用いる．また，Word2vec の学習の際に行う日本語文の形態素解析には，辞書に mecab-ipadic-NEologd[1] を使用した，MeCab [10] を用いた．今回モデルは 1981,1982 年の作品を学習したモデル, [11] [12] の研究を用いて，意見などの評価表現を抽出し，文の評価極性のスコアを計算する Python ライブラリである oseti[2] によって，作品の評価極性がが 0.49 以上のポジティブな作品のみを学習したモデルを作成した．

3.1 Word2vec で作品の特徴量を求める

Word2vec で作品の特徴量を求める場合の手順は図 1 となる．まず，青空文庫から条件を設けて抽出した作品を学習した 2 つの Word2vec モデルが単語ベクトルが 100 次元になるよう指定して学習を行う．次にモデル毎に作品の類似度を求めるために，青空文庫の作品から無作為に 1 作品を抽出し，もう 1 作品を先に抽出した作品と単語数が大きく異ならない範囲で無作為に抽出する．これは，後に得られる単語ベクトルに対して行う LDA が分類を行う群に対して等分散を仮定しているためである．次に，取り出した作品に存在する単語についてそれぞれのモデルでベクトル化するが，この時 Word2vec が同じ文脈で用いられる単語のベクトルを近くなるように学習を行うため，どの文章でも出現するような単語がノイズになるので，固有名詞の単語のみを抽出し，ベクトル化を行う．次に，得られた単語ベクトルについて，取り出した作品をクラスとして LDA を行い，1 次元へと次元削減を行われた単語ベクトルを取得する．得られた 2 作品の 1 次元ベクトルについて分離度を求める．作品の抽出から分離度を得るところまでを繰り返し，十分な数の作品間の分離度を求めたところで，同じ作品についてそれぞれのモデルの分離度をグラフに表して確認を行う．

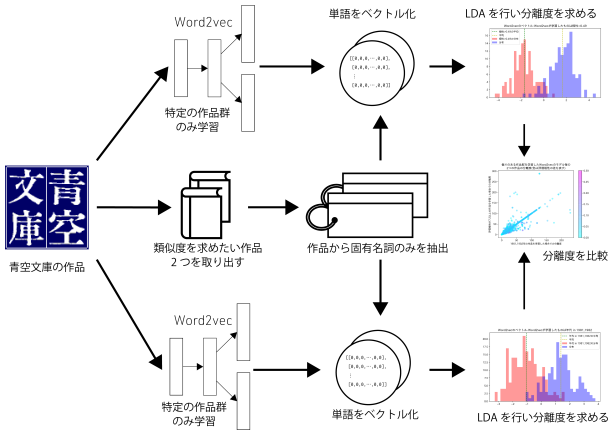

図 1　実験手順の流れ

3.2 VBGMM で作品の特徴量を求める

VBGMM で作品の特徴量を求める場合の手順は図 2 となる．まず青空文庫の作品全てを Word2vec にて単語ベクトルが 100 次元となるように指定して学習を行う．青空文庫の作品について固有名詞のみ単語ベクトルを求める．その単語ベクトルについて，条件を設けて抽出された作品のみを VBGMM が学習を行う．比較を行いたい 2 作品のみの単語ベクトルを抽出し，学習を行った VBGMM で単語ベクトルからクラスタに属する確率に変換を行う．得られたクラスタに属する確率について，取り出した作品をクラスとして LDA を行い，1 次元へと次元削減を行われたクラスタに属する確率を取得する．得られた 2 作品のクラスタに属する確率について分離度を求める．作品の抽出から分離度を得るところまでを繰り返し，十分な数の作品間の分離度を求めたところで，同じ作品についてそれぞれのモデルの分離度をグラフに表して確認を行う．

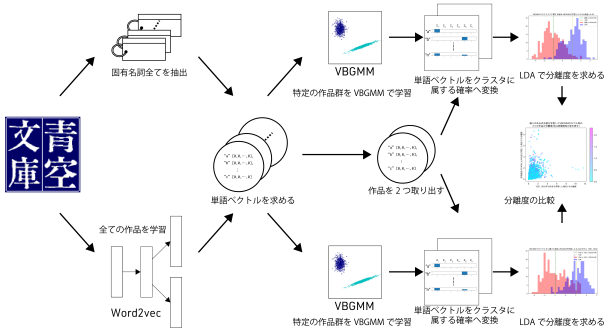

図 2　実験手順の流れ

1 : https://github.com/neologd/mecab-ipadic-neologd (2020/3/15 参照)

2 : https://github.com/ikegami-yukino/oseti (2020/3/19 参照)

4　結果と考察

図 3 は同じ作品の組み合わせについて 2 つの Word2vec モデルで求めたそれぞれの群間分散をプロットした図である．x 軸は 1981,1982 年の作品を学習したモデルでの分離度であり，y 軸は文章の評価極性が 0.49 以上の作品のみを学習したモデルでの分離度である．一方のモデルで群間分散が大きく，もう一方のモデルで群間分散が小さい作品の組み合わせがあることがわかる．これは一方のモデルでは作品が類似していると認識され，もう一方のモデルでは類似していないと認識されていることになる．つまり異なる学習によって，異なった作品の特徴を得ることができ，作品が類似しているかの基準も異なることが言える．

次に図 4 を見る．この図の軸が表すものも図 3 と同じである．VBGMM を用いた場合は Word2vec を用いた場合よりプロットが散らばっている事が確認できる．これは，Word2vec を用いた場合より VBGMM を用いた方がモデル毎の分離度の差が大きくなっていると言える．つまり，Word2vec を用いた手法より VBGMM を用いた手法の方が，異なる学習をさせた時に生じる類似度の差が大きくなり，どのモデルでも同じような小説が推薦されると言う状態が起きにくいと考えられる．

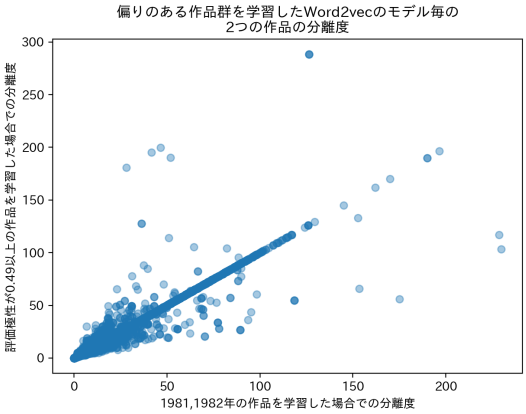

図 3　児童文学作家，童話作家の作品を学習した Word2vec モデルで LDA を行い 1 次元へ次元圧縮を行われた単語ベクトルの分布

5　ま　と　め

本論文では，個人毎に異なる小説に対しての類似度を反映した推薦システムを作成するために，異なる小説の学習を行ったモデルを用いて小説の類似度を求めた．Word2vec が単語をベクトル化することを用い，小説本文の学習を行い小説本文の特徴を求めた．また，モデルが小説本文を学習する際に個人が好んだ作品のみを学習するという環境を作ることによって，モデル毎によって異なる類似度が得られる事がわかった．個人差を出すモデルの学習には Word2vec を用いた手法と VBGMM を用いた手法 2 種類について実験を行ったが，Word2vec を用いた手法より VBGMM を用いた手法の方がモデル毎の学習差が大きく反映される事が確認された．今後の課題としては，他の

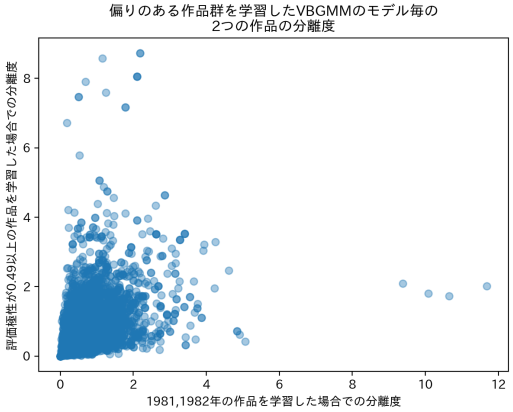

図 4　児童文学作家，童話作家の作品を学習した Word2vec モデルで LDA を行い 1 次元へ次元圧縮を行われた単語ベクトルの分布

基準を持った個人を想定しても正しく分離されるか，様々なパターンで検証を重ねる事を考えている．また，小説の類似度が妥当であるか調査を行い検討を重ね，具体的に人に対して推薦を行い，仮想の条件でなくとも正しく好んでいる作品の特徴を捉えられているか調査を行う必要もある．また類似度 1 つ 1 つの要素について，具体的な作品の特徴を調べ正しく意味を取り出せているか確認する必要もある．

文　　献

[1] 原田隆史. 感性パラメータを用いた類似する小説の提示. 情報知識学会誌, Vol. 21, No. 2, pp. 291–296, 2011.

[2] 純太増田, 徹杉本. E-015 小説推薦システムの構築に向けた検索表現と書評の分析 (検索・質問応答・抽出,e 分野:自然言語・音声・音楽). 情報科学技術フォーラム講演論文集, 第 11 巻, pp. 189–190, sep 2012.

[3] 泰広山本. レビュー及び周辺情報を用いた小説の感性情報の推定と小説検索システムの構築, 2016.

[4] 平良浩嗣, 當間愛晃. 感情推定に基づくコンテンツ推薦システムに向けた小説文感情調査. 第 76 回全国大会講演論文集, 第 2014 巻, pp. 143–144, mar 2014.

[5] 秦野智博. 利用者の評価基準に合致した文章推薦システムの構築. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 4J15–4J15, 2016.

[6] 山副睦実, 児玉英一郎, 王家宏, 高田豊雄. Linked data を用いた著者関連情報による小説推薦システムに関する考察. 電気関係学会東北支部連合大会講演論文集, Vol. 2014, pp. 251–251, 2014.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[8] Zellig Harris. Distributional structure. *Word*, Vol. 10, No. 23, pp. 146–162, 1954.

[9] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Vol. 7, No. 7, pp. 179–188, 1936.

[10] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[11] 東山昌彦, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 述語の選択

選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会, 2008, 2008.

[12] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.