人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第23回)
SIG-AM-23-01

隠れた良作を推薦可能な Web 小説レコメンドシステムの提案

Recommending Hidden Masterpieces: Towards an Attentive Web Novel Recommender System

小坂 直輝[1*] 小林 哲則[1] 林 良彦[1]
Naoki Kosaka[1] Tetsunori Kobayashi[1] Yoshihiko Hayashi[1]

[1] 早稲田大学 理工学術院
[1] School of Science and Engineering, Waseda University

Abstract: Websites accommodating user-generated content have been becoming popular among the general users who would like to share and enjoy the published works. An effective and efficient search/recommender system is however required to adequately link the publishing users and the consumer users. In the present study, we target a Website that provides an arena for sharing Japanese novels authored by general users and propose a system for making a personalized recommendation even for yet to popularized works based on the quality of the sampled sentences. This may contribute to discovering "hidden masterpieces." The proposed system may also be able to uncover works that might match a user's preference by measuring the similarity to already-read works. This paper specifically presents machine learning-based frameworks for estimating sentence qualities and predicting inter-works similarities. The preliminary experimental results suggest that the proposed frameworks are promising in implementing a recommender system that may satisfy the desiderata.

1 はじめに

1.1 研究背景と目的

近年，インターネット上で個人が小説を投稿できる，小説投稿サイトが多数開設され，数多くの作品が投稿されている．こうしたサイトに投稿される作品は一般に web 小説やオンライン小説と呼ばれ，これらの作品の中には商業用に書籍化・映像化される作品も存在している．

一方でこれら web 小説に関して，その作品数の多さから読者が好みの作品を探すのが難しいという問題や，人気の作品や最新の作品といった限られた少数の作品に大多数の読者が集中するという問題がある．

こうした問題は特に，有名な作品は一度は見かけたことがある，一定以上の利用経験がある読者にとっては重大であり，そういった読者が大量にあるマイナーな作品から，次に読むための良い作品を探そうとすると多大な労力を要求される．結果として誰にも読まれない良い作品，いわゆる隠れた良作ができてしまう．

書き手の立場からすればこれは，作品を読んでもらうことの困難さに繋がり，良い作品を書いてもその作品が読者の目に入らない可能性があると言える．さらにこうした問題は読者，筆者両方について小説投稿サイトの利用を止めたり，ひいては web 小説そのものから離れる要因になることも考えられるので，解決する必要がある．

本研究の目的は，上記問題点を解決し，ユーザが作品を探すのを助けるレコメンドシステムを実現することである．具体的には，日本最大級の小説投稿サイトであり，約 70 万の作品と 160 万人のユーザ数を有する「小説家になろう」[*1]の作品を対象に，作品の本文とジャンルやキーワードといった付属情報から，作品の類似度や質を推定するモデルを構築することで，読者が付与した情報の無い隠れた良作の推薦や，ユーザによる推薦基準の操作が可能なレコメンドシステムを提案する．

特に本稿では，類似度や質を定める有効な情報について検討し，機械学習を用いた推定実験の結果を議論する．

* 連絡先：早稲田大学理工学術院
〒 162-0042 東京都新宿区早稲田町 27 早稲田大学 40 号館 701 号室
E-mail: kosaka@pcl.cs.waseda.ac.jp

[*1] https://syosetu.com/

1.2 関連研究

一般にアイテムベース協調フィルタリングなどの代表的な推薦アルゴリズムは，新規アイテムはユーザの評価履歴が無いため推薦できない，コールドスタート問題を抱えている [1][2][3]．本研究における隠れた良作とは評価がついていれば高い評価となるであろうが，評価が無いためにユーザの目に入らない作品であり，この点でコールドスタート問題と関連がある．コールドスタート問題への対処法として，アイテム自身の特徴を利用する内容ベースの推薦アルゴリズムならユーザの評価履歴が無くとも推薦が可能であり [1][2][3]，本研究でも web 小説のレコメンドにおいて作品自体の情報を利用する．

また，文章を何らかの側面から評価するという試みについても様々な研究がある．例えば日本語の文章の読みやすさを評価する試みとして，[4] では，文の長さや各文字種の頻度から読みやすさを求める評価式が提案されていて，[5] では文章の難易度の基準となるコーパスを定め難易度ごとの言語モデルを作製し，文章の難易度を判定する手法が提案されている．本研究は読みやすさに限らない総合的な文章の質を，機械学習を用いて，様々な作品情報から予測する．

さらに [6] ではトピックモデルや，センチメント分析によるプロットライン判定，文章の文体の分析によりベストセラーを予測するという試みが行われている．高い評価を得る作品には潜在的に共通するパターンが存在すると考え，そういった作品をコンピュータに予測させるという点で本研究と関連がある．

web 小説のレコメンドに関連する研究としては，お気に入り登録のリンク構造を用いる [7] や，作品のあらすじを用いる [8]，文体の類似度を考慮する [9] があるが，本研究は作品の本文も用いるという点で [7][8] と，類似度以外の観点も考慮するという点で [9] と差異がある．また，評価者とシステム利用者の評価基準の不一致に着目した研究 [10] も存在する．

2 提案システム

2.1 読者が小説を選び読む際の観点

提案するシステムはレコメンドシステムであるので，当然ながら提示する作品はユーザが読みたいと思う作品や読んで良かったと感じる作品であるのが理想的である．そこでまず読者が小説を読むか決めるポイントや，読んで良かった，面白かったと感じる要因について考え，表 1に示すように主に三つの観点に分けてまとめた．

一つ目は小説の種類で，その小説がどのジャンルのどういった形式で書かれたものかといった観点である．この観点は，例えば SF が好きで SF しか読まない読者がいたり，女性が女性向け作品を読んだりというように，小説を選ぶ際に大きく影響し，また，好みのジャンルの方が読んで良かったと感じやすいといったように読書中・読後の感じ方にも影響すると考えた．この 1 の観点は，好みなどによる個人差が大きく，一方で機械による判断は比較的容易だと考えた．

二つ目は，ストーリーやキャラクター，世界観といった小説の内容であり，一般に小説の面白さといった部分に大きく関わるであろう観点である．この 2 の観点は，1 の観点程ではないにせよ好みよる個人差が大きく，一方でこの観点においてどれだけ優れているかを機械によって判断するのは非常に難しいと考えた．

三つ目は，読みやすさや表現の豊富さといった小説の文章に関わる観点であり，これも読書中・読後の感じ方に影響すると考えた．特に web 小説においては，編集者の手が入り一定以上の文章の質が保証される商業書籍と異なり，読みやすさや文法の誤り，誤字脱字といった部分でも，優れた作品からそうでないものまで幅があるので，この 3 の観点も重要である．そしてこの観点は，好みなどによる個人差は比較的小さく，機械による判断もある程度可能なのではと考えた．

表1　読者が小説を選び読む際の観点

1	**小説の種類**	（ジャンル，　形式， 男性向け女性向け　等）
	好みによる個人差：大	機械による判断：易
2	**小説の内容**	（ストーリー，　キャラクター， 設定，　世界観　等）
	好みによる個人差：中～大	機械による判断：難
3	**小説の文章**	（読みやすさ，　表現の豊富さ， 描写の丁寧さ，　文法　等）
	好みによる個人差：小	機械による判断：普～難

本研究で提案するレコメンドシステムでは推薦時に S スコアと Q スコアという二つの指標を用いるが，前者で主に 1 の観点に，後者で主に 3 の観点に対応することで，ユーザが読みたいと思うような作品や読んで良かったと感じる作品を推薦する．各指標については後の 2.3 節や 4，5 節で詳しく述べる．また提案するシステムにおいて，2 の観点に関しては機械による判断の難しさなどから現段階で直接的に対応する機能は考えていない．

2.2 現状の検索システムの問題点

レコメンドシステムを提案するにあたり，小説投稿サイトにおける現状の作品検索システムがどういったもの

であるかその問題点と共に説明する．小説投稿サイトにおける作品検索システムの概要図を図 1に示す．ユーザが作品を探す流れは，まずジャンルやキーワードといった検索条件を指定し，次に検索結果を表示する順序を指定すると，作品が指定した順に並べて複数ページに渡り表示されるといった流れである．

ここで，この作品の並べ方には複数の選択肢があるが，基本的には人気の作品や最近の作品が前に表示される．そのため小数の人気の作品がユーザの目に入りやすく，そういった作品に読者が偏るということが起こる．これはとりあえず人気の作品や最近の作品が読みたいというような，ライトユーザには適していて大きな問題はない．一方で，多くの作品を読んでいるユーザが，後ろに表示されている大量のマイナーな作品から良い作品を探そうとすると多大な労力を要求されるので，そういったユーザにとっては問題であり，より適した別の作品提示方法も必要だと考えられる．

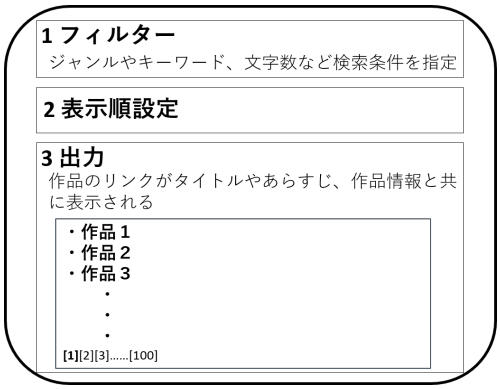

図1　現状の検索システムの概要図

また，別の問題として作品を並べて表示する際の人気度を示す指標が累積値のようなもので，作品を比較する指標として適切でないという点がある．例えば「小説家になろう」[*1]において作品を人気順に表示するために用いられている指標の一つに総合評価点があるが，これはユーザが着けた作品の評価やブックマークの合計値から計算した値であり，一桁の作品から数十万点の作品まで存在する，上限のない累積値である．累積値では 10 人が 10 点と評価した作品よりも 100 人が 5 点と評価した作品の方が値が大きいといったことが起こりえるので，作品を比較する指標として適切とは言えない．

小説投稿サイトには「小説家になろう」の他にも「カクヨム」[*2]や「エブリスタ」[*3]など多くのものがあるが，作品検索システムの形は類似していて，上記問題点は複数の小説投稿サイトにおいてある程度共通する問題でる．

[*2] https://kakuyomu.jp/
[*3] https://estar.jp/

2.3　レコメンドシステムの設計方針

本研究の目的や既存の検索システムの問題点を踏まえ，提案するレコメンドシステムの要件を次の二つにまとめた．一つ目が読者が少ない，いわゆるマイナーな作品もレコメンドの提示候補となりうること，二つ目がレコメンドにおいて何を重視するかという，レコメンドの基準をユーザが操作できることである．

以上二つの要件を満たすため，提案するレコメンドシステムでは、S スコアと Q スコアという二つの指標を用いる．S スコアはある作品が他の作品とどれだけ似ているかを示す指標で，これによりユーザの過去の読書傾向に似た作品を提示する．一方 Q スコアは，主に表 1の小説の文章の観点において，作品の質がどれだけ良いかを示すもので．これによりユーザがより満足できる作品を提示する．

ここで，上記のように作品の類似度や質を示すものとして定めた二つの指標，S スコアや Q スコアについて，その値をどう求めるかが重要である．単純な方法として，作品の類似度や質を読者の一致率や評価の高さと見なし，読者が付与したデータから求める方法が考えられるが，それでは，読者が付与したデータの少ないマイナーな作品には各指標が適切に求められない可能性がある．そこで S スコアと Q スコアのそれぞれにおいて，小説本文も含めた作品自体の情報から各指標を予測するモデルを構築し，これを用いることで読者が付与したデータのないマイナーな作品についてもレコメンドを可能にする．

また，最終的なレコメンドにおいては，S スコアと Q スコア，二つの指標を合わせた値の高さで作品の提示を行うが，その際，各指標の重みを調整可能にし，レコメンドの基準をユーザが操作できるようにする．例えば S スコアの重みを大きくすれば，提示する作品がより過去の読書傾向との類似性を重視したものになる．この仕組から，より各ユーザに適した形で，ユーザが作品を探すのを助けるシステムになると考えた．

3　データ

本研究では「小説家になろう」[*1]から API やスクレイピングを用いて取得した小説のデータを利用する．まず，本研究で取得するデータに含まれる情報を，その種類により分類したものを表 2に示す．表 2における読者が付与した情報は作品ごとにその有無や量が異なるが，作品自体の情報は基本的に全ての作品に共通して存在する．

レコメンドに用いる S スコアと Q スコアという二つの指標をモデルから予測させる際は，表 2における作品自体の情報を入力として用いることで，マイナーな作品についても対応する．

表2　取得データに含まれる情報の分類

読者が付与した情報		評価点，感想，レビュー　等
作品自体の情報		
	作者による付属情報	ジャンル，文字数　等
	作品の中身	本文，文章の品詞傾向　等

以下具体的なデータの取得状況を説明する．まず，「小説家になろう」で公開されている全作品を対象に，提供されている API を用いて取得できる 33 項目の作品情報を取得した．ここにはタイトルやジャンル，文字数やユーザの評価点などが含まれ，670,834 作品分の作品情報を取得した．このうち，368,576 作品が長編小説，長編小説の 85,649 作品が完結済であった．ここで長編小説とは，作者が小説を一部分ずつ投稿，更新していくタイプの作品であり，「小説家になろう」の作品はこの長編小説と，短い作品を一度に投稿する短編小説の二種類に大別される．本研究では，ユーザの需要やデータ数の観点から，5 万字以上の長編小説もしくは 1 万字以上で完結済の条件を満たす 116,529 作品を扱うこととした．

次に，上記 116,529 作品を対象に小説の本文を取得した．具体的にはスクレイピングにより，小説のランダムな三か所から 1,200 文字弱ずつの文章を取得した．このような取得の仕方をしたのは，全対象作品の全文章を取得するのはデータの規模から容易ではない一方，作品の先頭など特定の一か所だけから取得すると，その作品全体の文章傾向とずれ偏ったものとなる恐れがあるからである．

最後に，同じく上記条件を満たす長編小説を対象に，作品の文章評価とストーリー評価を取得した．「小説家になろう」には読者が作品の文章とストーリーを各 1〜5 点で評価できる仕組みが存在する．API から取得できるのは全評価点の合計のみで，文章評価とストーリー評価の割合は考慮出来ないが，作品ページにはそれぞれの評価の合計が公開されているのでこれをスクレイピングにより取得した．

データの取得には時間がかかり，その間に削除された作品や何らかのトラブルで取得が出来なかった作品もあるため，実際に取得したのは本文が 114,233 作品分，文章評価とストーリー評価が 112,851 作品分である．

その他の取得は行っていないが利用を検討しているデータとして，各作品に書かれている感想や，ユーザの評価履歴といったデータがある．

4　作品の質予測

ここからはレコメンドに用いる指標であり，作品の類似度や質を示すものとして定めた二つの指標，S スコアと Q スコアについて作品自体の情報から予測が可能かを調べるため行った，分析や実験の内容と結果について述べる．まずこの 4 節では作品の文章の質などを示す Q スコアについて触れる．

4.1　Q スコアの設定

Q スコアを定めるにあたり，この指標は作品の質の高さなどを示すもので，かつ作品の比較を行うのに適した指標であることが望ましい．そこで今回，式 1で求められる指標を Q スコアとして設定した．

$$16 \times \frac{\text{累積文章評価}}{\text{評価人数}} + 10 \times \frac{\log_2(1 + \text{レビュー数})}{\sqrt[4]{\text{評価人数}}} \quad (1)$$

取得したデータの中で，Q スコアを求める参考になる情報に文章評価があり，これは累積値であるが，取得したデータには評価人数も含まれているので，累積文章評価を評価人数で割ることで，比較が可能な 5 点を最大とする平均文章評価点を求められる．また，レビューというのは「小説家になろう」[*1]においては読者が作品を推薦する目的で書かれるもので，多いほど良い作品と判断できる．これらの観点から今回式 1で求められる値を Q スコアとして設定した．なお，対数や 4 乗根，定数倍はスケール調整のためのものであり，Q スコアの値の約 8 割が平均文章評価点により定まるようになっている．これはレビューが書かれていない作品も多くあることを考慮した結果である．

文章評価を取得した作品の内，評価人数が 10 人以上の作品に対し，上式 (1) を用いて Q スコアを計算した結果のヒストグラムを図 2に示す．縦軸が作品数，横軸が Q スコアで，最大値が 97.8，最小値が 35.2 と一定の範囲に収り，正規分布に近い分布となっているので，作品を比較する指標としてある程度適切だと考えた．

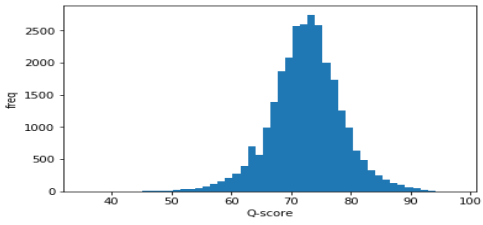

図2　Q スコアの分布

4.2 Q スコア予測実験

4.2.1 事前分析

まず，事前の分析として，取得できるデータに含まれる，作品の総文字数，会話率，完結済かの三つの特徴量について Q スコアとの相関を調べたが，明確な傾向は確認出来なかった．例として文字数と Q スコアの散布図を図 3に示す．

文字数と会話率のどちらにおいても Q スコアとの明確な相関は見られず，また完結済かにおいても，微かに完結済の作品の方が平均 Q スコアが高かったが有意な差は確認できなかった．

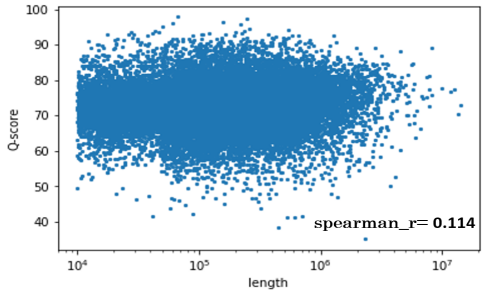

図3 文字数と Q スコアの相関

4.2.2 機械学習モデルによる予測

まず，事前分析で用いた三つの特徴量について，これを合わせて用いることで予測ができないかと考え重回帰分析とシンプルな Neural Network モデルによる学習と予測を行ったが，精度は悪かった．(結果は次の 4.2.3 節に示す)

そこで，使用する作品の特徴量を増やし改めて Neural Network モデルを構築し，学習・評価を行った．構築したモデルの構造を図 4に示す．

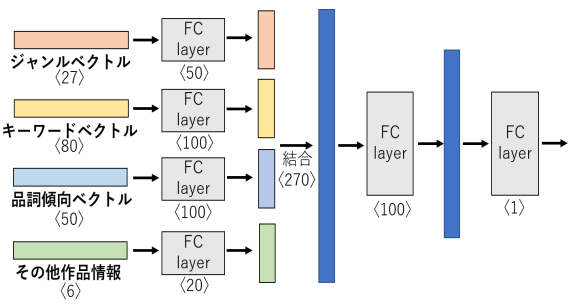

図4 Q スコア予測モデル

入力ベクトルと含まれる特徴について，まず，ジャンルベクトルは「小説家になろう」におけるジャンル分類において，作品がどのジャンルに含まれているかを表している．キーワードベクトルは作者が作品に設定するキーワードの内，今回用いたデータセットにおいて出現回数が一定数以上だったキーワードにおいて，各キーワードが含まれているかを表したバイナリベクトルである．品詞傾向ベクトルは作品の本文を形態素解析し，各品詞の出現頻度を表したベクトルである．形態素解析には日本語自然言語処理ライブラリの GiNZA[*4]を用いた．その他の作品情報としては，事前の分析で用いた三つの特徴量に加え，タイトルの長さ，漢字の割合，文密度 (文字数/行数) を用いた．

また，ここまでは Q スコア予測モデルの入力として用いる作品のベクトルを取得したデータから直接生成していたが，文書のベクトル化手法である Doc2Vec[11] を用いて，作品の本文から教師なし学習を行い生成した作品のベクトルを，入力として用いる場合についても実験を行った．Doc2Vec を用いるにあたっては，GiNZA を用いて分かち書きした単語全てを用いる場合と，助詞や助動詞といった一部品詞を除いた場合，[11] で紹介されている二つの手法である dmpv と DBoW のそれぞれを用いた場合，[12] で紹介されている最適パラメータを参考に大まかにパラメータを変化させた場合など様々な組み合わせを試した．

また，図 4における入力ベクトルと Doc2Vec によって生成した作品ベクトルを合わせて Neural Network の入力として用いる場合や，それぞれを単独に用いたモデルの出力をアンサンブルした場合についても実験を行った．

4.2.3 実験結果と考察

今回行った機械学習モデルによる Q スコア予測実験の結果を，学習において最小化を目指した平均二乗誤差と，モデルの精度を考える目安となるスピアマンの相関係数と共に表 3に示す．Q スコアの値の大小が正しく予測出来ていることが望ましいので，スピアマンの相関係数が 1 に近いほど良いモデルだと言える．表 3における MLP (Doc2Vec のみ) と記したものが Doc2Vec による作品ベクトルから Neural Network モデルによる予測を行ったもので，MLP (Doc2Vec 併用) が図 4の入力ベクトルと Doc2Vec による作品ベクトルを共に Neural Network の入力としたもの，MLP (アンサンブル) がそれぞれ単独に用いた予測値を 7:3 で重み付き平均した場合である．

なお Doc2Vec については複数試した組み合わせの中で最も精度が良かった，前処理で一部品詞を除き，dmpv

[*4] https://megagonlabs.github.io/ginza/

のモデルで 35 エポック学習させたものを用いた結果について掲載している．実験では評価人数が 10 人以上の作品で本文の取得が完了している 26,859 作品を利用し，そのうち 8 割を訓練データ，訓練データの 1 割を validation 用として利用した．

表3　Q スコア予測精度

モデル	特徴の種類数	誤差	相関
重回帰分析	3	34.35	0.126
MLP	3	34.29	0.143
MLP	9	27.02	0.468
MLP (Doc2Vec のみ)	1	30.64	0.363
MLP (Doc2Vec 併用)	10	28.27	0.436
MLP (アンサンブル)	10	26.43	0.487
訓練データの平均を出力		34.66	−

事前の分析で用いた，文字数，会話率，完結済かの三つの特徴量だけでは，Q スコアの予測精度は，訓練データにおける Q スコアの平均を出力した場合と大差ない．一方で，ジャンル，キーワード，品詞傾向，タイトルの長さ，漢字の割合と文密度というさらに六種類の特徴を用いたモデルでは，予測精度に改善が見られた．また Doc2Vec による作品ベクトルのみを用いた場合でも多少は予測が行えているが，その精度は図 4の九種類の特徴を用いたモデルより悪く，九種類の特徴と併用した場合でも同様だった．だが，九種類の特徴と Doc2Vec による作品ベクトルを個別に用いた予測値をアンサンブルした場合は精度が上がっているため，パラメータ調整やモデルの構造の検討が十分でない可能性も考えられる．

次に正解値と予測値の関係を確認するため，最も精度が良かった場合における正解値と予測値の散布図を図 5に示す．

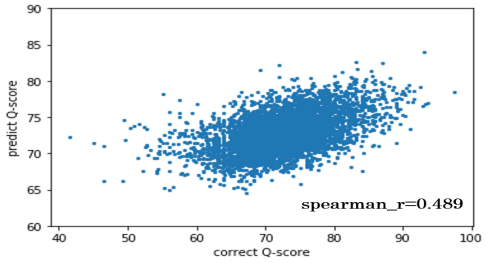

図5　正解値と予測値の相関

図 5と表 3の結果から正解値と予測値にある程度相関があることが確認できた．また図 5から，正解値に対し予測値の分散が小さく，学習データが少ない Q スコアが 65 以下や 85 以上などの作品に対し正確に予測できず予測誤差が下がらなかったのだと考える．今後精度を改善する方法として，一つにはさらに別の特徴を用いることが考えられる．今回利用出来なかった特徴として，様々な自然言語処理タスクで成果を上げている BERT[13] を用いて求めた文章の分散表現があり，これを用いることで予測精度が向上する可能性がある．また，各特徴量の予測値に対する影響度を調べたり，それに応じてモデルの構造やパラメータを見直すことでも，精度を向上させられると考える．

さらに，今回設定した Q スコア自体についても，ユーザの評価履歴や感想といったデータの取得後，改めて適切な指標設定を検討する余地がある．

5　作品の類似度予測

この 5 節では作品の類似度を示す S スコアについてその求め方や，妥当性を確認するために行った実験について述べる．

5.1　S スコアの設定

S スコアを求めるにあたり，作品の類似度の人手による正解データは存在せず，類似度の正解データを新たに作成するのも困難である．一般にアイテムベースの協調フィルタリングや，二つの文の類似度を求めるタスクである Semantic Textual Similarity では，それぞれのアイテムや文の多次元ベクトルによる表現を求め，そこからコサイン類似度やユークリッド距離，Jaccard 係数などを計算することで類似度を求めるという手法が多く使われる．同様に作品の多次元ベクトルを求める方法として，同じユーザが同じ評価をしている作品は似ているという観点から，ユーザの評価履歴を用いる方法が考えられるが，「小説家になろう」[*1]のユーザ数は 160 万人に上り，その一部にしても取得には膨大な時間がかかるためこの方法は現実的でない．またデータの偏りやスパース性といった問題もある．

そこで今回，4 節でも用いた Doc2Vec による作品ベクトルを利用し，この作品ベクトル同士のコサイン類似度を求めこれを S スコアとして利用することとした．「小説家になろう」において作品ジャンルやキーワードは作者が自由に設定できるが，こうして求めた S スコアにより，設定されたジャンルやキーワードは異なるが実際は似ている作品も探せると考えている．

5.2　S スコア予測実験の結果と考察

S スコアに関する実験においては本文の取得が完了している作品のうち，ジャンルがノンジャンルではない

87,237 作品を用いた．ここでノンジャンルの作品とは「小説家になろう」で 2016 年に行われたジャンル分類の変更後にジャンルが再設定されていない作品である．ノンジャンルの作品を除いたのはこの後で述べる S スコアの予測実験や結果の分析で作品のジャンルデータを利用するためである．

S スコア自体は Doc2Vec により生成した作品ベクトルのコサイン類似度を計算することで簡単に求められるが，作品の類似度の正解データが存在しないため，求めた S スコアがどの程度適切なものか直接的には評価が出来ない．そこで今回は作品のクラスタリングを行った結果の確認や，同ジャンル間と別ジャンル間での差異の確認，S スコアを元に同一作品の判定を行う実験などを通して，ある程度の妥当性を確認した．

まず，Doc2Vec により生成した作品ベクトルから Ward 法による作品のクラスタリングを行った結果のデンドログラムを図 6に示す．ここで S スコアの妥当性を確認するという観点から，S スコアが高い作品同士が近いクラスタとなるよう，各作品ベクトルを正規化してからクラスタリングを行った．Ward 法ではクラスタ間の距離を求める過程でユークリッド距離を利用するが，正規化を行うことでユークリッド距離が小さい作品と cos 類似度が高い作品が対応するためである．

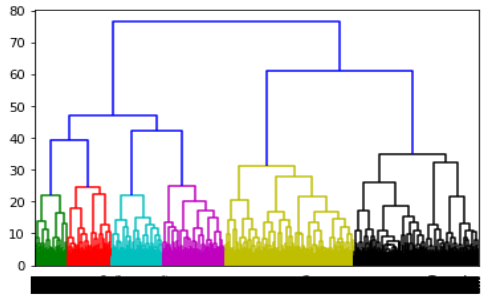

図6　Ward 法による作品のクラスタリング

また図 6だけではクラスタリングが適切に行われているか判断できないため tf をクラスタ内単語出現頻度，idf を逆作品頻度として tf-idf を求め各クラスタの特徴語抽出を行った結果を図 7に示す．加えて「小説家になろう」における大ジャンル五つにおいて，各ジャンルがどのクラスタに多く含まれているか求めた結果を表 4に示す．ここから同一クラスタ内や隣接するクラスタに含まれる作品にある程度類似性があることが確認できた．

例えば図 7からクラスタ 2 は「情報」や「通信」といった科学と関連するような特徴語が，クラスタ 5 には「好き」や「一緒」といった恋愛と関連するような特徴語が含まれていて，実際に表 4からクラスタ 2 の 21% が SF ジャンル，クラスタ 5 の 30% が恋愛ジャンルであることが確認できる．また図 6と図 7，表 4から合わせて判断することでクラスタ 1 から 4 にファンタジー，クラスタ 5 と 6 に恋愛系の作品が多く集まっていると判断でき，これはこの二つのジャンルが「小説家になろう」において人気で作品数が多いことを考慮すると妥当な結果だと考えられる．

図7　各クラスタの特徴語

表4　主要ジャンルとそのジャンルを含む割合の多いクラスタ

恋愛	クラスタ 5: 30%	クラスタ 6: 29%
ファンタジー	クラスタ 3: 68%	クラスタ 1: 64%
文芸	クラスタ 5: 35%	クラスタ 2: 28%
SF	クラスタ 2: 21%	クラスタ 4: 10%
その他	クラスタ 2:　8%	クラスタ 5:　5%

次に「小説家になろう」の小ジャンル 20 分類について，同ジャンルの作品間と別ジャンルの作品間で求めた S スコアの平均について表 5に示す．表 5から同ジャンルの作品間で求めた S スコアの方が高い傾向があることが確認できる．

表5　同ジャンル間と別ジャンル間の平均 S スコア

同ジャンル間の平均 S スコア	別ジャンル間の平均 S スコア
0.197	0.137

最後に S スコアを用いて同一作品判定を行った結果の精度を表 6に示す．同一作品判定においてはまず，取得した各作品の文章を半分ずつに分け，それぞれ別々に Doc2Vec のモデルを学習させ作品ベクトルを求めた．その後異なる作品の文章から求めたベクトルと，同じ作品の別の場所の文章から求めたベクトルを同量ずつ用意し，それぞれの S スコアを計算して，中央値以上なら同一作品と判定した．

別の場所の文章から求めたベクトル同士であっても，それが同じ作品のベクトルなら S スコアは高くなるべきで，実際にある程度同一作品判定が行えていることが確認できた．

表6　S スコアによる同一作品判定

判定方法	同一作品判定正答率
S スコアを利用	0.866
ランダム	0.501

以上いくつかの実験や分析を通して，今回設定し，実際に作品の本文から求めた S スコアが，作品の類似度を示すものとしてある程度妥当なものだと確認できた．今後より適切に S スコアを求める方法として，今回用いた Doc2Vec による作品ベクトル以外の特徴を用いることが考えられる．例えば 4.2.3 節でも触れた BERT[13] を用いて作品の文章の分散表現を求め，そこから同じくコサイン類似度を計算することで S スコアを求める方法なども検討している．

6　おわりに

隠れた良作の推薦や，ユーザによる推薦基準の操作が可能な web 小説レコメンドシステムとして，作品の類似度を示す S スコアと作品の質を示す Q スコアの二つの指標を用いて，この二つの指標を作品自体の情報から予測することで推薦を行うシステムを提案した．加えて，実際にそれぞれの指標を設定し機械学習モデルによる予測実験を行うことで，各指標の妥当性や予測可能性を確認しレコメンドシステムが作製できる見通しが立った．今後は今回用いたモデルをベースとしてシステムを作製しながら，モデルを改善し各指標の予測精度の向上を目指す．平行してデータの追加収集を行い，再実験の実施や，今回用いてないデータの利用も検討する．また，最終的な評価は完成したシステムをユーザに使ってもらい行う予定だが，S スコアと Q スコア，各指標の予測モデルについてもより適切な評価方法がないか考えていきたい．

参考文献

[1] Xiaoyuan Su, Taghi M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques”, Advances in artificial intelligence, 2009.

[2] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock, “Methods and Metrics for Cold-Start Recommendations”, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.253-260, 2002.

[3] 神嶌敏弘, “推薦システムのアルゴリズム（2）”，人工知能学会誌，Vol.23, No.1, pp.89-103, 2008.

[4] 建石由佳, 小野芳彦, 山田尚勇, “日本文の読みやすさの評価式”, 情報処理学会研究報告ヒューマンコンピュータインタラクション 25(1988-HI-018), pp.1-8, 1988.

[5] 近藤陽介, 松吉俊, 佐藤理史, “教科書コーパスを用いた日本語テキストの難易度推定”, 自然言語処理学会第 14 回年次大会発表論文集, pp.1113-1116, 2008.

[6] ジョディ・アーチャー, マシュー・ジョッカーズ, 川添節子訳, “ベストセラーコード”, 日本 BP 社, 2017.

[7] 清水一憲, 伊東栄典, 廣川佐千男, “集合知に基づくオンライン小説のランキング手法の提案と評価”, 情報処理学会研究報告, 2013.

[8] 飯田委哉, 伊東栄典, “セレンディピティを考慮した CGM 小説推薦”, 人工知能学会合同研究会 2018 第 15 回データ指向構成マイニングとシミュレーション研究会, 2018.

[9] 高田叶子, 佐藤哲司, “文体の類似度を考慮したオンライン小説推薦手法の提案”, DEIM Forum 2017 B5-2, 2017.

[10] 秦野智博, 阿部明典, “利用者の評価基準に合致した文章推薦システムの構築”, 2016 年度人工知能学会全国大会 (第 30 回), 2016.

[11] Quoc V. Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, Proceedings of the 31st International Conference on Machine Learning(ICML 2014), pp.1188-1196, 2014.

[12] Jey Han Lau, Timothy Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”, Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, pp.78-86, 2016.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.