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

高田　叶子†　　佐藤　哲司††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学大学院 図書館情報メディア系　〒 305-8550 つくば市春日 1-2
E-mail: †{ktakada,satoh}@ce.slis.tsukuba.ac.jp

あらまし　オンライン小説とは，web 上で誰でも投稿・閲覧することができる小説である．このような小説を投稿・閲覧する小説投稿サイトでは，投稿者数や投稿小説数が日々増加することなどから，一度埋もれてしまった小説を推薦することが困難である．本研究では，実世界での図書選択手法として広く用いられる「立ち読み」に着目し，立ち読みにおける図書選択の指標をアンケート調査によって明らかにするとともに，指標の有効性を利用者実験により検証する．調査の結果，同一著者や表紙の印象だけでなく，文章の構成や語彙など文体も重視されることが明らかとなった．小説投稿サイトに投稿された小説から文体に該当するいくつかの指標を抽出し，マハラノビス距離を用いて小説間の類似度を算出するオンライン小説推薦手法を提案・実装し，利用者実験により提案手法の有効性を検証した．

キーワード　小説推薦，オンライン小説，文体，小説投稿サイト

1. はじめに

オンライン小説とは，web 上で誰でも投稿・閲覧できる小説であり，オンライン小説を投稿・閲覧できるサイトを小説投稿サイトという．小説投稿サイトは，新しい作品発表の場としての役割を果たしている一方，投稿者数や投稿小説数が日々増加することによる課題も存在する．

利用者が小説投稿サイトを利用する目的は，人に読んでもらうことや書籍化を目指した小説の投稿と，無料かつ自分の好みにあう小説の閲覧の二つに分けられ，それぞれに課題が挙げられる．投稿者からすると，新規投稿者の小説や新規に投稿された小説，一度評価のつかなかった過去の作品は再び人目に触れづらいため，一度人気作家にならない限り，多くの人に自分の小説を読んでもらうことが難しい．また，後から参入した投稿者ほど，ランキングにも入りづらくなるため，人気作家になることが難しくなる．読者からすると，次々に新しい小説が増えるため，一度見つけそびれた作品を再び発掘することが困難である．また，出版された書籍において小説の選択を行う際，同著者のものや表紙が気に入るものを読むということを多く行うが，オンライン小説には同著者の小説が少ないことや，表紙がないことも多いため，出版された書籍に比べ小説選択のための手がかりが少ない．

つまり，小説投稿サイトは小説を投稿する利用者と小説を読む利用者双方に向けたサービスであるという特徴があり，オンライン小説の推薦においては読者に合った小説推薦を行うと同時に，新規に投稿された小説や評価数の少ない小説の推薦も必要となると言える．

従来の小説推薦には，同著者の著作を推薦手法，協調フィルタリングを用いた手法 [1]，表紙などの見た目による推薦手法 [2] などがある．しかし，オンライン小説には小説数が少ない著者が多く存在すること，評価数やブックマーク数が少ない小説が多くあること，新規投稿者の著作も多くあること，表紙や大きさ，重さが存在しないことなどにより，従来の推薦手法をそのままオンライン小説に用いることは難しい．

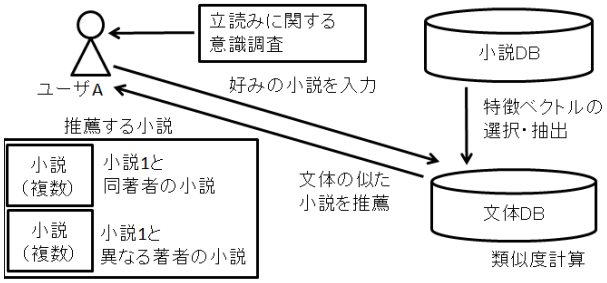

図 1　提案手法の概要

そこで，本研究では，立ち読みという小説選択手法に着目する．web 上で書籍を購入できるようになっても，現実世界において小説の購入を決める際，書店で立ち読みを行う人は多い．このとき読者は，web 上で得られる著者やレビュー，表紙といった特徴以外の何らかの情報を得ていると考えられる．そのため，立ち読みで重視される特徴量は，著者や他者の評価だけでない，新たな小説推薦の指標となると仮定できる．

本研究の目的は，小説推薦サイトにおける小説推薦手法の提案である．本研究では，読者が立ち読みという手法を用いて新しい小説の選択を行うことに着目し，立ち読みの効用を用いたオンライン小説における小説推薦手法を提案する．第一段階として，読者が立ち読みでどのような項目を見るのかについて，アンケート調査を行い立ち読みの効用を明らかにする．調査結果から明らかとなった立ち読みの効用に基づき，小説推薦手法を提案し，利用者実験を行い評価する．

以下，本論文では 2 章で関連研究を紹介し研究の位置づけを明確にする．3 章で立ち読みに関する調査を行い，4 章で提案手法を詳述する．5 章で評価した結果を述べ，考察する．最後に 6 章でまとめを示す．

2. 関 連 研 究

オンライン小説の推薦を目的とする本研究は，特に，ブックマークやレビューをはじめとした評価の数が少ない小説を推薦する点に課題がある．

2.1 書籍の推薦に関する研究

出版された書籍においては，協調フィルタリング，レビューや見た目による推薦など，本文の特徴を使わない研究が盛んに行われている．協調フィルタリングを用いた推薦として，原田ら [1] は，図書館の貸出履歴に重み付けした図書の推薦手法を提案している．オンラインショッピングサイトである amazon(注1) でも，「この商品を買った人はこんな商品も買っています」といった協調フィルタリングで図書の推薦をしている．

レビューを用いた推薦として，増田ら [3] は，感性的な表現を含んだ自然言語によるユーザとの対話を通した小説推薦システムの構築のため，アンケートと書評の分析を行っている．

見た目から書籍の推薦を行う研究としては，親泊ら [2] の研究がある．親泊らは表紙の好みを利用して同じ嗜好を持つユーザを発掘し，ユーザの好みに合った書籍を推薦する手法を提案している．

感情表現を扱った研究も存在する．原田 [4] は，指定した図書と感性パラメータの分布が類似する図書を提示するシステムを試作している．児童書・ヤングアダルト図書 1425 冊を対象として，10 名の被験者に入力した図書と関連する図書 30 冊を提示したところ，約 53%の図書について強い興味があるという結果を確認している．

辻ら [5] は，協調フィルタリングやアソシエーションルールと，タイトルや NDC などの情報を，SVM によって統合的に利用し，情報を単独で用いた場合や Amazon による推薦と比較検証している．結果，「NDC ＋タイトル＋貸出履歴」と「タイトル＋貸出履歴」の組合せが有効であること，Amazen の推薦の方が評価が高いことを確認している．

小野寺ら [6] は，レビュー分析に基づいて選定したオノマトペに対してユーザが持つ印象を調査し，オノマトペを入力としてそれに合った小説を出力とする小説推薦システムを開発することで，直感的に小説を推薦できかつユーザの嗜好に合った小説を推薦できることを示している．しかし，この方法では，あらかじめ各小説に対し印象値を割り当てておく必要がある．

2.2 小説投稿サイトに関する研究

小説投稿サイトを対象とした研究には，将来ランキング上位になる小説を早期にを推定する研究がある．清水ら [7] は，読者のお気に入り登録を集合知として扱い，お気に入り登録のリンク構造に基づく小説ランキング手法を提案することで，幾つかのジャンルについて，5 月時点で 7 月の人気上位の小説の推定に成功している．

2.3 文体の類似度に関する研究

文章を数値化する研究は，計量文体学，計量文献学と呼ばれ，その活用方法としては，著者推定，特定の種類の文書の特徴分析が主流である．

金川 [8] は，構文構造に着目して、文体の類似性を数値化する手法を提案している．石田ら [9] は既往研究で扱われている文体的指標をまとめ，そのうちいくつかを学術論文，日記，新聞記事に対して調査を行い，文頭・文末表現，接続詞で特有の表現が多く用いられていることを明らかにしている．

2.4 本研究の位置づけ

本研究では，文体の類似度と小説のジャンルを組み合わせることで，小説投稿サイト上の小説を読者に推薦する．そのため，事前に読者の少ない小説を推薦できるという点，貸出履歴などの事前情報が少ない小説を扱うという点において，2.1 で述べた手法と異なる．オンライン小説に関する研究として特に推薦を扱ったものはなく，その点で 2.2 と異なる．また，文体の類似度に関する特徴量の選出においては，2.3 で述べた既存の研究で使われた手法に加え，独自の特徴量を追加する．

本研究は，読者への小説推薦に加え，投稿者への支援を加味した小説推薦システムである点に新規性がある．

3. 小説の選択に関する意識調査

3.1 調査の目的と方法

1 章で述べた通り，書籍を対象とする小説推薦手法をそのままオンライン小説に適用することは難しい．一方，小説の購入を決める際に書店で立ち読みを行う人は多く，立ち読みで重視される特徴量は，著者や他者の評価だけでなく，新たな小説推薦の指標となると仮定できる．そこで，立ち読みの有効性の調査，及び，立ち読みで考慮されている特徴量を明らかにすることを目的としてアンケート調査を行った．

対象は，筑波大学で情報系の基礎科目を受講する主に 1 年生であり，調査手法は web 上でのアンケート調査である．アンケートの依頼用紙を受講生 120 名に配布し，有効回答数は 91 であった．

3.2 結果と考察

「小説をよく読みますか」という項目に対し，「全く読まない」と回答した 7 名，及び，「読まない」と回答し，月に読む冊数が 0 または無回答であった 5 名の回答を除外した，79 名を対象に調査結果を示す．

小説を選ぶ方法

小説を選ぶ際の方法と頻度を問うた．複数回答可の設問である．結果を図 2 に示す．

「よく使う」と回答した人と「たまに使う」と回答した人を合わせると，著者を使うと答えた人が一番多く 93.7%という結

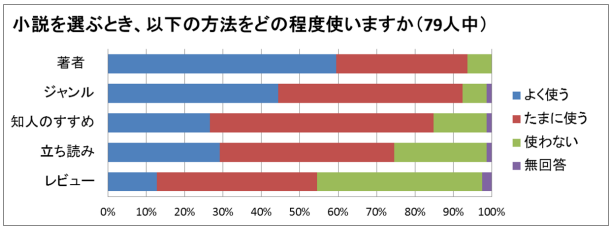

図 2 小説を選ぶ方法

(注1)：https://www.amazon.co.jp/

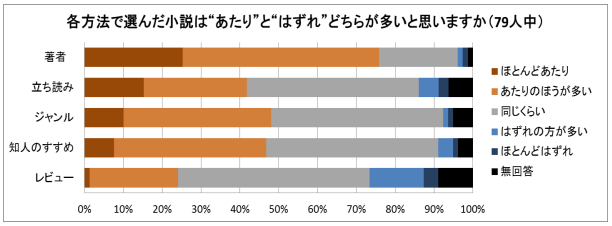

図 3　各小説選択方法の有効性

表 1　立ち読みの頻度と立ち読みによる小説選択の相関

		立ち読みで選んだ小説のあたりとはずれ						
		ほとんどあたり	あたりの方が多い	同じくらい	はずれの方が多い	ほとんどはずれ	無回答	合計
立ち読み	よく使う	9 39.1%	6 26.1%	6 26.1%	1 4.3%	0 0.0%	1 4.3%	23 100.0%
	たまに使う	3 8.3%	12 33.3%	19 52.8%	2 5.6%	0 0.0%	0 0.0%	36 100.0%
	使わない	0 0.0%	3 15.8%	10 52.6%	1 5.3%	2 10.5%	3 15.8%	19 100.0%
	無回答	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 100.0%	1 100.0%
	合計	12 15.2%	21 26.6%	35 44.3%	4 5.1%	2 2.5%	5 6.3%	79 100.0%

果であり，それ以下は，ジャンル，知人のすすめ，立ち読み，レビューの順である．立ち読みを使うと答えた人は，レビューより 20%ほど高い結果である.「よく使う」と回答した人のみ見ると，著者，ジャンル，立ち読み，知人のすすめ，レビューの順である．この結果から，同著者の小説を読むなど著者による小説選びがかなり高い比率で行われていること，多くの人が小説を選ぶときに複数の方法を使うことが明らかになった．選択肢以外の方法を記入する自由記述欄を設けたところ，表紙という回答が 14 あった．そのほか，勘，あらすじ，編集後記，背表紙，タイトル，映像化の原作，裏の解説，といった回答も一定数あった．

小説選択方法の有効性

各方法で選んだ小説に“あたり”と“はずれ”どちらが多いと思うかについて調査した．結果を図 3 に示す．

著者による選択を除いたいずれの方法においても，同程度という結果が一番多かった．著者に関しては，75%が「ほとんどあたり」または「あたりの方が多い」との回答であった.「ほとんどあたり」または「あたりの方が多い」と答えた人の割合は，著者 50.6%，ジャンル 48.1%，知人のすすめ 46.8%，立ち読み 46.8%と立ち読みが若干少なかったものの,「ほとんどあたり」と答えた人の割合は，著者 25.3%，ジャンル 10.1%，知人のすすめ 7.6%，立ち読み 15.2%と，立ち読みは二番目に多い結果となった．選んだ小説が「ほとんどあたり」であるとの回答が比較的多く得られた立ち読みは，小説推薦において有効な手法になりうると考えられる．

「小説を選ぶとき、立ち読みをどの程度使いますか」という項目と「立ち読みで選んだ小説は“あたり”と“はずれ”どちらが多いと思いますか」という項目のクロス集計を表 1 に示す．

立ち読みで購入した小説に「ほとんどあたり」または「あたりのほうが多い」とした人は，全体で 41.8%，立ち読みを「よく使う」人で 65.2%,「たまに使う」人で 41.6%,「使わない」人で 15.8%であった.「ほとんどはずれ」または「はずれのほうが多い」とした人は，全体で 76%，立ち読みを「よく使う」人で 4.3%,「たまに使う」人で 5.6%,「使わない」人で 15.8%であった．

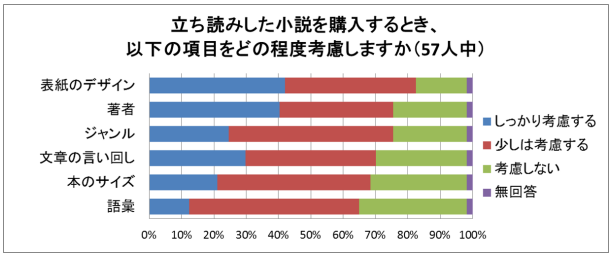

図 4　小説を立ち読みをするときに考慮する項目

立ち読みで考慮される特徴量

前項までの対象者 79 名から,「小説を選ぶとき、立ち読みをどの程度使いますか」という項目が「使わない」または「無回答」であった 20 名を除外し，そこからさらに「小説を立ち読みしたことがありますか」という項目に「いいえ」と回答した 22 名を除外した 57 名を対象に，小説を立ち読みをするときに考慮する項目を調査した．結果を図 5 に示す．複数回答可の設問である．

「しっかり考慮する」と「少しは考慮する」を合わせた数では，表紙のデザイン 82.5%，著者 75.4%，ジャンル 75.4%，文章の言い回し 70.2%が高い値となった．自由記述欄には，登場人物達の結末，面白そうかどうか，翻訳者が挙げられた．表紙のデザインや著者など従来の推薦手法に加え，文章の言い回し，すなわち文体が購入の際に高い割合で考慮されていることがわかる．

3.3　調査のまとめ

小説を選ぶ際，立ち読みを使う人は，74.7%という結果となった.「よく使う」と「使う」を合計した結果では，著者やジャンル，知人のすすめより低い値ではあるものの，小説推薦の手法としてよく用いられるレビューよりも 20%ほど高い数字である．また,「よく使う」と答えた人の割合では知人のすすめを上回っており，このことから，立ち読みは小説を選ぶ際の主要な方法であるといえる．

また，立ち読みで選んだ小説が「ほとんどあたり」と答えた割合は 5 つの方法の中でも著者に次いで 2 番目であり，小説を選ぶ際に立ち読みを「よく使う」と回答した人の 65%以上が立ち読みで選んだ小説について「ほとんどあたり」または「あたりの方が多い」と答えていることから，立ち読みは高い有効性があると考えられる．

立ち読みで考慮される主な特徴として，70%以上の人が考慮すると答えた，表紙のデザイン，著者，ジャンル，文章の言い回しなどが考えられる．本研究では，文章の言い回しに約 65%の人が考慮すると答えた語彙を加えて文体と定義し，オンライン小説に適用できる，ジャンルと文体を使用することとした．

表 2　使用する特徴量

特徴量	説明
一文あたりの読点数	読点数/文数
一文あたりの読みの文字数	読みでの文字数/文数
一文あたりの文字数	文字数/文数
一文あたりの品詞数	品詞数/文数
一文あたりの文節数	文節数/文数
一文あたりの助詞数	助詞数/文数
漢字の割合	漢字の数/文字数
ひらがなの割合	ひらがなの数/文字数
カタカナの割合	カタカナの数/文字数
句読点間の読みの文字数	読みでの文字数/句読点の数
品詞の割合	各品詞数/全品詞数
読点の前の品詞の割合	読点前の各品詞数/読点前の全品詞数
句点の前の品詞の割合	文末の各品詞数/文末の全品詞数
文頭の品詞の割合	文頭の各品詞数/文頭の全品詞数
一文あたりの直喩数	「よう」と「みたい」の数/文数
一文あたりのルビの数	ルビの数/文数
一文あたりのオノマトペ数	オノマトペ数/文数
語彙（TTR）	異なり語数/延べ語数
一行あたりの文字数	全文字数/全改行数

4.　文体の類似度を考慮した小説推薦手法

4.1　提案手法の概要

オンライン小説本文から文体に関するいくつかの特徴量を選択，抽出し，マハラノビス距離を用いて小説間の距離を算出することで，文体の類似度を求める．読者が選んだ小説と同ジャンル内の小説の中から，文体の類似度が高いものを推薦小説として提示する．

4.2　使用する特徴量の選択

既存研究で使われている特徴量から，英語論文の場合日本語に応用できるものであること，大量の小説を対象とするため目視を使わず抽出できること，を基準に選択を行う．具体的には，望月ら [10]，齋藤ら [11]，石田ら [9]，土山ら [12]，土山 [13]，劉ら [14]，小西 [15]，工藤ら [16] の既存研究より，読点数，読みの文字数，文字数 ，品詞数，文節数，助詞数，漢字の割合，ひらがなの割合，カタカナの割合，句読点間の読みの文字数，品詞の割合，読点の前の品詞の割合，句点の前の品詞の割合，文頭の品詞の割合，オノマトペ数（声喩数），直喩数，語彙（TTR）を使用する．作品により文字数に差があるため，特徴量によっては一文あたりの数を計算に用いる．

また，独自に追加する特徴量として，ルビの数，一行あたりの文字数を使用する．これは，オンライン小説には，ルビが多用される小説が見られること，好きなところで好きな回数改行を入れることができることによる．

使用する特徴量を表 2 に示す．各品詞の割合に関しては，助詞，助動詞，名詞，動詞，形容詞，接頭詞，感動詞，連体詞，副詞，接続詞の 10 品詞を用いるため，特徴量は全部で 55 である．

4.3　使用する特徴量の収集

小説投稿サイトから小説が書かれたページの html ソースを収集する．それぞれの特徴量を抽出するにあたり，表 3 に示す 5 つのテキストを使用する．

表 3　特徴量の抽出に使用するテキスト

使用するテキスト	説明	例
サイト上での表示		今日からここが君の舞台（いばしょ）だ！一緒に頑張ろう。
原文テキスト	小説投稿サイトから収集した html ソースから本文のみを抜き出したテキスト	今日からここが君の ＜ ruby ＞＜ rb ＞ 舞台 ＜ /rb ＞＜ rp ＞ (＜ /rp ＞＜ rt ＞ いばしょ ＜ /rt ＞＜ rp ＞) ＜ /rp ＞ だ！一緒に頑張ろう。
本文テキスト	原文テキストからルビ部分と挿絵部分を削除したプレーンテキスト	今日からここが君の舞台だ！一緒に頑張ろう。
読みテキスト	原文テキストのルビ部分を残し，ルビに対応する本文と挿絵を削除したプレーンテキスト	今日からここが君のいばしょだ！一緒に頑張ろう。
文テキスト	本文テキストの末尾記号を削除し，一文ごとに改行したテキスト	今日からここが君の舞台だ 一緒に頑張ろう
カタカナテキスト	読みテキストの文章を全文カタカナに直したテキスト	キョウカラココガキミノイバショダ!イッショニガンバロウ。

文数，句点数

本研究では，簡便のため，文数と句点数を同一のものとして扱う．つまり,「！」や句点なしで文章が終わる文であっても，そこには句点がついているものとして扱う．

まず，ルビ部分を削除したプレーンテキスト「本文テキスト」を作成する．次に，本文テキストに対し，文ごとに改行を行う．『。』,『.』,『.』,『？』,『！』,『」』,『?』,『!』,『)』,『…』を文末記号として扱い，これらの文末記号を改行に置換する．その後，2 回以上続く改行を削除することで，文ごとに改行されたテキスト「文テキスト」を作成する．文テキストの改行数を数えることで，文数，句読点数とする．

読　点　数

本文テキストから,『,』,『,』,『、』の数をカウントする．

読みでの文字数

ルビ部分を残しルビに対応する本文を削除したプレーンテキスト「読みテキスト」を作成する．読みテキストに対して，形態素解析エンジン Mecab(注2) の-Oyomi オプションを用いる．-Oyomi オプションを用いることで，入力したテキストファイルの読みがカタカナで出力される．このとき得られたテキストファイルをカタカナテキストとよぶ．カタカナテキストの文字数をカウントしたものを読みでの文字数とする．

文　字　数

文テキストの文字数をカウントする．そのため，記号はカウントされるが，文末表現はカウントされない．

(注2)：http://taku910.github.io/mecab/

品詞，読点の前品詞，文頭の品詞，文末の品詞

形態素解析エンジン Mecab を用いて文テキストの形態素解析を行い，各品詞数をカウントする．Mecab では，一行一文を前提として解析が行われ，間に EOC という文字列が入るため，EOC の前の形態素の品詞と EOC の後の形態素の品詞を，それぞれ文頭の品詞と文末の品詞として用いる．読点の前の品詞に関しては,『,』,『,』,『、』の前の形態素の品詞を用いる．

文　節　数

日本語係り受け解析器 Cabocha(注3) を用い，文テキストに対し係り受け解析を行う．Cabocha の出力データには文節の区切り情報が付与されているため，文節数をカウントする．

漢字，ひらがな，カタカナの数

正規表現を用い，それぞれの文字数をカウントする．

直　喩　数

Mecab で解析した形態素の中から，直喩表現として，名詞「よう」と名詞「みたい」の数を合計した数を直喩数とする．

ル ビ の 数

収集した html ソースから本文部分のみを抜き出した「原文テキスト」を作成する．原文テキストから，ルビを示す html タグの数をカウントする．

オノマトペ

渡辺ら [17] の研究では 126 語のオノマトペを使用している．オノマトペごとの印象推定システムを提案する清水ら [18] の研究では 312 語のオノマトペを使用している．小説・コラム・ブログなど物書きの参考書を目指している，日本語表現インフォ(注4) にもオノマトペが多く掲載されている．渡辺ら [17]，清水ら [18] の研究にあるオノマトペに加え，日本語インフォから 201 語，独自にジョキジョキ，シャビシャビ，ヒョーヒョーの 3 語を使用する．これらのオノマトペを全てカタカナに直し，重複を除き，2 文字のオノマトペは削除した．この結果，538 語のオノマトペを得た．これらのオノマトペがカタカナテキストに出現する回数を，その小説のオノマトペ数とする．

語　　彙

内容語である，名詞，形容詞，動詞，副詞の語彙を取得する．語彙の計算方法には，もっとも良く知られる指標である TTR（Type-Token Ratio）を用いる．

対象となる文章の延べ語数を N，異なり語数を V と定義したとき，$TTR = \frac{V}{N}$ である．

改　行　数

空行も含め原文テキストの改行数をカウントする．

4.4　類似度の計算方法

小説 i から抽出した各特徴量を x_{ih} とし，小説 i を H 個の特徴量からなる特徴ベクトル $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,H})$ で定義する．本研究では，各特徴量ごとに分散（標準偏差）の規模が大きく異なる．その差異を考慮するため，以下のマハラノビス距離により，小説 i と小説 j 間の距離を定義する：

$$Dist(i,j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)}.$$

(注3)：https://taku910.github.io/cabocha/

(注4)：http://hyogen.info/

表 4　「小説家になろう」内のジャンル分けと作品数

大ジャンル	作品数
恋愛	28,603
ファンタジー	41,810
文芸	40,555
SF	7,504
その他	28,191
ノンジャンル	293,892
合計	440,556

ここで，Σ^{-1} は共分散行列の逆行列を意味する．本研究では，簡便のため特徴量間の共分散を無視した対角行列を Σ とする，すなわち，h 番目の特徴量の全小説における分散を σ_h^2 とすると，小説 i の h 番目の特徴量 x_{ih} と小説 j の h 番目の特徴量 x_{jh} の間の距離を

$$d_h(i,j) = \frac{(x_{ih} - x_{jh})^2}{\sigma_h^2}$$

により計算し，

$$Dist(i,j) = \sqrt{(\sum_{h=1}^{H} d_h(i,j))}$$

としている．マハラノビス距離は距離の公理を満たすため，$D(i,j) = D(j,i)$ となる．

算出されたマハラノビス距離に基づき，距離の近い順に順位付けを行い，その順位を文体の類似度が高い順とする．

5.　評 価 実 験

5.1　評 価 方 法

小説投稿サイトでは投稿者と読者の両方が利用者となる．そのため，評価においても上記の 2 つの観点から評価を行う．

投稿者の支援という観点では，感想数やブックマーク数や総合評価の値が低い小説に関しても上位に提示することができているかを確認する．利用者の満足という観点では，利用者実験により，実際に小説を読む際，どの程度提案手法で推薦された小説を好ましいと思うかを評価する．

5.2　評 価 対 象

ヒナプロジェクト社が提供する日本の小説投稿サイトである「小説家になろう」に投稿された小説を使用する．表 4 に 2016 年 12 月 15 日時点での「小説家になろう」内のジャンル分けと作品数を示す．

評価実験に使うジャンルの選定は,「小説家になろう」内において，ノンジャンルを除き一番大きなジャンルであることから，ファンタジーを選定した．ノンジャンルに関しては，作品ジャンル再編成により 2016 年 5 月 24 日時点で投稿済み作品が全てノンジャンルに編成されているため，ジャンルとして扱わないこととした．ファンタジージャンル内には「ハイファンタジー」「ローファンタジー」という一段階小さいジャンルが存在するが，今回は大きいジャンル編成のみ用いることとする．

5.2.1　推薦基準とする 10 小説の選定

「小説家になろう」に投稿された小説を検索・閲覧できるサイト「小説を読もう」(注5) では，詳細検索において，読了時間指定，文字数指定，範囲検索（年月日の指定），抽出条件，除外条件，並び替え，種別，ジャンル，検索キーワード指定，除外キーワード指定，キーワード検索指定が可能である．

推薦基準とする 10 小説の選定にあたっては，まず，文字数指定を 16800 文字以上，年月日を 2016 年 06 月 30 日以前，除外条件として「長期連載停止中」「R15」「残酷な描写あり」「ボーイズラブ」「ガールズラブ」，並べ替えを総合評価の高い順，ジャンルにファンタジーを指定し検索を行う．2016 年 11 月 25 日 14 時の時点で条件に当てはまる 2,412 小説のうち，4 小説以上ジャンルを除いた条件が同じ小説を書いている著者の小説を，上から 10 小説選ぶ．

このうち，シリーズの続編であるものに関しては，代わりに同シリーズの 1 作目を使用する．

5.2.2　評価に用いるデータセット

詳細検索において，文字数指定を 16800 文字以上，年月日を 2016 年 06 月 30 日以前，除外条件として「長期連載停止中」「R15」「残酷な描写あり」「ボーイズラブ」「ガールズラブ」，並べ替えとして新着順，ジャンルにファンタジーを指定して検索を行う．2016 年 11 月 15 日 13 時 30 分の時点で条件に一致した 2,397 作品から，新着順に 2,000 作品を収集する．そこに，推薦基準の 10 小説と同著者かつジャンル以外同条件の作品を加えた，2,043 作品を使用する．

5.3　投稿者支援に関する評価

5.3.1　方　　法

「小説家になろう」には，読者による評価制度が存在する．具体的には，感想，レビュー，ブックマーク数，ポイント評価である．著者，及び，読者は，評価件数やポイント数を見ることができる．また，ブックマーク登録数，総合評価，一定期間内のポイント数，週間ユニークアクセス数，新着などは「小説を読もう」で小説を検索する際のソートに使うことができ，レビューはレビュー一覧からみることができる．つまり，これらの評価指標の値が高いほど人目に触れやすいといえる．

本研究では，小説推薦における投稿者支援の側面として，新規の著者の著作や，投稿された時点で評価数がさほどつかなかった著作など，人目に触れる機会が少ない作品についても推薦できることを目標としている．そこで，基準の小説に対して，提案手法で上位に推薦される小説に対し，その小説の「小説家になろう」内での各指標数を示し，評価指標が低いものも推薦できることを確認する．

5.3.2　結果と考察

基準の 10 小説中 2 小説に対する，提案手法で上位になる小説と投稿サイト内での評価を表 5 に示す．推薦順位に色がついているものは，基準の小説と同著者の小説である．小説投稿サイト内での評価が高いものに混ざって，評価が少ないものも推薦できていることがことがわかる．

(注5)：http://yomou.syosetu.com/

表 5　提案手法で上位になる小説と投稿サイト内での評価

推薦順位	感想 (件)	レビュー (件)	ブックマーク数 (件)	総合評価 (pt)	文章評価 (pt)	ストーリー評価 (pt)
基準	1,105	1	10,623	32,792	5,741	5,805
1	265	1	2,897	7,936	1,071	1,071
2	137	0	1,712	4,689	635	630
3	84	1	4,621	11,086	908	936
4	1	0	24	57	4	5
5	38	0	2,734	6,464	493	503
6	221	0	3,704	10,137	1,358	1,371
7	129	1	3,478	8,595	815	824
8	0	0	3	29	12	11
9	73	5	100	331	66	65
10	0	0	1	2	0	0
基準	65	1	2,673	9,343	1,984	2,013
1	405	1	4,848	15,426	2,828	2,902
2	45	0	1,851	4,900	601	597
3	0	0	28	63	3	4
4	22	0	44	212	62	62
5	0	0	22	102	29	29
6	0	0	1	2	0	0
7	214	0	377	1,042	141	147
8	1	0	0	10	5	5
9	0	0	15	30	0	0
10	0	0	0	0	0	0

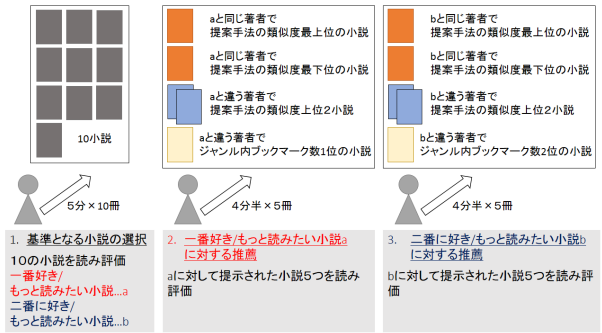

図 5　評価実験方法の概要

5.4　読者の満足度に関する評価

5.4.1　方　　法

ファンタジー小説が好きな筑波大学の学生 21 名を対象に，利用者実験を行った．概要を図 5 に示す．

実験参加者は，まず，推薦基準となる 10 小説を 5 分ずつ読み，一番好き/もっと読みたい小説と，二番目に好き/もっと読みたい小説を選択する．次に，一番好き/もっと読みたい小説に対し提示された 5 小説をそれぞれ 4 分 30 秒読み，アンケートに答える．最後に，二番目に好き/もっと読みたい小説に対し提示された 5 小説をそれぞれ 4 分 30 秒読み，アンケートに答える．推薦基準の小説に対し提示する 5 冊は次のとおりである．

（1）　同著者で提案手法最上位の小説（以下，同・高）
（2）　同著者で提案手法最下位の小説（以下，同・低）
（3）　異著者で提案手法最上位の小説（以下，異・高 1）
（4）　異著者で提案手法 2 位の小説（以下，異・高 2）

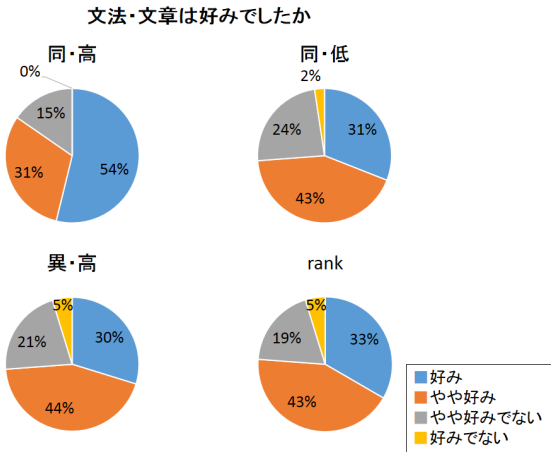

図 6　文体の好み

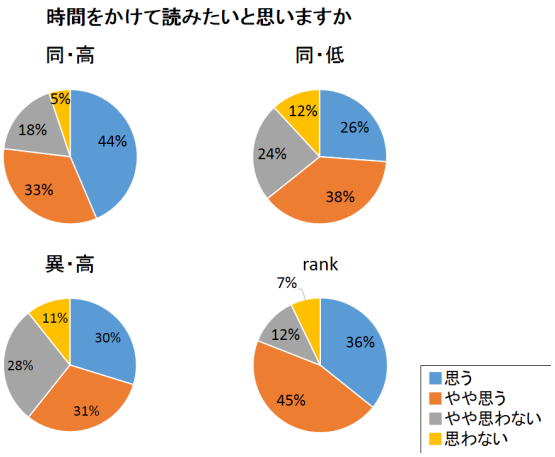

図 7　小説推薦の満足度

表 6　同著者提案手法低類似と異著者提案手法高類似の文体の好み比較

	一番好きな小説に関して	二番目に好きな小説に関して
合計 (21 名)	9	-11

（5） 異著者でブックマーク数が多い小説（以下，rank）

ブックマーク数に関しては，使用した小説の中で最もブックマーク数が多いものを「一番好き/もっと読みたい小説と」との比較に対して提示し，次にブックマーク数が多いものを「二番目に好き/もっと読みたい小説」との比較に対して提示した．

5.4.2 結　　果

「文法・文章は好みでしたか」，及び,「時間をかけて読みたいと思いますか」に対する回答の割合を図 6，及び，図 7 に示す．

両方の設問において，同・高のほうが同・低より優位な値である．これは，提案法の有効性を示しているといえる．また，同・低と異・高では同程度の結果となった．

5.4.3 同著者の小説と異著者の小説の比較と考察

同著者低類似度の小説と異著者高類似度の小説

次に，同程度の結果が得られた同・高と異・低を比較する．

表 6 に，以下の式の回答値を示す．それぞれ，bun は「文法・文章は好みでしたか」，$jikan$ は,「時間をかけて読みたいか」の設問に対する実験参加者 i 答えの値である．

表 7　基準小説の物語（ストーリー）と文法・文章の好み

	一番好きな小説		二番目に好きな小説	
	物語 (ストーリー)	文章・文法	物語 (ストーリー)	文章・文法
平均 (21 名)	1.381	1.333	1.524	1.619

表 8　文体の好み

	好み	やや好み	やや好みでない	好みでない	合計
同著者	34	30	16	1	81
	42.0%	37.0%	19.8%	1.2%	100.0%
異・高 a	2	2	0	0	4
	50.0%	50.0%	0.0%	0.0%	100.0%
異・高 b	1	3	0	0	4
	25.0%	75.0%	0.0%	0.0%	100.0%
異・高 c	3	1	0	0	4
	75.0%	25.0%	0.0%	0.0%	100.0%

表 9　小説推薦の満足度

	思う	やや思う	やや思わない	思わない	合計
同著者	28	29	17	7	81
	34.6%	35.8%	21.0%	8.6%	100.0%
異・高 a	1	3	0	0	4
	25.0%	75.0%	0.0%	0.0%	100.0%
異・高 b	2	2	0	0	4
	50.0%	50.0%	0.0%	0.0%	100.0%
異・高 c	4	0	0	0	4
	100.0%	0.0%	0.0%	0.0%	100.0%

$$
\begin{aligned}
diff_i = &(\text{同・低}_{bun} - \text{異・高 } 1_{bun}) \\
&+(\text{同・低}_{jikan} - \text{異・高 } 1_{jikan}) \\
&+(\text{同・低}_{bun} - \text{異・高 } 2_{bun}) \\
&+(\text{同・低}_{jikan} - \text{異・高 } 2_{jikan})
\end{aligned}
$$

結果が正の値となると，異・高の小説が優位であること，負の値となると同・低の小説が優位であることを示す．

また,「一番好き/もっと読みたい小説と」と答えた小説，及び,「二番目に好き/もっと読みたい小説」として答えた小説に対して，物語（ストーリー）と文法・文章が好みであったかをそれぞれ 4 段階の評価スケールで質問した結果とその平均を表 7 に示す．1 が好み，2 がやや好み，3 やや好みでない，4 が好みでないの 4 段階の評価尺度であり，したがって値が 1 に近いほど好ましく，4 に近いほど好ましくない．

一番目に好き/もっと読みたい小説に関しては，物語の好ましさよりも文章の好ましさの方が高い値で選ばれており，二番目に好き/もっと読みたい小説に関しては，文章の好ましさよりも内容の好ましさの方が高い値で選ばれている．そして，文章の好ましさをより重視して選ばれていると考えられる一番目に好き/もっと読みたい小説への推薦においては，異・高の小説の方がより多くの参加者から好まれた．

同著者の小説全体と異著者高類似度の小説

全体の結果を見ると，異著者高類似度の小説より同著者高類似度の小説のほうが満足度が高い結果となったが，小説ごとに

見た場合，同著者高類似度の小説と同程度の満足度を得られた異著者高類似度の作品も存在した．

提案手法で高類似度となった異著者の小説の中でも，特に満足度の高かった小説を表 8，表 9 に示す．表 8，表 9 における同著者の値は同・低，及び，同・高のに対するアンケート結果を合わせたものであり，表 8，及び，表 9 における異・高 a から異・高 c はそれぞれ同じ小説である．ここから，部分的にではあるものの，同著者高類似度の小説と同程度以上に，読者の満足度が高い異著者の小説を提案法により推薦できていることがわかる．

3 章で示したとおり，同著者による小説選択は約 94%の支持を得ており，76%が著者で選んだ小説はあたりの方が多いと答えるなど，他の方法に比べ圧倒的に有効な小説推薦方法である．著者による推薦と同程度以上の結果を得られるということは，文体の類似度を考慮した小説推薦の有効性を示唆しているといえる．

6. ま と め

本研究では，小説投稿サイトにおける小説推薦を目的に，ジャンルと文体の類似度を考慮したオンライン小説推薦手法を提案した．小説投稿サイトでは，同著者の作品が比較的少ないこと，一般的に表紙が存在しないこと，投稿者の支援と読者の満足度の 2 つの観点での推薦が求められることから，提案手法では，小説本文における文体の類似度に着目した推薦を行う．文体に着目することで，異なる著者の作品であっても推薦が行えること，過去の埋もれた作品であっても推薦の候補とできることなど，オンライン小説サイトの課題を解決できると期待される．

実際に投稿されているオンライン小説を用いて利用者実験を行い，提案手法の有効性を検証した．実験は投稿者支援と読者の満足度の 2 つの観点から評価を行った．投稿者支援の観点では，提案手法での推薦では小説投稿サイト内での評価数が少ない小説も推薦できることを明らかにした．読者の満足度の観点では，同著者の著作の中でも提案法の類似度が高いほうが満足度が高いこと，一部の小説に関しては，提案法の類似度が高い異著者の小説が，同著者の小説と同程度の満足度を得られることを明らかにした．

謝　　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです。

文　　献

[1] 原田隆史, 増田浩佑. 貸出記録を用いた図書推薦システムにおける重みづけの変更. ディジタル図書館, No. 38, pp. 54–66, Mar. 2010.
[2] 親泊広直, 菊地佑介, 岸野文郎, 中島康祐, 伊藤雄一. 表紙の好みに基づく書籍推薦システムに関する検討. 電子情報通信学会総合大会講演論文集, 2014 年__基礎・境界, p. 169, Mar. 2014.
[3] 増田純太, 杉本徹. 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. 情報科学技術フォーラム講演論文集, Vol. 11, No. 2, pp. 189–190, Sep. 2012.
[4] 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, May. 2011.
[5] 辻慶太, 滝沢伸也, 佐藤翔, 池内有為, 池内淳, 芳鐘冬樹, 逸村裕. 図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性. 図書館界, Vol. 65, No. 4, pp. 253–267, Nov. 2013.
[6] 小野寺祐貴, 杉本徹. オノマトペを利用した小説推薦システムの開発. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 114, No. 444, pp. 23–28, Jan. 2015.
[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. 研究報告データベースシステム, Vol. 2012-DBS-156, No. 18, pp. 1–6, Dec. 2012.
[8] 金川絵利子, 佐原諒亮, 岡留剛. 作家の文体の類似性：情報量木カーネルの導入による構文間距離を用いた分析. 人工知能学会全国大会論文集, Vol. 29, pp. 1–4, Sep. 2015.
[9] 石田栄美, 安形輝, 野末道子. 文体からみた学術的文献の特徴分析. 三田図書館・情報学会研究大会発表論文集, pp. 33–36, 2004.
[10] 望月朝香, 泰博鈴木. 小説における文体印象解析の試み. 情報処理学会研究報告数理モデル化と問題解決（MPS）, Vol. 2007, No. 128, pp. 179–182, Dec. 2007.
[11] 齊藤雄大, 長谷川大, 佐久田博司. 文章のリズムを考慮した小説執筆支援システムの作成. 第 75 回全国大会講演論文集, Vol. 2013, No. 1, pp. 145–146, Mar. 2013.
[12] 土山玄, 村上征勝.『源氏物語』第三部の成立に関する計量的な考察. じんもんこん 2014 論文集, Vol. 2014, No. 3, pp. 213–220, Dec. 2014.
[13] 土山玄. 文学作品の計量分析：その方法と歴史. 研究報告人文科学とコンピュータ, Vol. 2015-CH-107, No. 7, pp. 1–6, Aug. 2015.
[14] 劉雪琴, 金明哲. 宇野浩二の文体変化に関する計量的分析. 日本行動計量学会大会発表論文抄録集, Vol. 43, pp. 214–215, Sep. 2015.
[15] 小西光. 近代口語文翻訳小説コーパス構築の概要と計量的分析. 国立国語研究所論集, No. 11, pp. 37–61, Jul. 2016.
[16] 工藤彰, 村井源, 徃住彰文. 計量分析による村上春樹文学の語彙構成と歴史的変遷. 情報知識学会誌, Vol. 20, No. 2, pp. 135–140, May. 2010.
[17] 渡辺知恵美, 中村聡史. オノマトペロリ：味覚や食感を表すオノマトペによる料理レシピのランキング. 人工知能学会論文誌, Vol. 30, No. 1, pp. 340–352, 2015.
[18] 清水祐一郎, 土斐崎龍一, 坂本真樹. オノマトペごとの微細な印象を推定するシステム. 人工知能学会論文誌, Vol. 29, No. 1, pp. 41–52, 2014.