オンライン小説の検索に有効なタグの推薦手法の提案

山崎　睦月†　　佐藤　哲司††

† 筑波大学情報学群　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †{yamazaki19,satoh}@ce.slis.tsukuba.ac.jp

あらまし　数多くのオンライン小説が投稿されている小説投稿サイトでは．投稿小説の増加で，投稿者は自身の小説が多くの小説に埋もれ読まれづらくなる，読者は多くの小説の中から趣向にあった小説を探さなければならなくなってきている．多くの小説投稿サイトでは，投稿小説にタグを設定することで，新たな情報を小説に付与できる．そのため，検索に有効なタグを設定することで，利便性を大幅に向上できると考えられる．ここで，有効なタグとは，小説の内容を合致していてかつタグ間で意味の重複がない，小説の数に対して一つのタグが小説に設定された数が十分に小さいタグとする．本研究では，一つのタグが小説に設定された数が十分に小さい範囲を求め，Doc2Vec によってベクトル化した小説から，小説に設定されたタグは小説と同じ性質を持つという仮定に基づいて，タグをベクトル化し，タグ間で類似度比較を用いることで意味の重複を防ぎ検索に有効なタグを推薦する．

キーワード　オンライン小説，情報推薦

1 研 究 背 景

小説投稿サイトとは，オンライン小説専用の CGM(Consumer Generated Media) サイトである．小説投稿サイトの登場・発展により誰もが容易にインターネット上に自身の作品を投稿し，全世界に向けて発表可能である．ヒナプロジェクト社が運営している日本の小説投稿サイト の一つである「小説家になろう」[1] にはおよそ 6 か月間の間に約 39,000 件の小説が投稿されている（2019 年 6 月 5 日では 657,825 件だった投稿数が同年 12 月 12 日には 696,658 件に増加している）．

ここで，投稿作品が増加することによって，サイトの利用者である投稿者，読者それぞれの立場で問題が発生している．投稿者にとっては，自身の投稿した小説が，多くの小説の中に埋もれてしまい，読者の目に触れる機会が少なくなっている．一方、読者にとっては，多くの小説の中から自身の趣向に適合した小説を探すことが難しくなっている．

小説投稿サイトの多くでは投稿した小説にタグを設定する機能がある．タグには，サイト側が用意したタグ（システムタグ）と利用者（多くの場合は投稿者）が自由に設定できるタグ（自由タグ）の 2 種類がある．タグは，作品の持つテーマ，舞台，世界観などの特徴や作品に登場する人物や要素などについての情報を小説に付与する．検索する際，その対象となるのが主にタイトル，あらすじ，タグである．そのため，タグによって付与された情報が，読者が作品を検索する際の大きな手掛かりとなる．特に自由タグは投稿小説の内容を反映した記述が可能なため，適切な自由タグを小説に設定することは，検索支援に有効であると考えられる．しかし，タグの設定は投稿者に依存するため，タグを有効に活用できていないケースがある．例えば，極少数の限られた小説にのみ付与されている特定性が高く検索の対象となりにくいタグや多くの小説に付与されているため絞り込みができないタグといったようなタグが設定されているケースである．また，タグ単体では問題がない場合でも，小説の持つ特徴を他の小説と差別化できるようなタグが設定されていない，一つの小説に酷似した情報を持つタグが複数付与されている，といったようなタグを有効的に活用できていないケースも存在する．

本研究では，読者の検索を支援するという視点から，検索時の絞り込みに効果的なタグを小説投稿者に推薦する手法について提案する．この研究で推薦するタグは，自由タグより選別することとする．システムタグは，各サイトの仕様により差異が存在するが，選択形式などで感覚的に利用者が付与するかどうかを判断できるものや作品の内容を警告するために付与することが義務付けられているもの，TRPG（テーブルトークロールプレイングゲーム）を行ったときのセッション（ゲームの進行やプレイヤーの行動）を文章に起こした，リプレイや既存の創作物の世界観や人物などを利用して創作を行う，二次創作などの作品の分類を表すものである．以上から，投稿者に依存するわけではないと考えられるため，本研究では推薦の対象としてはとらえないこととする．

2 関 連 研 究

オンライン小説に関する推薦の研究には高田 [1] らの研究がある．オンライン小説の投稿数が増加することによって発生する問題に対して，高田らの研究では，読者に小説を推薦するというアプローチで問題解決を図っている．オンライン小説には，評価数やブックマーク数が少ない小説が多く存在することや著作の少ない投稿者や新規投稿者が多いこと，一般的に表紙や大きさなどの形状目的要因が存在しないことから，同著者の著作推薦，協調フィルタリングを用いた手法，表紙からの推薦など

1：https://syosetu.com/

既存の手法を用いることが難しいことを指摘している．そこで、高田らは立ち読みという選択手法に注目した．立ち読みを行う時，人々が考慮する特徴量として文章の言い回しと語義の 2 つを合わせて，文体と定義した．そして，文体を新たな小説推薦の指標として，ジャンルと組み合わせて推薦手法を提案した．高田らは，投稿者の支援と読者の満足度の 2 つの観点から評価を行っている．投稿者の支援としては，サイト内の評価レビュー数の少ない小説でも推薦可能なことを示した．読者満足度の観点からは，異なる著者でも文体類似度の高い小説の推薦は，同著者の著作を推薦されたときの満足度と同程度の高い満足度を獲得できることを示している．

推薦以外にものオンライン小説を対象としたさまざまな研究が行われている．飯田 [2] らは，オンライン小説のあらすじを Doc2Vec を用いてベクトル化し，Cos 類似度の総和を求めることによって，投稿されたオンライン小説の多様性について定量的に評価を行った．そこで，月別新規小説投稿数が増加するにしたがって，Cos 類似度の総和の平均値が増加していることから，オンライン小説の多様性が減少していることが定量的に明らかになった．

清水 [3] らの研究では，現状の小説投稿サイトが提供するランキングでは現在の人気作品しか示せないことを問題点にあげ，読者のつけたブックマークのリンク構造を利用して新たなランキング手法を提案し，将来ランキングに入る人気作品の予測を行っている．清水らの実験では，複数のジャンルで 2 か月後に人気ランキング上位になる小説の推定に成功している．

伊藤 [4] らの研究では，投稿されているオンラインの中には章や節で分けられていないものも多く，紙の書籍などに比べて，読者が一度に読む量を自身で決定することが難しいことを問題点にあげている．そこで，オンライン小説の本文を 1 つのテキストデータとして扱い，それを意味のある位置で区切る，段落分割の手法の提案を行っている．

実崎 [5] らの研究では，オンライン小説の人気度について，読者のつけたブックマークから人気度の推定を行おうとしている．

浦川 [6] らは，小説投稿サイト上に投稿された小説に設定されたキーワード（タグ）の出現頻度によって，小説のジャンルの流行について分析を行った．

オンライン小説以外を対象としたタグの推薦には，井上 [7] らの研究や Wang [8] らの研究がある．井上らは，SNS の投稿を対象に，ハッシュタグが付与された投稿をひとつの文書として扱い，文書の TF-IDF を求めクラスタリングし，推薦する投稿文書と TF-IDF ベクトルの比較を行い，ハッシュタグのクラスタの推薦を行っている．Wang らの研究では，SNS 上で投稿される画像の類似度だけでなく，投稿の人気度と投稿ユーザの人気度，タグの人気度を考慮し，タグランキングを作成した．作成したタグランキングをもとに推薦を行うことで，従来のタグ推薦よりも閲覧数を伸ばす推薦が可能であるという結果を示している．

本研究は，投稿サイト上のオンライン小説の本文とタグの類似度を比較することで，オンライン小説の投稿者に検索に有効なタグの推薦を行う．オンライン小説の小説で，上記の高田らのような小説を推薦する研究はあったが，投稿者にタグを推薦する研究は見当たらない．小説をベクトル化するときに，Doc2Vec を用いて小説本文を学習させることで，オンライン小説本文の意味を考慮することに重きを置いているという点で特徴的である．

3 検索に有効なタグを推薦する手法の提案

本研究では，小説投稿サイトへと投稿されるオンライン小説に対して，検索に有効なタグの推薦を行う手法について提案する．ここで，検索に有効なタグとは，設定された小説の内容に合致した情報を持つこと，設定されたタグの間でタグの表す情報に重複が生じないこと，小説投稿サイト上の全ての小説数に対してそのタグが小説に設定されている数が十分に小さいこと，の 3 つの要件を満たしているタグとする．

3.1 提案手法の概要

ここでは，提案手法全体の概要を説明し，それぞれの詳細については，各節で説明を行う．まず，オンライン小説の本文とそれに設定されたタグのデータの収集を行う．次に，収集したタグのデータから，小説の数に対して一つのタグが小説に設定された数が十分に小さくなる範囲を求める．そして，本文と抽出したタグを用いて，ベクトル化されたタグデータベースを作成する．最後に，タグデータベースを用いて，オンライン小説とタグの類似度を比較し，タグの推薦を行う．

3.2 データ収集

小説投稿サイトよりオンライン小説の本文およびオンライン小説に設定されたタグの収集を行う．本研究では，オンライン小説の完結、連載中を問わずに収集を行う．また，短編小説と連載小説などの区分は行わない．連載小説やオムニバス形式の小説では，話，章，節を結合し，一つの小説本文として取り扱うこととする．本研究では，前述した通り自由タグのみを推薦の対象とする．

3.3 検索に有効なタグ範囲の抽出

有効なタグとは，小説投稿サイト上に投稿されているすべてのオンライン小説の件数に対して，そのタグが小説に設定されている数（タグの出現数）が十分に小さいタグである．また，全小説中 1 件に設定されていないタグなどは，投稿者に依存している度合いが高く，読者が検索する際に検索の対象外となるタグであると考えられる．したがって，タグの出現数が小さすぎるタグは，検索に有効なタグとはいえない．以上のことを踏まえて，検索に有効なタグの範囲を求めていく．手法としては，実際に収集したタグの出現数の分布を作成する．システムタグは，小説投稿サイトが公式に用意したため，投稿者が感覚的に選択し，小説に設定できるという特徴を持つ．そのため，多くのシステムタグはタグの出現数の上位に位置すると推測できる．この分布を参考にし，タグの出現数の分布からシステムタグのものを取り除いた分布を作成し，比較を行う．システムタグと同様な位置に分布している自由タグは，サイト利用者にとって

は非常に一般的なタグであり，推薦する必要性は低いと推測されるため，取り除く．また，この分布はタグ出現数が十分に小さいとは言えないタグであると推測される．出現数が小さすぎると考えられるタグも，実際の分布と参考にし，検索に有効なタグの範囲を求め，タグの抽出を行う．

3.4 タグベクトルリストの作成

タグは小説の世界観やテーマ，登場人物の持つ要素，全体の展開，大ジャンルやジャンルで説明しきれていないより詳細な分類等の情報を表すものである．したがって，類似した情報を持つ小説同士では，同様なタグが設定される．しかし，投稿された小説は非常に多いため，直接小説同士の類似度を行うという手法では，推薦を行うたびに類似度計算に時間を消費することになり，入力から推薦までに大きなラグが生じることになる．そのため，提案手法では，小説をもとにタグのベクトルを作成し，リストに格納することで，実際に処理を行う時の計算量を減らし，効率化を図る．

3.4.1 オンライン小説本文のベクトル化

文書の分散表現を可能とする Doc2Vec を用いて，オンライン小説の本文をベクトルに変換する．はじめに Doc2Vec に入力する学習データを作成する．事前に収集したオンライン小説の本文を形態素解析エンジンである MeCab を通して形態素解析を行う．今回，本文から取り出す単語は名詞，形容詞，動詞，形容動詞の 4 つの品詞に該当するものである．数詞や助動詞などの 4 つの品詞に該当しない単語はストップワードとして除外した．また小説の特徴として，文章中には多くの固有名詞を含む．しかし，これらは形態素解析の辞書には記載されていない．そのため，名詞が連続した際などは一つの固有名詞として単語を連結する処理を行う．例えば「暗黒騎士」という言葉が本文中に現れたとする．MeCab でそのまま処理を行うと「暗黒」と「騎士」と二つの名詞に分類される．しかし，小説本文中で「暗黒騎士」は一つの固有名詞であるため,「暗黒」と「騎士」という二つの語としてとらえるときとは，異なる性質を持つと考えられる．したがって，前述した方法で結合を行い一つの単語として処理する．作成した学習データを Doc2Vec に適用し，学習モデルを作成する．学習モデルを利用し，各オンライン小説本文のベクトル化を行う．ベクトルの次元数は 300 とした．

3.4.2 タグのベクトル化

タグは設定されたオンライン小説の世界観やテーマなどの情報を示すものである．したがって，タグは設定された小説と同様な性質を有していると考えられる．以上のことから本研究で提案する手法では，タグは設定された小説の持つベクトルの性質を持っていると仮定する．そこで，タグ$\overrightarrow{T}$は，そのタグが設定された i 個の小説$\overrightarrow{N}$を合成したものとし，以下の式とする．

$$\overrightarrow{T} = \overrightarrow{N_0} + \overrightarrow{N_1} \cdots + \overrightarrow{N_i} \tag{1}$$

計算より求められたタグのベクトルをリストに格納していき，タグベクトルのリストを作成する．

3.5 推薦するタグの提示

3.4.1 節で作成した Doc2Vec の学習モデルを利用して，推薦の対象となるオンライン小説の本文（入力文書）をベクトル化する．ベクトル化したオンライン小説を 3.4 節で作成したリストに格納されたタグベクトルと Cos 類似度を用いて比較を行う．入力文書のベクトルを$\overrightarrow{d}$，比較するリスト内のタグベクトルを$\overrightarrow{t}$とすると，Cos 類似度は以下の式で求められる．

$$\cos(\overrightarrow{d}, \overrightarrow{t}) = \frac{\vec{d} \cdot \vec{t}}{|\vec{d}||\vec{t}|} \tag{2}$$

Cos 類似度は −1 から 1 の範囲で表される．より 1 に近い値のものが，入力した小説に類似したタグである．よって，小説本文の持つ情報に合致したタグを推薦するときは，Cos 類似度が 1 に近い順にタグの推薦を行えばよい．しかし，そのまま類似度が高い順にタグの推薦を行うと，タグ同士でも類似度が高いタグが推薦されると考えられる.. タグ間で類似度が高いということは，同じような情報を持っているということになる．タグ間で意味が重複しては，推薦前の状況と変化なく，検索に有効なタグであるとはいえない．したがって，文書と高い類似度を持つが，タグ間では異なった情報を持つタグを検索に有効なタグとして推薦する．

以上のことを考慮し推薦するタグを決定していく．まず，入力した小説と Cos 類似度の高い順に N 件を推薦候補タグとして抽出する．入力した小説 $\overrightarrow{D}$ は $\overrightarrow{T_0}$ とする．推薦するタグを第 1 位から $\overrightarrow{T_1}$，$\overrightarrow{T_2} \ldots \overrightarrow{T_N}$ まで順位付けする．

$$\begin{aligned} \overrightarrow{T_0} &= \overrightarrow{D} \\ \overrightarrow{T_1} &= \arg\min_i |\overrightarrow{DT_i}| \end{aligned} \tag{3}$$

タグ $\overrightarrow{T_{i-1}}$ が定められいる時，既に推薦されたタグと異なるタグを推薦するために D を挟んで直近に推薦された 2 つのタグの中点と反対にあるタグ $\overrightarrow{T_i}$ を求める方法を図 1 に示す．図中の $\overrightarrow{T_i}$ は、$\overrightarrow{T_{i-2}}$ と $\overrightarrow{T_{i-1}}$ の中点ベクトルである．

$$\overrightarrow{T_i} = \frac{\overrightarrow{T_{i-2}} + \overrightarrow{T_{i-1}}}{2} \tag{4}$$

推薦するタグはすでに推薦したベクトルと異なることが望ましいため，$\overrightarrow{T_iD}$ 上にあり，小説と高い類似度を持つタグの推薦を行いたい．しかし，実際には $\overrightarrow{T_iD}$ に推薦するべきタグがあるとは限らない．そのため，$\overrightarrow{T_i}$ を $\overrightarrow{T_iD}$ 上に投影した $\overrightarrow{T_{i'}}$ を求める．

$$\begin{aligned} \overrightarrow{T_i'} &\leftarrow \arg\min_i |\overrightarrow{DT_{i'}}| \\ \overrightarrow{DT_i} &= \frac{|\overrightarrow{DT_{i'}}|}{\cos(\frac{\theta_{i'}}{2})} \end{aligned}$$

(5)

θ_j は $\overrightarrow{T_iD}$ と $\overrightarrow{T_i}$ とのなす角度である。θ の範囲は $0 \leqq \theta \leqq \pi$ となる．

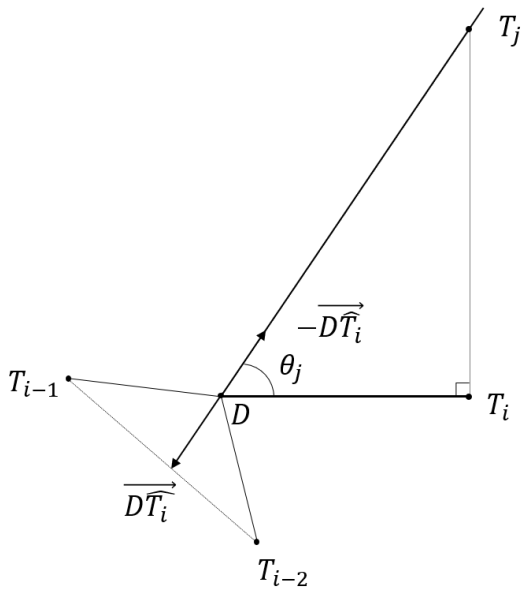

図 1　推薦するタグ $\overrightarrow{T_i}$ の図解

4　実験と評価

4.1　データセット

ヒナプロジェクト社が運営する，日本の小説投稿サイトの 1 つである「小説家になろう」を対象にオンライン小説の本文とタグの収集を行った．

4.1.1　小説本文の収集

「小説家になろう」では，小説を恋愛，ファンタジー，文芸，SF，その他という 5 つの大ジャンルに区分している．さらに大ジャンルの下で，それぞれより詳細な区分が設けられており，合計 20 のジャンルが設定されている．本研究では，大ジャンルである恋愛の中に区分される異世界ジャンルに該当する小説 23,036 件を収集した．本文の収集には，Beautiful Soup を用いた．Beautiful Soup とは，HTML や XML ファイルからデータを抽出することができるスクレイピングに特化した Python ライブラリである．

4.1.2　タグの収集

本研究でタグと呼ぶものは,「小説家になろう」上でのキーワード，登録必須キーワードに該当する．キーワードは，以下の 8 つに分けられる．

- 公式シナリオ用キーワード
- リプレイ用キーワード
- 二次創作キーワード
- 管理キーワード
- おすすめキーワード
- 公式キーワード
- 企画キーワード
- 手動入力キーワード

キーワードは，10 字以内で記述され，1 つの作品に合計 15 個まで設定することができる．また，キーワードを設定できるのは，その小説の投稿者だけである．本研究で定義する自由タグには，手動入力キーワードが該当し，それ以外の 7 種類のキーワードがシステムタグに該当する．手動入力キーワードは，前述した字数制限以外に制約なく記述が可能である．登録必須キーワードとは,「小説化になろう」が指定している 6 つの要素を含んだ小説を投稿する場合に，設定が要求されるタグである．6 つの要素とは，R15，ボーイズラブ，ガールズラブ，残酷な描写あり，異世界転生，異世界転移，である．登録必須キーワードは，手動入力キーワードを設定するキーワード欄に入力することでも，設定が可能であるがキーワードとは別個にとらえられるため，キーワード設定数には含まれない．そのため，タグは一つの小説に 21 個設定できる．本研究では，登録必須キーワードはシステムタグとして，他のシステムタグに該当するキーワードと同様に扱った．

タグの収集には,「小説家になろう」を運営しているヒナプロジェクト社が提供している開発ツール「なろうデベロッパー」から「なろう小説 API」[2]を用いて収集を行った.「小説家になろう」全体の傾向を把握するため，収集した本文以外に設定されたタグの収集も行った．2019 年 6 月 27 日に，661,495 件の小説に設定された 3,833,470 個のタグを収集した．タグの種類は，319,277 種類である．平均すると 1 つの小説には 5.80 個のタグが設定されている．実際に小説に設定されたタグの個数を示したものを図 2 に示す．図 2 から読み取れるように，タグが 4 つ設定された小説が最も多いことがわかる．そして，4 を最大とし，それより多くのタグが設定されたもの，それより少なくタグが設定されたものはそれぞれ上限，下限に接近するにつれて順に減少していることが読み取れる．以上の値を参考にし，実験では 1 つの小説に 5 つのタグを推薦することにする．

4.2　検索に有効なタグ範囲の抽出

学習データとなるオンライン小説本文に設定されたタグの出現頻度を図 3 に示す．図 3 を参照すると，タグの出現頻度はおよそべき乗に分布していることが読み取れる．また，システムタグを含むタグすべての分布と自由タグのみの分布を比較すると，システムタグが上位を占めていることを読み取ることができる．

ここで，4.1 で収集したおよそ「小説家になろう」に投稿されたオンライン小説に設定されたすべてのタグについてタグの出現頻度を表した図を図 4 に示す．図 4 を参照すると，学習データの小説に設定されたタグの分布図 3 と同様にタグの出現頻度はおよそべき乗になっていることが読み取れる．システムタグを含むタグすべての分布と自由タグのみの分布を比較すると，システムタグが上位を占めていることを読み取ることがで

2 : https://dev.syosetu.com/man/api/

きる．このことから，図 3.3 で行った推測がおよそ妥当であったことが確認できる．

今回，実験では学習データに設定されたタグの出現数が 100 ～ 1,000 を検索に有効なタグである範囲として定める．該当する範囲のタグ 901 個を抽出する．

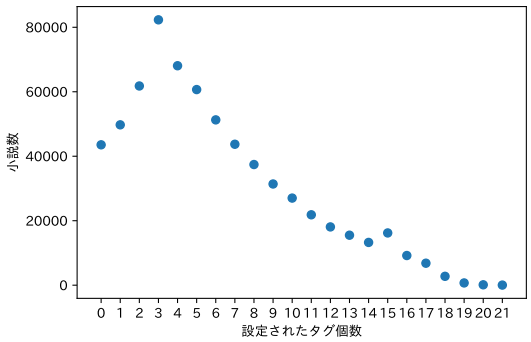

図 2　1 つの小説に設定されているタグ数

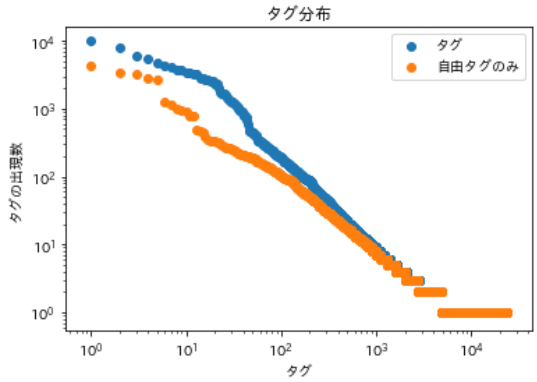

図 3　学習データのタグ分布

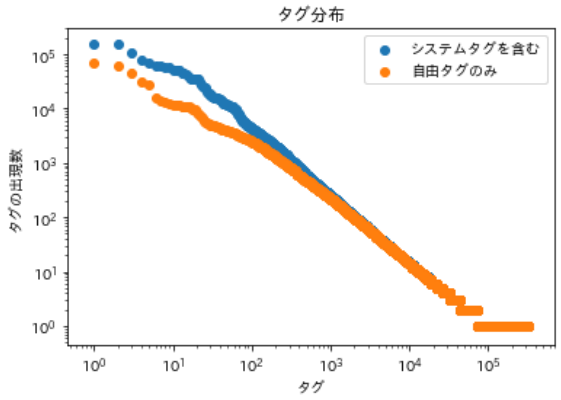

図 4　収集したタグ全体の分布

4.3　タグベクトルリストの作成

4.2 節で抽出したタグもベクトル化を行うために，学習データとなる小説本文の選定を行う．小説の定義は，著者の心情を自由に書き記した散文である．散文は．非常に自由度が高いものであるため，本文の大半部分が顔文字や記号などでで記されていたり，意味がない語の連続であったりすることがある．このような小説は，その他の小説と文章の形式が大きく異なっているため，取り除く必要がある．また，収集する際には字数などに指定を設けていないため，本文が 1 行に満たないものなど，非常に文量が少ない小説などが見られた．これらの小説からは，ベクトル化するときに特徴量が十分でないと考えられ，学習データとして不適だと考えられる．そのため，本研究では，収集した小説の内本文が 1,000 字以上の小説のみを抽出した．2 からもわかるように，小説にはタグが 1 つも設定されていないものも含んでいる．タグの推薦を行う際に，タグが設定されていない小説を用いることは，推薦の精度を下げることにつながる可能性があるため，タグが設定されていない小説も学習データから取り除いた．以上の前処理を行い，学習データには 20,962 件の小説を用いた．

作成した Doc2Vec の学習モデルから，学習データの小説本文の 300 次元のベクトルを生成する．生成した小説本文のベクトルをもとに，4.2 節で抽出した自由タグのベクトル化を行う．ベクトル化したタグをリストに格納した．ベクトルはすべて正規化し，単位ベクトルとした．小説本文とタグを結びつける ID には，「小説家になろう」で小説一つ一つにつける固有の ID である N コードを利用する．

4.4　タグの推薦

4.3 節で，作成したタグのリストを用いて，推薦対象のオンライン小説と比較を行い検索に有効なタグの推薦を行う．推薦対象として入力するオンライン小説は，4.1 節で学習データとなる小説本文を収集した方法と同様な手順で，「小説家になろう」に投稿されている大ジャンルの恋愛，ジャンル異世界に属するオンライン小説とし，学習データと同様に，特徴量が十分であるように，本文が 1,000 字以上のものとした．上記の条件で収集したオンライン小説 158 件を実験データとして用いた．Doc2Vec の学習済みモデルを使用し，MeCab で形態素解析した 158 件分の小説本文をそれぞれベクトル化した．

3.5 節で，説明したタグの意味の重複を防ぐことを意識した提案手法でタグの推薦を行う．ベースラインの手法として，タグを推薦する小説とタグのベクトルの Cos 類似度が高い順にタグを推薦する手法での推薦も行う．

4.4.1　評 価 実 験

提案手法とベースラインそれぞれの手法で推薦されたタグと小説本文の Cos 類似度を求め，実験データすべての平均を求めた．それをまとめたものが表 1 である．Cos 類似度は，1 に近いほど類似しているといえる．そのため，1 に近いほど推薦されたタグは実際の小説の内容と合致していると考えられる．ベースラインと比較して，提案手法で推薦したタグの類似度の平均値は総じて低いことが読み取れる．また，提案手法とベー

スラインの手法ともに類似度は全体的に低く，推薦されたタグは検索に有効なタグといえない小説の内容と異なるものが推薦されている可能性がある．

表 1　推薦したタグと小説の Cos 類似度の平均

	提案手法	ベースライン
1	0.395	0.395
2	0.151	0.374
3	0.161	0.364
4	0.161	0.358
5	0.155	0.352

そこで，推薦されたタグが小説の内容と合致しているのかをそれぞれ比較した．小説の内容に合致しているタグとして，実際に小説に投稿者の手によって設定されたタグと比較をする．実験データの 1 つである『魔導人形物語～左手の薬指にはあなたの指輪を～』を例に比較する．それぞれ『魔導人形物語～左手の薬指にはあなたの指輪を～』に推薦されたタグと実際のタグを表 2 に表す．実際に設定されているタグは 8 個であり，すべてが自由タグである．表 2 を見ると，どちらにも人形という同一のタグが設定されていることがわかる．また，提案手法によって推薦されたタグには愛，実際のタグには恋愛とある．これらのタグは，どちらも作品の内容，要素に恋愛や愛がテーマになっていることを指すタグである．このように，提案手法によって推薦されたタグの中には，作品の内容に合致しているタグも見られた．しかし一方で，幻獣やオッサンの 2 つのタグは作品の内容とは関係がないタグの推薦も行われていた．

表 2　『魔導人形物語～左手の薬指にはあなたの指輪を～』に推薦されたと実際のタグ

提案手法	実際のタグ
人形	人形
幻獣	職業モノ
愛	恋愛
オッサン	王道ヒロイン
オリジナル	幼なじみ
	三角関係
	主人公以外病んでる
	ブクマ５０で続編書く

次に，推薦されたタグ同士で持つ情報が異なっているかを検証するために，それぞれの手法で推薦されたタグの間で Cos 類似度を求め，1 つの小説に推薦されたタグ間での Cos 類似度の平均値を求めた．それぞれ求めた平均値から実験データすべてを合算して，平均値を求めた．その結果を 3 に記す．提案手法の方がベースラインの手法よりも値が小さくなっていることがわかる．このことから，提案手法はタグ間で意味の重複が少ないタグを小説に推薦できていることがわかる．

表 3　推薦タグ間での Cos 類似度の平均の平均

提案手法	ベースライン
0.501	0.639

これらの実験により，提案手法によって，設定されたタグの間で情報に重複が生じないとういう検索に有効なタグの 1 つの要件を満たすことができた．また，2 の表にある職業モノというシステムタグの 1 つである職業ものの表記ゆれのタグや主人公以外病んでるという絞り込み過ぎるタグを推薦対象から除いたことで，小説投稿サイト上の全ての小説に対してそのタグが小説に設定されている数が少ないタグという要件も満たしている．しかし，一部では小説の内容と合致したタグを推薦できたが，内容に関係ないタグも推薦された．また小説と推薦されたタグの Cos 類似度が全体的低くなった．このことから，提案手法では小説の内容と合致したタグの推薦には課題が残ったといえる．

5 考　　察

推薦されたタグのベクトルと小説本文のベクトルの類似度が低くなった原因として，タグのベクトルがそのタグの持つ特徴を表すことができていなかった可能性が考えられる．提案手法では，タグが設定された小説と同様な性質を持つという仮定の下で，小説の本文から作成したベクトルをもとにタグのベクトルを作成した．小説の本文からベクトルを作成する際に，タグの性質を表す特徴量の抽出ができていなかったと推測される．今回，学習データとして用いた小説は特徴量が十分あるように 1,000 字以上の小説と設定したが，実際に小説の文字数をしらべると，1,000 字に近いものから 100,000 字を超えるものもあり，小説ごとにさまざまであった．これらの小説を一様に取り扱ったため，正確な特徴量の抽出に失敗したと考えられる．小説の特徴量を正確に抜き出す方法としては，文章の量ごとにクラスタをわけて，学習を行いベクトル化をする，小説本文ではなく，飯田 [2] らのようにあらすじを小説の特徴を表すものとして扱いあらすじをもとに Doc2Vec で学習を行う，小説を話や章などの区切り一部もしくは場面ごとにわけて特徴量を求める方法などが考えられる．また，小説の情報というものは場面ごとに切り替わるものなので，その小説の特徴的な場面を切り抜くことは，より正確な情報をもったベクトルを生成する助けなるといえる．

小説の特徴量を求める際に，小説の内容や意味を考慮に入れるために，今回は Doc2Vec を用いたが，より精度が高いと考えられる Sent2Vec などの手法を用いることによっても特徴量の抽出の改善につながると考えられる．ここで，今回推薦対象として選択した大ジャンルは恋愛である．そのため，小説のテーマが愛であることは自明であり，話の要素として恋愛を含むことも自明であるといえる．よって，大ジャンルとタグの意味で重複してしまっているため，これらのタグは大ジャンル恋愛に属する小説には推薦する必要性が薄いといえる．タグを推薦するときに，タグ同士での意味の重複だけではなく，大ジャンルやジャンルなどの要素にも考慮することでより検索に有効なタグを求めることができると考えられる．

4.2 節より，今回用いたデータセットのタグ分布がべき乗になっていた状態，と同様な傾向が「小説家になろう」サイト全

体のタグの分布には当てはまる．本研究では，検索に有効なタグの推薦は行えなかったが，推薦するタグの精度を向上することで本研究で用いた手法は，他の大ジャンルやジャンルに属する小説，サイト上のすべての小説に適用できると期待される．

これらの小説にタグを推薦することが可能になれば，検索をよりスムーズに行うことができると推測できる．

6 おわりに

本研究では，小説投稿サイトへと投稿されるオンライン小説に対して，小説本文のベクトルとタグのベクトルを比較することで，検索に有効なタグを推薦する手法について提案を行った．

具体的には，「小説家になろう」から収集した小説の本文をDoc2Vec で学習を行い，学習モデルを作成し，そのモデルを用いて小説の本文をベクトル化した．小説に設定されたタグは小説と同様な性質を持つという仮定のもと，小説のベクトルからタグにベクトルを付与し，タグベクトルのリストを作成し，推薦する小説とリスト内のタグベクトルで類似度比較を行った．さらに推薦するタグ同士で意味が重複することを防ぐために，タグ間での類似度の比較も行った．実際に，小説投稿サイトの1 つである「小説家になろう」に投稿されているオンライン小説とタグを用いて，提案手法の実験を行ったが，推薦されたタグは小説との類似度が低く，小説の持つ特徴を表すタグの推薦としては不十分な側面が残った．タグ間で類似度の比較を行い，類似度の平均値を求めることで，タグ間の重複を防げているか検証を行った．提案手法の方が類似度が低い傾向にあることから，提案手法の有用性が見られた．

今後の課題としては，5 節で挙げた改善する方法を参考に小説の特徴量を求め，タグに正しい情報を付与できるようにすることで提案手法の実現に近づけたい．また，推薦されたタグが実際に検索に有効的に働くかの評価を行うことや推薦するタグを求める式の改善などが今後の課題である．

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] 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*, No. B5-2, pp. 1–8, 2017.
[2] 飯田委哉, 伊東栄典, 佐嘉田悠樹. クラスタリングによるオンライン小説の多様性動向分析. 火の国情報シンポジウム論文集, Vol. 2018, pp. 1–7, 2018.
[3] 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, pp. B–3–2, 2013.
[4] 伊藤志暢, 松村敦, 宇陀則彦. 語の結束度と感情を考慮したオンライン小説の段落分割手法の提案. 情報処理学会第 81 回全国大会, Vol. 4, p. 03, 2019.
[5] 実崎直人, 伊東栄典. 回帰分析によるオンライン小説の人気度推定. 情報処理学会第 81 回全国大会, Vol. 2019, No. 1, pp. 333–334, 2019.
[6] 浦川隆寛, 伊東栄典. オンライン小説におけるキーワードの時系列傾向分析. 情報処理学会研究報告, pp. B–3–2, 2013.
[7] 井上優作, 若林啓. 表記の多様性を考慮したハッシュタグ推薦. 第 14 回日本データベース学会年次大会, 2016.
[8] Xueting Wang, Yiwei Zhang, and Toshihiko Yamasaki. User-aware folk popularity rank: User-popularity-based tag recommendation that can enhance social popularity. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1970–1978. ACM, 2019.