小説の特徴量を用いたオンライン小説の検索ワード推薦手法の提案

山崎　睦月†　　佐藤　哲司††

† 筑波大学大学院 人間総合科学学術院　〒 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　オンライン小説投稿サイトでは．投稿小説の増加に伴い，ユーザが自分の読みたい小説を探すことが難しくなってきている．多くの小説投稿サイトでは，小説本文だけでなく小説のテーマや世界観を表すタグも合わせて投稿されている．本論文では，小説とタグの 2 部グラフ構造を用いて小説の特徴量からタグの特徴量を算出し，読了した小説の特徴を反映したタグを推薦する手法を提案する．推薦されたタグの中からユーザが選択した検索ワードを用いた評価実験を行い，提案手法の有効性を検証した．

キーワード　検索クエリ，推薦モデル，オンライン小説

1　はじめに

オンライン小説を専門に扱う小説投稿サイトの登場・発展によって，誰もが容易にインターネット上に自身の書いた作品を投稿できるようになった．投稿された小説はウェブサイトの利用が可能な状態であれば，いつでも読むことができ，多くの人々がサイトを利用している．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 8 か月間の間に約 29,000 件の小説が投稿され，登録ユーザ数は約 114,000 人増加している．[2] 登録ユーザは小説の投稿，ブックマーク機能などのサイト利用が可能である．登録ユーザ以外でも，小説を読むことはできるため，より多くの人々がサイトを利用していると推測される．

小説投稿サイト上に投稿された小説が増加することは，サイトのさらなる発展や利用ユーザの獲得などのメリットがあると考えられる．他方で，小説の増加に伴い，サイト上の小説を利用するユーザ，読者は非常に多くの小説の中から自分の趣味嗜好に適合した小説を探索する傾向が高まっている．

小説投稿サイトでは，読者が小説を探しやすくするために様々な工夫がされている．例えば，小説のジャンルをわかりやすく掲示する方法や，サイト内の小説全体やジャンル別で PV（閲覧数）や評価などをもとに，週間，月間など期間に分けてランキングを作成し掲示する方法などがあげられる．これらの方法は，読者のオンライン小説探索の支援になると考えられるが，読者の小説探索の支援には十分であるとは言えない．前者の方法では，ジャンルだけを示すだけではジャンル内での小説の絞り込みが更に必要であり，小説の増加に対する根本的な解決には至らない．後者の方法では，ジャンル別で人気のあるもしくは評価されている小説を探すことには適している．しかし，投稿されて間もない小説や閲覧数や評価の少ない小説は当然探索の候補から外れてしまう．また，期間ごとにランキングは分けられているものの，同じ小説が上位にランクインしているケースも多くみられる．よって，読者が自分の趣味嗜好にあった小説を探す方法にはランキングの手法だけでは十分でないと考えられる．

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明示的に付与するものである．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（多くの場合は著作者）が自由に設定できるタグ（自由タグ）の 2 種類がある．小説投稿サイトで，キーワード検索をするときにその対象となるものは，主にタイトル，あらすじ，タグである．タイトルはやあらすじだけでは，小説の特徴的な要素を十分に示すことは難しい．タグを設定することは，小説のタイトルとあらすじで足りない情報を補完することができる．そのため，タグは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，著作者の手によって投稿小説の内容をより詳細に反映した記述が可能なため，自由タグを活用して小説の探索を行うことは，読者の嗜好に沿った探索に効果的であると考えられる．

実際に小説投稿サイトでも，このタグを活用した小説の検索は推奨されている．「小説家になろう」では，人気のキーワード（上記のタグに相当するもの）として過去 60 日のうちに投稿小説に設定されたキーワードを掲示し，そこから小説を探せるようになっている．一方で，一部の表示されるキーワードには「ざまあ」や「もう遅い」などといったサイトのライトユーザでは理解が難しい，一般的な語ではないものも散見される．

本研究では，小説投稿サイトを利用する読者の小説検索支援を目的として，読者が読了した小説の本文をもとに，その小説と類似した要素を持った小説を検索するために効果的に機能するタグを推薦する手法について提案する．読者は推薦されたタグを選択し，検索ワードとして用いることができる．

推薦するタグは，自由タグからのみ選出する．これは，シス

1：https://syosetsu.com/
2：2022 年 4 月 8 日では 657,825 件だった投稿数が同年 12 月 6 日には 995,423 件に増加し，登録ユーザは 2,245,258 人から 2,358,933 人へと増加している．

テムタグは，サイト側が用意したタグであるため，非常に一般的な語であると考えられ，読者に推薦する必要性が少ないためである．

2 関連研究

多くの人々が利用する小説投稿サイトやオンライン小説について多くの研究が行われている．オンライン小説を推薦する研究の 1 つとして，高田 [1] らの研究がある．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既存の手法を用いることが難しいことを指摘している．この研究では，立ち読みという小説の選択手法に注目している．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た推薦手法を提案している．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，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，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している．

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．

オンライン小説以外を対象としたタグの推薦には，井上 [4] らの研究や，Wang [5] 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，ユーザが読了したオンライン小説と類似した特徴を持つ小説の探索を支援するタグの推薦を行う．小説の特徴量とタグの特徴量の類似度を直接比較をして推薦するタグを選ぶだけでなく，タグ同士で類似度の比較を行うことで，検索に効果的なタグを推薦するというところに本研究の新規性がある．

3 提案手法

本研究では，小説投稿サイトに投稿されている小説を読んだユーザが，その小説と類似した要素を持つ小説を検索するときに効果的な検索ワードを推薦する手法について提案を行う．

提案手法の基本的なアイデアは，過去に会議で発表 [6] [7] しており，本論文は，基本アイデアを精緻化するとともに，より発展した内容となっている．

3.1 概　　要

以下では，提案手法全体の概要について説明を行い，それぞれの詳細については，各節で説明を行う．まず，小説とその小説に設定されたタグの対から，タグの出現回数を基にタグの情報量を求める．求めたタグの情報量から，情報量の極めて小さいものと情報量の極めて大きいものを推薦の候補から取り除き，推薦するタグ候補の範囲として設定する．次に，小説本文の特徴量を生成し，その小説本文の特徴量を基に推薦候補の範囲にあるタグの特徴量を生成する．最後に，入力となる小説本文に対して，検索に効果的なタグの推薦を行う．検索に効果的なタグの選出には，上記で求めたタグの情報量と類似度を用いて行う．

本提案手法によるシステムは，以下のような文脈のシナリオにおいて活用されると想定している．まず，ユーザが小説投稿サイトの 1 つのオンライン小説を読み終えたと仮定する．ユーザはその小説の内容に，面白い，感動したなどの感情を得たり，興味関心を覚えたりし，新たにその小説と類似した小説を読んでみたいと考える．しかし，小説投稿サイトに存在する数多くの小説の中から探索を行うことは困難である．そこで，ユーザの探したい小説にマッチした検索キーワードが必要になる．しかし，ユーザが自分の力で見つけ出すことは，小説投稿サイトのコンテキストを理解している必要があるが，日々新たな小説が投稿されることから正確なコンテキストを把握することは困難である．この時，システムに読了済みの小説を入力することで，その小説と類似した小説が効果的に検索されるタグを複数推薦する．ユーザは，この推薦されたタグを選択して検索に用いることで，自分の読みたいと考える小説が探しやすくなると考えられる．

本研究における効果的な検索タグとは，以下の 3 つの要件を満たすタグである．

- 入力された小説の内容と合致したタグ
- 推薦されたタグ同士でタグの表す情報に重複がないタグ
- 複数のタグで絞り込みができるタグ

上記の要件を満たすタグとする理由をそれぞれ説明する．1 つ目の「小説の内容と合致したタグ」という要件は，小説の内

容に合致するすなわち，その小説もしくはその小説に類似した小説にはそのタグと同じ内容が含まれていると考えられるためである．そのため，推薦されるタグは，ユーザが入力した小説と似ている必要がある．2 つ目の「推薦されたタグ同士でタグの表す情報に重複がないタグ」というものは，推薦されたタグ同士で内容が異なっているほうが検索の際により効果的に絞り込みが行えると考えられるからである．小説とは，1 つの作品中においても，作品内での場面や時間における変化など，様々な要素を持っていると考えられる．そのため，検索のためのタグを推薦するときに，同じ要素を示すタグが複数推薦されても，それで示すことができるものは，その小説の一部分に偏ってしまう．推薦するタグが異なる内容を表すようにすることで，小説の含む要素をまんべんなく示すことができる．また，推薦されたタグの中からユーザが特に好ましく考える要素を選出して検索に活用することができる．3 つ目の「複数のタグで絞り込みができるタグ」は，単一のタグでは絞り込みが難しいケースが想定されうるためである．タグの中には，多くの小説に設定された出現数が大きいものがある．そのため，仮に 1 つだけタグを推薦するとき，出現数が大きいタグが推薦されると，そのタグだけでは十分な絞り込みが行えず，結局多くの小説の中から類似した小説を探すことになり，ユーザの検索を支援するという目的を達成することが難しくなる．そこで複数のタグを推薦することで，絞り込みが可能な検索を実現する．検索ワードによる探索はユーザによって行われるため，ユーザの負担が大きくなり過ぎないような絞り込み件数が望ましい．よって，複数の推薦されたタグを用いた検索ワードで検索した時に，数十件程度の小説に絞り込めることが必要である．

読了した小説と類似した別の小説を直接推薦するのではなく，検索するためのタグを推薦する理由の 1 つは，計算量を抑え効率化を図るためである．オンライン小説は，日々内容を更新していく連載作品が多い．また，新たな作品が頻繁に投稿されている．これらをすべて推薦する候補とするために，毎回類似度比較のために計算をすることは非効率的であると考えられる．そこで，本論文では，事前にタグの特徴量を算出し，推薦時の計算量を削減できると考えた．もう 1 つの理由は，タグを用いることで読了済みの小説と全体的に類似している小説でなく，部分的に類似した小説を探すことが可能となるためである．1 つの小説の中でも，小説は様々な側面を持っている．そのため，ユーザが好みとする部分は小説の局所的な部分であることが考えられる．タグの推薦によって，ユーザの好みによる偏りを反映した小説の検索が可能となる．

3.2 タグの情報量

ある事象が発生したとき，その事象がどの程度まれに発生するかを示す尺度を情報量という．事象 e が発生する確率が $P(e)$ のとき，その事象の情報量 $I(e)$ は以下の式で示される．

$$I(e) = -\log P(e) \tag{1}$$

したがって，事象が発生する確率 $P(e)$ が低くなるほど，事象 e の情報量 $I(e)$ はより大きな値となる．逆に，その事象 e が頻繁に発生し確率 $P(e)$ が高くなるにつれて，事象 e の情報量 $I(e)$ は小さくなっていく．

ユーザが入力した情報量を調整した研究に文献 [8] がある．Ou らはオンラインショッピングサイトでのクエリの出現という事象を確率的な事象とし，検索クエリのワードの情報量を求めている．この手法を参考に，本研究では，この情報量の概念を小説投稿サイト内でのタグの出現数とし，タグの情報量を求める．タグが出現する，すなわち 1 つの小説に設定されることを 1 つの事象として扱う．タグ t が 1 つの小説に設定されるという事象が起きたとき，情報量 $I(t)$ は以下の式で表せる．

$$I(t) = -\log P(t) \tag{2}$$

タグの出現確率 $P(t)$ は，タグ t が出現する回数 $n(t)$，すべてのタグの総出現数 N としたとき，以下の式で求められる．

$$P(t) = \frac{n(t)}{N} \tag{3}$$

タグの出現回数の値が大きくなるとそれに伴い，タグの情報量は小さくなる．逆に，タグの出現回数の値が小さくなるほど，タグの情報量は大きくなる．

タグの情報量が小さくなるとき，このタグは非常に多くの小説に設定されている，普遍的な性質を持ったタグであるといえる．このようなタグは普遍的であるゆえに検索で絞り込みが難しいと考えられる．また，普遍的なタグは，小説投稿サイトを利用するユーザにとって，既知なタグであると想定される．これらのことから，情報量の極めて小さいタグは，ユーザに推薦する候補とはならない．逆に，タグの情報量が非常に大きくなるとき，そのタグはごくわずかな小説にのみ設定されているものであり，特定性の強いという性質を持っている．このようなタグは，検索に用いてもそのタグが設定されているごくわずかな小説のみがヒットし，新たな作品の発見には適していないと考えられる．したがって，情報量の極めて大きなタグも推薦するタグの候補とはならない．以上のことから，タグの情報量をもとに推薦するタグの範囲を定める．

3.3 タグの特徴量

小説投稿サイトにおいて小説とタグは 2 部グラフの関係にある．タグは，その設定されたオンライン小説の世界観，テーマ，登場人物等の要素を示すものである．そのため，当然それらのタグが設定されている小説はタグが示す要素を内包しているといえる．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，タグは設定された小説の特徴量を合成することで，そのタグの特徴量を求めることができると仮定する．タグの特徴量 $\overrightarrow{T_j}$ は，そのタグが設定されている i 個の小説のベクトル $\overrightarrow{D_i}$ を合成したものである．$\overrightarrow{T_j}$ は以下の式で与えられる．

$$\overrightarrow{T} = \sum_{i \in D_i \to T_j} \overrightarrow{D_i} \tag{4}$$

推薦を行う前に，3.2 節で定めた推薦する候補となるタグの特徴量を全て求めておく．タグの大きさを統一するために，単

位ベクトルに変換を行う．

$$\overrightarrow{D} \leftarrow \frac{\overrightarrow{D}}{|\overrightarrow{D}|} \qquad \overrightarrow{T} \leftarrow \frac{\overrightarrow{T}}{|\overrightarrow{T}|} \tag{5}$$

3.4 タグ推薦

ユーザが入力する読了した小説の本文の特徴量を求め，3.3 節で求めておいたタグの特徴量と類似度を比較し，タグの推薦を行う．類似度の比較には Cos 類似度を用いる．入力された小説本文のベクトルを$\overrightarrow{D_i}$，類似度比較を行うタグのベクトルを$\overrightarrow{T_j}$とすると Cos 類似度は以下の式で求められる．

$$\cos(\overrightarrow{D_i}, \overrightarrow{T_j}) = \frac{\overrightarrow{D_i} \cdot \overrightarrow{T_j}}{|\overrightarrow{D_i}||\overrightarrow{T_j}|} \tag{6}$$

Cos 類似度は -1 から 1 の範囲で表される．より 1 に近い値を取るものが，入力した小説本文により類似したタグである．よって，小説本文の持つ性質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った場合，タグ同士で Cos 類似度比較を行うと，高い類似度の値を示すタグが推薦される可能性が考えられる．タグ間で Cos 類似度が高いということは，小説の同じ要素を示すタグであるということになる．このままでは，検索に効果的なタグの要件を満たすことができない．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手法について考える必要がある．

3.4.1 手　法　A

小説と高い Cos 類似度を持ちつつ，推薦されるタグ同士では類似性が低いタグの推薦をする．端的に示すと，タグの特徴量を座標上に位置した時，既に推薦されたタグと小説を挟んで反対側に存在し小説本文に類似したタグを推薦することで上記の推薦を実現する．以下にその詳細を記す．

入力した小説本文 $\overrightarrow{D}$ は推薦の基準として，$\overrightarrow{T_0}$ とする．推薦するタグは 1 番目に選ばれるタグから N 番目のタグを，$\overrightarrow{T_1}$, $\overrightarrow{T_2} \ldots \overrightarrow{T_N}$ とする．$\overrightarrow{T_1}$ は，小説本文と最も類似度が高いタグを選出する．

$$\begin{aligned} \overrightarrow{T_0} &= \overrightarrow{D} \\ \overrightarrow{T_1} &= \arg\min_i |\overrightarrow{DT_i}| \end{aligned} \tag{7}$$

新たに推薦するタグが既に推薦されたタグと類似性の低い特徴量を持つようにするためには，今まで推薦されたタグ全てを考慮する必要があるが，計算量を抑えるために，直近に推薦された 2 つのタグを考慮する対象とする．$\overrightarrow{T_i}$を次の推薦するタグとすると，直近 2 つに推薦されたタグは$\overrightarrow{T_{i-1}}$と$\overrightarrow{T_{i-2}}$と示せる．$\overrightarrow{T_{i-1}}$ と $\overrightarrow{T_{i-2}}$ をつなぐ線を，小説本文 D から引いた二等分線の交点を $\hat{T}$ とする．この $\hat{T}$ の逆位置にあたる $-\hat{T}$ に近いものから推薦タグを選ぶ．$\overrightarrow{T_i}$ を $-\hat{T}$ と異なる象限へと写像するために，$-\overrightarrow{DT}$ と $\overrightarrow{DT_i}$ がなす角 θ_i (θ_i は $0 \leqq \theta_i \leqq \pi$ の範囲) を二等分した位置に T_i を写像し T_i' とする．T_i' から-$\overrightarrow{DT}$へと伸ばした垂線の交点を T_i'' とする．このとき，以下の式で T_i'' を求められる．

$$\begin{aligned} \overrightarrow{T_i''} &\leftarrow \arg\min_i |\overrightarrow{DT_i''}| \\ |\overrightarrow{DT_i''}| &= \frac{|\overrightarrow{DT_i}|}{\cos(\frac{\theta_i}{2})} \end{aligned} \tag{8}$$

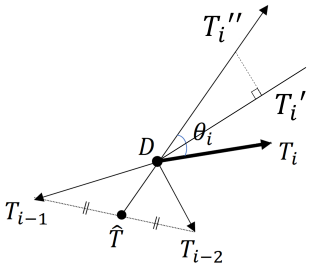


図 1: 推薦タグ T_i の図解

求められた T_i'' の元となったタグが推薦されるタグへと選出される．図 1 が上記の図解である．この手順を N 番目まで繰り返し，ユーザに推薦するタグを決定する．

3.4.2 手　法　B

手法 A では，タグの特徴量と入手された小説本文を比較することで，検索に効果的なタグの推薦を行った．この手法では，検索に効果的なタグの要件である「入力された小説の内容と合致したタグ」と「推薦されたタグ同士でタグの表す情報に重複がないタグ」の 2 つに重点を置いてタグの推薦を行っている手法である．しかし，もう 1 つの要件である「複数のタグで絞り込みができるタグ」については，3.2 で行った，タグの情報量によるタグの推薦範囲を定めただけに留まっている．そのため，手法 A では推薦したタグの中で情報量に偏りがみられる可能性がある．このとき，複数あっても検索するときにあまり絞り込みができないケースなどが推測される．

手法 A では，推薦するタグを選出するときに，最も重要視したものは，入力された小説本文の特徴量とタグの Cos 類似度の高さであった．そこで，手法 B ではタグの情報量という軸からの視点を，手法 A へと取り入れる．タグの情報量を I とし，以下の式からタグ $\overrightarrow{T_i}$ を求める．

$$|\overrightarrow{T_i}| \sim \alpha \cos(\overrightarrow{D}, \overrightarrow{T_i}) + (1-\alpha)\frac{1}{I} \tag{9}$$

このとき α は $0 \leqq \alpha \leqq 1$ とする．手法 A の小説本文の特徴量とタグの特徴量を比較するときに，上記の式を当てはめて，タグの推薦を行う．

4 実験・評価

本章では，第 3 章で提案した手法について実験及び評価を行う．

4.1 データセット

本研究の実験・評価には，「小説家になろう」に投稿された小説およびそれに付随するタグを用いる．

4.1.1 小説本文

本研究では，小説の特徴量（ベクトル）を生成するために小説の本文を用いる．本研究では，「小説家になろう」に投稿された小説のうち大ジャンル「恋愛」に区分されているジャンル「異世界」に分類される小説を対象とした．小説本文の収集期間は．2019 年 7 月 18 日から 2020 年 1 月 22 日である．上記期間で，サイトの負荷を考慮し断続的に収集を行った．

4.1.2 タグ

小説に設定されているタグ（小説家になろうサイト上では，キーワードが相当する．）の収集を行った．タグの収集には，なろう API[3]を利用した．小説家になろう全体の傾向を明確にするため，サイト内の小説を網羅的に収集を行い，661,495 件の小説に設定されたタグ 3,833,470 個を収集した．タグの収集日は，2019 年 6 月 27 日である．

1 つあたりの小説にどのくらいタグが設定されているかを調査した．図 2 はその結果である．図 2 から読み取れるように，タグが 4 つ設定されている小説が最も多くなっていることがわかった．また，設定されているタグの平均値を求めると 5.80 であった．以上のことから，タグはおよそ 5 個あればその小説の性質を表すことができると考えられる．そこで，本研究で効果的なタグとして推薦するタグの個数を 5 個とした．

小説家になろうでは，様々な種類のタグが存在する．本研究では，これらのタグをその性質によって，自由タグとシステムタグという 2 つの枠組みを作成，その 2 つに大別した．自由タグとは，著作者（ユーザ）の手によって決められる，自由度の高いタグのことである．小説家になろうにおいては，手動入力キーワードを自由タグに分類した．システムタグとは，サイト側が用意したタグのことである．手動入力キーワード以外のキーワード，登録必須キーワードをシステムタグとした．システムタグは，サイト側が用意したという特徴から，普遍的なもので多くの人にとって既知であったり，ある要素を含むものをゾーニングさせるためのものであったりするものが多い．図 3 は，小説家になろう上のタグの出現頻度を，すべてのタグ，自由タグのみそれぞれの様子をあらわしたものである．比較してみると，システムタグが上位に存在していることがわかる．これは，システムタグは選択式のため，情報量が小さいもの多いと推測でき，検索の絞り込みには向いていない．以上から，本実験で検索に効果的なタグとして推薦するものは，自由タグから選ぶものとした．

図 3 を見ると，自由タグのみの出現頻度も全体の出現頻度と同様に分布している様子がわかる．そのため，システムタグを取り除き自由タグのみを推薦する場合でも，情報量の極めて大きいものと小さいものは，推薦候補から除外する必要がある．本実験では，タグの出現頻度が 10,000 より大きいもの，出現頻度が 5 回未満であるものを，それぞれ情報量の極めて大きいものと小さいものとした．

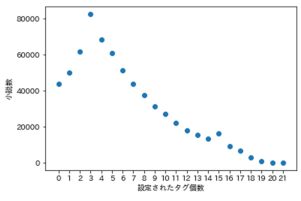


図 2: 1 つの小説あたりに設定されたタグ数

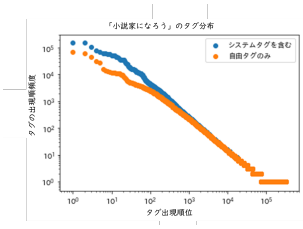


図 3: 「小説家になろう」内のタグの出現頻度分布

4.2 ベクトル生成

4.1 で準備したデータセットを基に小説本文及びタグのベクトルを生成する．

4.2.1 本文のベクトル生成

4.1.1 節で収集した小説本文の特徴量を求めるために，本文のベクトル生成を行った．タグの推薦に用いるためには，小説とタグが 2 部グラフの関係である必要がある．したがって，収集した本文のうち，タグと 2 部グラフの関係にない，すなわち，タグが 1 つも設定されていない小説を取り除いた．また，小説本文が数行程度など，非常に短い作品もみられた．これらの小説は，ほかの小説と比べて特徴量を抽出するために，分量が十分であるとはいえない．そのため，小説の本文が 1,000 文字以上の小説を対象とした．以上の条件を満たす小説本文 26,904 件の特徴量を求めた．

まず，小説本文の形態素解析を行った．形態素解析には，形態素解析エンジンである MeCab [9] を用いる．形態素解析では，本文から取り出す単語を，名詞，形容詞，動詞，形容動詞の 4 つの品詞を取り出した場合と，上記に加えて連体詞，副詞，接続詞，感動詞の合計 8 つの品詞を取り出した場合の 2 通りで行った．このパターンをそれぞれ前者をパターン 1．後者をパターン 2 と称する．この 2 通りの品詞取り出しパターンを比較することで，小説の特徴量をより適切に表すのに必要な品詞について検証する．品詞を取り出すときには，原形に統一し取り出しを行った．オンライン小説は，幅広い年代の人々が執筆し投稿されているため，型にはまらない様々な表現が用いられている．小説の内容から適切に特徴量を抽出するために，形態素解析には新語に対応可能な辞書である NEologd を用いた．ストップワードとしては，SlothLib [10] が提供しているストップワードリストを利用した．

形態素解析した本文は，文書から分散表現を求める手法である Doc2Vec [11] を用いて 300 次元のベクトルを生成した．小

3：https://dev.syosetu.com/man/man/

説本文の文量の差が，小説本文の特徴量へと影響を与えないようにするために，単位ベクトルへと変換した．

4.2.2 タグのベクトル生成

4.2.1 で生成した小説の本文ベクトルをもとにタグのベクトル生成を行った．タグのベクトルもその出現頻度に特徴量が影響を受けないように単位ベクトルへと変換を行った．タグのベクトルは，4.1.2 で作成したタグの識別子である T コードに紐づけることで管理を簡潔に行えるようにした．

4.3 タグの推薦実験

4.2.2 節で生成したタグのベクトルと，小説本文のベクトルを類似度比較し，検索に効果的なタグの推薦を行った．大ジャンル「恋愛」・ジャンル「異世界」に属する小説から無作為に選んだ 10 件の小説を対象とした．小説本文は，4.2.1 節と同様な手順で特徴量を算出した．それぞれ手法 A，手法 B ごとに推薦実験をパターン 1，パターン 2 それぞれで行う．

4.3.1 手　法　A

手法 A，B の手順に従い，各小説にタグの推薦を行った．手法 B では，タグの特徴量の類似度とタグの情報量のどちらに重点を置くかパラメータ α で決定する．α は，$0.1, 0.5, 0.9$ の 3 つの値に分けて実験を行った．表 1，表 2 は，それぞれ小説本文のベクトルと推薦したタグのベクトルの Cos 類似度を求め，全体で平均したものである．表 1 のそれぞれパターン 1 とパターン 2 で，Cos 類似度に大きな違いは見られない．表 2 の結果は，手法 A と比べて，全体的に Cos 類似度の値が小さくなっていることがわかる．パターン 1 とパターン 2 によった目立った差異は見受けられない．Cos 類似度は 0 から 1 の値であり，値が 1 に近づくほどそれらのベクトルは類似しているといえる．すなわち，Cos 類似度が高いほど，小説とタグは同様な性質を持っていると考えられる．そこで，推薦したタグと各小説に著作者の手によって設定されたタグ（著作者タグ）と比較を行うことにした．著作者タグは，小説を作成した当人によって設定されたタグであるため，小説の持つ性質と同じ性質のタグであると考えられる．

手法 A の結果を表 3 に，手法 B の結果を表 4 に示す．表 3 から両パターンで，著作者タグと合致したタグが推薦できていることがわかる．このことから，手法 A で推薦したタグの中には小説と同様な性質を持っているタグが含まれていることが分かった．手法 A のパターン 1 とパターン 2 を比較すると，パターン 2 のほうが著作者タグと合致したタグを推薦できていることが読み取れる．手法 B では，パターン 1 とパターン 2 での大きな差異は見られなかった．手法 A の結果である表 3 と比べると，著作者タグと合致した推薦タグが少なくなっている．しかし，検証の際に手法 A では，著作者タグと合致する推薦タグがなかった小説に対して，著作者タグと合致したタグを推薦できていたケースもあった．著作者タグと合致した推薦タグの小説本文との Cos 類似度は，平均値より高いものもみられたが，Cos 類似度が高くても著作者タグと合致しないタグも見られた．したがって，Cos 類似度が高いタグが著作者タグと合致するとは一概にいえない．

表 1: 小説ベクトルと手法 A の推薦タグベクトルの Cos 類似度平均値

推薦順	パターン 1	パターン 2
1	0.726	0.707
2	0.593	0.578
3	0.579	0.566
4	0.568	0.559
5	0.531	0.524

表 2: 小説ベクトルと手法 B 推薦タグベクトルの Cos 類似度平均値

	パターン 1			パターン 2		
α	0.1	0.5	0.9	0.1	0.5	0.9
推薦順 1	0.514	0.511	0.647	0.500	0.503	0.633
推薦順 2	0.516	0.517	0.504	0.502	0.500	0.500
推薦順 3	0.521	0.525	0.507	0.508	0.517	0.502
推薦順 4	0.534	0.521	0.514	0.513	0.506	0.503
推薦順 5	0.518	0.512	0.512	0.514	0.507	0.497

表 3: 手法 A の推薦タグと著作者タグの比較

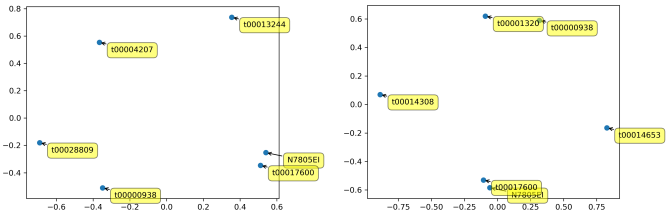
	パターン 1	パターン 2
総合致タグ数	6	11
最大合致数	3	4
合致小説数	4	6

表 4: 手法 B の推薦タグと著作者タグの比較

	パターン 1			パターン 2		
α	0.1	0.5	0.9	0.1	0.5	0.9
総合致タグ数	4	4	5	4	4	5
最大合致数	3	2	3	2	2	3
合致小説数	4	3	3	3	3	3

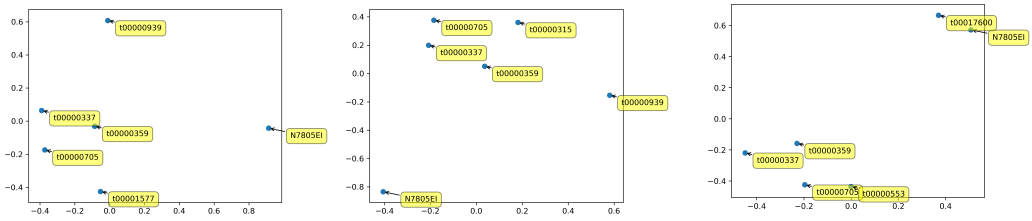
推薦したタグ同士で異なった性質を持っているかを検証するために，小説と推薦したタグのユークリッド距離を算出し，多次元尺度構成法により可視化を行った．図 4 は，N コードが N7805EI の小説に対して手法 A で推薦されたタグの距離を可視化したものを一例として挙げる．大きな傾向を把握することは難しいが，タグ同士である程度離れて位置していることが読み取れる．同様に手法 B でも，形態素解析の品詞パターンと α の値ごとに図を作成した．図 5 および図 6 の (a) (c)α を変えた結果を示している．これらは共通して，小説とタグの位置が離れたところに位置しているように読み取れる．タグ同士は，ある程度の距離を持って分布していると読み取れる．このことから，手法 A，B ともにタグ同士でタグの表す情報に重複が少ないタグを推薦できたと考えられる．

推薦したタグが，複数のタグで絞り込みができることを満たしているかを検証するために，推薦タグが設定されている小説を求めそれらを 1 つの集合とし，各集合同士の共通集合を求めた．実際の小説投稿サイト内での検索を想定し，今回収集した 661,495 件の小説全てを対象とした．比較対象には，著者タグを用いた．ここでは，各手法のパターン 2，手法 B は $\alpha = 0.1$ の結果のみを示す．以下の表 5〜7 がその結果である．表 5 にあるベースラインの結果を見ると，1 万件以上の大きさを持つ共通集合や該当が 1 件のみの共通集合など小説によって結果が大きく異なっている．このことからベースラインでの絞り込みは著者が付与したタグに依存すると考えられる．表 6 の手法 A で推薦したタグの結果を見ると，共通集合が空集合となっている状態が多くみられ、空集合でないものも件数が少ない集合が



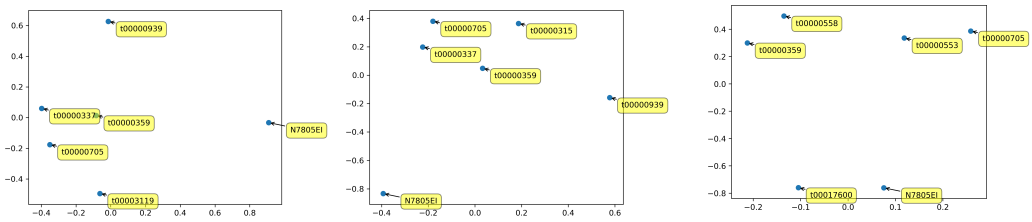
(a) パターン 1　　**(b)** パターン 2

図 4: 小説 N7805EI と手法 A で推薦されたタグの距離



(a) $\alpha = 0.1$　　**(b)** $\alpha = 0.5$　　**(c)** $\alpha = 0.9$

図 5: 小説 N7805EI と手法 B パターン 1 で推薦されたタグの距離



(a) $\alpha = 0.1$　　**(b)** $\alpha = 0.5$　　**(c)** $\alpha = 0.9$

図 6: 小説 N7805EI と手法 B パターン 2 で推薦されたタグの距離

多くみられる．表 7 の手法 B で推薦したタグの結果を見ると，手法 A と比較して全体的に空集合が減少し，1 つの共通集合に含まれる小説数も多くなっている．情報量を考慮しない手法 A の時，推薦タグ単体では数件から千件越えの小説に絞り込めるタグが推薦されるが，何件になるかはタグ依存で安定した手法とは言えない．情報量を考慮した手法 B の場合，タグ単体では数千件が該当するが，2 つのタグを選択・指定すると数十件に絞り込めるケースが多く現れ，絞り込み手法として安定した絞り込み件数を得ることができる．このことから，情報量を考慮したタグ推薦は，ウェブ小説の絞り込み検索を効率的に行うことができる．

4.4 考　　察

本節では，4 章で行った実験・評価について考察を行う．Mecab の形態素解析で用いた品詞によって 2 つのパターンで分けてタグの推薦を行ったが，推薦したタグの Cos 類似度には大きな違いは見られなかった．一方で，パターン 2，名詞，形容詞，動詞，形容動詞，連体詞，副詞，接続詞，感動詞の 8 つの品詞を特徴量に用いたときのほうが，著作者タグと合致したタグを推薦できていたため，より小説の内容に沿ったタグを推薦でいていたと推測できる．4.3 節の実験では，入力の小説を無作為としてしまったため，小説の性質による推薦の違いなどの検証はできていない．文量や同じシステムタグが設定されているなどの制約をかけて分類した小説を入力とすることで，どのような小説に提案手法が有効であるかの検証が可能となる．著作者タグの個数もそれぞれで異なっていたため，統一することでより正しく評価ができると考えられる．

手法 A のほうが著作者タグと同一のタグをより多く推薦できるという点で優位性があったが，絞り込み検索においては手法 B の方が効率的に行える．これらの手法の利点を組み合わせた手法を考案することにより効果的な推薦が行うことができると考えられる．

5 おわりに

本研究では，小説投稿サイトに投稿されている小説を読了したユーザに対して，その読了した小説と類似した要素を持つ小説を検索するために効果的なタグを推薦する手法について，提

表 5: ベースライン（著者タグ）

n コード	タグ 1	タグ 2	タグ 3	タグ 4	タグ 5	タグ 6	1 ∩ 2	1 ∩ 3	1 ∩ 4	1 ∩ 5	1 ∩ 6	2 ∩ 3	2 ∩ 4	2 ∩ 5	2 ∩ 6	3 ∩ 4	3 ∩ 5	3 ∩ 6	4 ∩ 5	4 ∩ 6	5 ∩ 6
N7805EI	772	3833	2679	69379	13	4884	13	13	87	1	9	116	1047	1	27	736	1	31	1	523	-
N4698EF	69379	4146	582	1883	59971	-	438	112	1243	10532	-	5	15	892	-	3	131	-	462	-	-
N9215DX	7	1	41	59971	582	63	1	1	4	1	1	1	1	1	1	7	2	1	131	18	2
N7491DE	69379	59971	393	120	1	110	10532	25	18	1	30	122	38	1	17	2	1	1	1	1	1
N8796FC	45126	1577	-	-	-	-	160	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 6: 手法 A パターン 2

n コード	推薦タグ 1	推薦タグ 2	推薦タグ 3	推薦タグ 4	推薦タグ 5	1 ∩ 2	1 ∩ 3	1 ∩ 4	1 ∩ 5	2 ∩ 3	2 ∩ 4	2 ∩ 5	3 ∩ 4	3 ∩ 5	4 ∩ 5
N7805EI	13	772	135	28	11	1	0	0	0	0	0	0	0	0	0
N4698EF	87	1883	120	10	19	2	0	0	0	6	0	0	0	0	0
N9215DX	7	18	697	8	17	0	0	0	0	0	0	0	0	0	0
N7491DE	120	393	157	11	17	2	0	0	0	0	0	0	0	0	1
N8796FC	30	52	455	84	66	0	1	1	0	0	0	0	0	0	0

表 7: 手法 B パターン 2

n コード	推薦タグ 1	推薦タグ 2	推薦タグ 3	推薦タグ 4	推薦タグ 5	1 ∩ 2	1 ∩ 3	1 ∩ 4	1 ∩ 5	2 ∩ 3	2 ∩ 4	2 ∩ 5	3 ∩ 4	3 ∩ 5	4 ∩ 5
N7805EI	3833	2679	1976	1761	1475	116	62	26	84	89	47	115	9	131	8
N4698EF	3927	3194	2762	1883	1486	140	11	47	0	17	134	23	43	109	42
N9215DX	3194	1883	1154	1032	1004	134	9	14	66	2	6	18	0	1	18
N7491DE	8611	4814	4684	3929	3776	105	123	189	229	158	93	147	55	64	84
N8796FC	9572	7986	7475	5299	4859	91	824	135	86	21	65	245	67	40	109

案を行った．

具体的には，小説本文と小説に付与されているタグを行い，個々の小説の特徴量（ベクトル）は，小説本文に Doc2Vec を適用し算出した．一方，タグの特徴量は，そのタグが設定されている 2 部グラフの関係にある小説のベクトルを合成して算出した．本研究における検索に効果的なタグとは，入力された小説の内容と合致したタグ，推薦されたタグ同士でタグの表す情報に重複がないタグ，複数のタグで絞り込みができるタグの 3 つの要件を満たすタグとした．提案手法を小説家になろうに投稿されているオンライン小説に適用・評価し，提案手法の有効性を検証した．限定された範囲の小説に関してではあるが，効果的な検索タグを推薦できることができた．

今後の課題としては，提案手法で推薦したタグについてのユーザ実験による評価を行い，手法の改善を行いたい．また，今回実験でもちいた小説と異なるジャンルにも手法を適応させ，その結果を考察することで小説投稿サイト全体でのユーザの検索行動を支援していくことが重要である．

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. 第 9 回データ工学と情報マネジメントに関するフォーラム, DEIM2017, B5-2, 高山グリーンホテル 岐阜県高山市, March 3 - 9, 2017. IEICE データ工学研究専門委員会.

[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, 2013.

[4] 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 8 回データ工学と情報マネジメントに関するフォーラム, DEIM2016, B6-5, ヒルトン福岡シーホーク 福岡県博多市, February 29 - March 2, 2016. IEICE データ工学研究専門委員会.

[5] 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.

[6] 山崎　睦月, 佐藤 哲司. オンライン小説の検索に有効なタグの推薦手法の提案. 第 11 回データ工学と情報マネジメントに関するフォーラム, DEIM2020, C2-4,2020.

[7] Mutsuki Yamazaki, Kazufumi Inafuku, and Tetsuji Satoh. Tag recommendation method for enhancing web novel retrieval. In Tokuro Matsuo, Kunihiko Takamatsu, Yuichi Ono, and Sachio Hirokawa, editors, *9th International Congress on Advanced Applied Informatics, IIAI-AAI 2020, Kitakyushu, Japan, September 1-15, 2020*, pp. 43–48. IEEE, 2020.

[8] Genkou Ou, Kei Wakabayashi, and Tetsuji Satoh. Searching behavior analysis of online shopping based on information content of query words. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 43–48, 2019.

[9] 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.

[10] 大島 裕明, 中村 聡史, 田中 克己. Slothlib web サーチ研究のためのプログラミングライブラリ. 日本データベース学会 letters, Vol. 6, No. 1, pp. 113–116, 06 2007.

[11] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.