Web小説のCGMにおける読者の読書行動についての一考察

安田幸生（慶應義塾大学大学院）
植原啓介（慶應義塾大学）
佐藤雅明（東海大学）

概要:近年，様々なジャンルにおいてCGM（Consumer Generated Media）が台頭している．とりわけ，小説分野における投稿サイトでは，作者・読者の双方が年々増加傾向にある．これは，読者が読むべき対象を多数から選ぶことが困難であると共に,作者にとっても作品を読者に届ける競争が激化していることを意味する．　本研究では，小説投稿サイトを対象に，読者の行動における第一段階である読むかどうかを決めるフェーズにおいて，読者の作品の選択行動について作品に紐づく情報と閲覧数などのデータから分析し，考察をおこなった．
キーワード:CGM，選択行動，自然言語処理，機械学習

An Analysis of Reading Behavior in CGM of Web Novels

Koki Yasuda /Keisuke Uehara（Keio University）
Masaaki Satoh（Tokai University）

Abstract: In recent years, CGM（Consumer Generated Media）has emerged in various genres. In the field of fiction, in particular, the number of authors and readers in posting sites has been increasing year by year. This means that it is difficult for readers to choose works, and for authors to deliver their works to readers. In this study, we analyzed the selection behavior of readers on a novel posting site based on data such as information related to the work and the number of views.
Keywords: CGM, Selection Behavior, Natural Language Processing, Machine Learning

1．はじめに

近年，様々なジャンルでCGM（Consumer Generated Media）が台頭している．とりわけ，小説分野における投稿サイトは，作者の数・読者の数の双方において増加傾向にある．これは，読者が読むべき作品の選択が困難になると共に，　作者にとっても，自身の作品を読者に届けるための競争が激化していることを意味する．

こうした流れに付随して，小説等の創作を対象とした研究は増えているが，小説の評価について分析をおこなう論文の多くは，読後の評価を対象とした分析やその内部的な評価を一体として処理している．しかし，実際に読んで評価するまでのプロセスは，「まず読むかどうかを決める」「読んでみてどうかを判断する」という2つの段階に分かれていると考えられる．

本研究では，読者の行動における第一段階である，読むかどうかを決めるフェーズにおいて，どのような要因が決定的な要因となっているのかを，データを用いた機械学習による手法とアンケート調査結果を比較検討することで分析し，考察を行った．

2．関連研究

本研究では，タイトル等の作品を読む前に取得可能な情報を用いて，読み始めるかどうかを決める読者の行動の分析を目的としており，この分析の手段として機械学習による手法を用いている．そこで，本章では特に創作分野に機械学習手法を用いた関連研究について触れる．

2.1 小説の自動生成

Osoneら[1]は，AIとの共創をテーマとし，GPT-2を物語文章で学習することにより，日本語の物語共創システムの作成手法を提案している．また，Alabdulkarimらによる物語文章の自動生成に関するサーベイ論文[2]では，物語の操作性やコモンセンスの活用方法等にフォーカスしており，創造的な文章と物語の構造的なそれらしさが現状ではまだトレードオフの関係であることや，確立された評価指標が物語生成に存在しないことを明らかにしている．

2.3 小説の内容評価

ジョン・マッケンジーらによる書籍[3]では，計量文献学の観点から様々な特徴量を文章中の情報等を基に,NYタイムズの小説の作品がベストセラーとなるか否かについて分類を行っており，例えば「will」という言葉の割合が多い作品では主人公の意思があらわになっている点で評価がなされている可能性がある，等の示唆的な考察が行われている．

本研究では内容に踏み込む前段階として，作品に関する表層の情報をもとに閲覧数の予測を試みているため，内容を読んだ後の評価等を考慮する研究とは目的が異なっている．

小坂ら[4]による研究では，読み手の立場から作品に対するアプローチが難しい点から，ユーザーによる推薦基準の設定が可能なレコメンドシステムを提案している．しかしながら，作品の質を表すスコアの計算に総合ポイントとレビュー数を扱っている点から，その

作品を読むかどうかを決めるための行動の分析ではない.

本研究では, 読者が小説に対して評価を付けるまでの行動を「読むかどうかを決める」と「読んでみてどうかを判断する」という二段階に分かれていると仮定し, 前者の「読むかどうかを決める」部分において重要な要因を機械学習的なアプローチによって導き出すことを目的としており, 本研究の趣旨とは異なる.

3. データセット

本章では, 本研究において利用したデータセットの収集方法を3.1章にて行い, 機械学習手法による分析のために行ったデータセットの加工方法について3.2章以降で述べる. なお, 利用した機械学習のモデルは2種類であり, それぞれのモデルで想定される入力の形式が異なる. 表データとして加工した方法を3.3章にて述べ, シーケンスデータとして加工した方法を3.4章にて述べる.

3.1 データセットの収集

本研究では, 国内最大の小説のCGMサイトである「小説家になろう」[5]に, 2017年から2021年の間に投稿された作品の中で, 以下の条件を満たす作品のユニーク閲覧数を取得した.

・短編小説である
・文字数が1000字を超えている
・2017年から2021年の間に投稿されている

まず, ユニーク閲覧数を取得した理由について述べる. 「小説家になろう」において公式が作品評価として用いている「総合ポイント」は, ある作品に対して読者が「ブックマーク」や「感想」といったリアクションをとることで算出される値である. 「総合ポイント」は, 読んでみてどうだったかを基にしており, 本研究での目的である読むかどうかを決める段階の分析において活用することは不適切である.

よって, 本研究の分析では「総合ポイント」ではなく, 「小説家になろう」の運営会社がアクセス解析システムとして公開している外部サイトであるKASASAGI[7]より, 各作品の累計の「ユニーク閲覧数」を取得し, 分析に用いた.

次に、本研究で対象を短編小説とした理由について述べる. これは, 投稿される短編小説は投稿時点で完結していること, 作品ごとに読者の目に触れる回数が同じであること, この2つの点で長編の作品を対象にするよりも個々の作品に対してフェアな比較が可能であると考えたからである.

抽出の条件として文字数を1000字以上としたのは, 事前調査としていくつかの条件を無作為に変更して抽出された作品の内容に差がないかを調査した結果からである. 事前調査の結果, 1000字以下の作品にはポエムや日記等の, 物語要素を含まない作品の割合が高いことがわかった. そのため, 分析において外れ値となることを考慮し文字数に閾値を設定した.

2017年から2021年の間に投稿されている作品という条件の是非に関して述べる. 2021年の12月に投稿された作品のうち「総合ポイント」が1,000を超える短編小説303件について, 2022年5月までの閲覧数に対する2022年6月の閲覧数の割合を算出したところ, 平均値で2.9%, 中央値で1.7%であり, これらの作品の2022年7月以降の閲覧数は, 2022月6月に比べて単調減少の傾向にあった. この結果から, 投稿から半年以上程度経過している短編作品に関しては, 投稿された期間による閲覧数の伸びを考慮せずとも問題はないと考える.

上記の条件に該当する短編小説は102,336件であったが, 削除済みの作品の閲覧数の取得がシステムの仕様上できなかったため, 閲覧数を取得できた101,747件を分析の対象とした.

3.2 目的変数の準備

本研究における機械学習手法の目的変数には, KASASAGIから取得したユニーク閲覧数のカテゴリ変数を用いた. この変換は5クラスになるように行っており, それぞれのユニーク閲覧数とユニーク閲覧数ラベルとの対応関係は0:0~200, 1:200~800, 2:800~3200, 3: 3200~12800, 4: 12800~とした. この対応関係は, ラベルごとに同数のサンプル数となるようにカテゴリ変数化を行う場合, 図1の分布から見て取れるように, 極端に小さい閲覧数の値でデータを分割することとなる. この問題を避けつつ, 元の分布の様子にある程度一致するように対応関係を設定した.

この処理の結果, 各ユニーク閲覧数ラベルごとのデータ数の分布は図1のようになった.

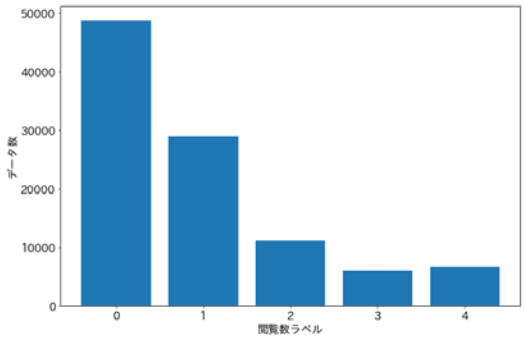

図1 ユニーク閲覧数ラベルごとのデータ数

3.3 表データの事前準備

ユニーク閲覧数ラベルの予測タスクに対して利用したモデルは2種類であり, それぞれのモデルで想定される入力は異なる. 本章では表データのデータの準備の手法について述べる.

データとしては, ジャンル, 投稿年, 作品本文の文字数, タイトルの長さ, あらすじの長さ及び, キーワード・タイトル・あらすじに含まれる名詞など(以後, タイトルキーワード・あらすじキーワードと表記する)を用いた. これらのデータの形式を表1に示す.

表1 表データとして利用した特徴量

特徴量	カテゴリ変数処理	データ例 (範囲)
ジャンル	無	VR ゲーム
投稿年	無	2019, 2021
文字数	有	0~9
タイトル長	有	0~9
あらすじ長	有	0~9
キーワード	有	0~299
タイトルキーワード	有	0~299
あらすじキーワード	有	0~299

カテゴリ変数化処理において，作品本文の文字数，タイトルの長さ，あらすじはそれぞれのデータを10クラスのカテゴリ変数に変換した．このカテゴリ変数への変換はそれぞれデータ量がおおよそ均等になるように行っており，分析を行う観点で多すぎず少なすぎない値にするためにカテゴリ変数のクラス数を10に設定した．

続いて，キーワード・タイトルキーワード・あらすじキーワードの取得処理は以下の手順により行った．

1. 全作品に付随しているキーワードを抽出する
2. 1で得られた単語群を頻度順に並べ，頻度上位300語を抽出する
3. それぞれの作品情報に付随しているキーワード群のうち，頻度上位300個の単語に含まれているものを分析対象のキーワードとする
4. 2で取得した頻度上位300語を追加の単語として登録したMeCab[10]を用いて形態素に分割する
5. 4の操作により獲得した単語群を頻度順に並べ，名詞・形容詞・動詞とキーワードの中の上位300個を抽出する
6. それぞれの作品情報に付随しているタイトル・あらすじに対して，5で得られた300個の単語に含まれているものをタイトルキーワード・あらすじキーワードとして抽出した

ここで，2及び5の操作において頻度上位300語を対象としたのは，事前の実験によって頻度上位300語より多くの語を使っても，ユニーク閲覧数の予測モデルの学習の精度向上に寄与しないことがわかっていたためである．

上記の操作により，キーワードは元の形式が単語のリストのため，頻度の少ない語をフィルタリングしたもの，タイトル及びあらすじは元の形式が自然言語であるため，形態素に分割することで一度文字列のリストに変換し，その後頻度の少ない語をフィルタリングすることで最終的な単語のリスト(タイトルキーワード及びあらすじキーワード)を得た．また，実験の際にはそれぞれのキーワードを300個使う方法だけではなく，徐々に利用するキーワードの種類を増やしていくことでどのように精度向上に対して寄与するのかについても調査を行った．

3.4 シーケンスデータの事前準備

取得したデータを自然言語的に処理するためにシーケンスデータを用意する方法について述べる．BERTのfine-tuningを行う上で標準的な入力の方法に則り，ジャンル・文字数・キーワード・タイトル・あらすじを，数字や文字列等の形式が何であるかに関わらず文字列に変換し，それぞれをBERTのモデルにおいて文字列のブロック間の明示的な分割の意味合いを持つ疑似的な単語であるseparation tokenを介してつなげたものを入力とした．なお，表データの場合と異なり，カテゴリ変数化の処理は目的変数である閲覧数ラベルを除き，どの特徴量にも実施していない．

実際の学習及び推論の際には，キーワード・タイトル・あらすじの，それぞれの分類性能への寄与度を調査するため，ジャンル・文字数・投稿年度を必ず入力に含める情報とした上で，キーワード・タイトル・あらすじのそれぞれの情報を入力に含める場合と含めない場合でいくつかのパターンの学習方法を試行した．

4. 予備分析

本章では，前章のデータセットに対して機械学習手法を適用する前に，作成したデータセットのみから言える事項について4.1章にて述べ，「小説家になろう」利用者を対象とした読書行動に関するアンケート結果と分析を4.2章にて述べる．

4.1 データセットの分析

小説家になろうのAPI[1]によって取得が可能な情報のうち，作品の詳細ページから確認できるデータ等は用いておらず，プラットフォーム上で作品一覧が並んでいる際に確認が可能な情報と，読者や著者の数に関連する可能性のある年度に絞ったものを分析の対象とした．

具体的には，表2に示す情報が分析に用いたものである．これらの情報の他，作品一覧からはあるユーザーの閲覧時点で既に獲得している総合ポイントが確認可能だが，ユーザーごとに閲覧のタイミングが異なるため，トラッキング及びデータ化が難しい．また，後述する4.2章のアンケート結果から，作品選びにおいて総合ポイントは重要視されない傾向がみてとれたため，分析対象外とした．

表2 本研究で分析対象とした特徴量の種別

名称	説明
タイトル	作品のタイトル
作者 ID	作者を表す固有 ID
大ジャンル	「恋愛」や「SF」などの粒度のジャンル
小ジャンル	「異世界恋愛」や「VR ゲーム」などの粒度のジャンル
あらすじ	作品のあらすじ
キーワード	著者が編集可能な作品に関するキーワード
作品文字数	作品本文の文字数
作品投稿年度	作品が投稿された年度

視認性の観点から閲覧数が20,000以上の作品を除いた作品の閲覧数の分布を図2に掲載する．投稿される短編作品全体の78.8%程度の作品はユニーク閲覧数が1,000以下となっており，多くの作品が一定

[1]https://dev.syosetu.com/man/api/

以上の閲覧数を得るために苦労をしていることが伺える.

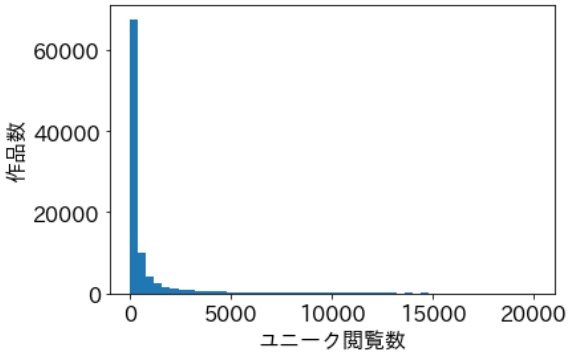

図2 短編小説のユニーク閲覧数の分布

また, 図3に2022年8月13日時点で取得したユニーク閲覧数と総合ポイントの関係を示す.

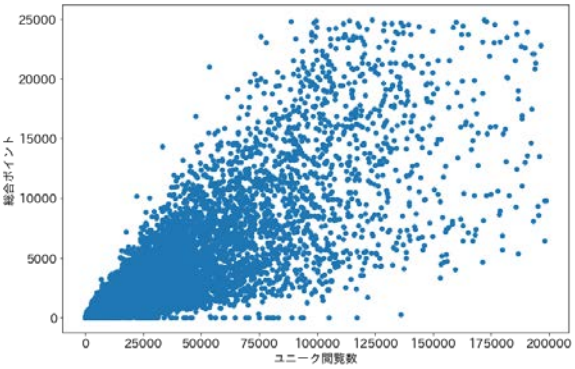

図3 ユニーク閲覧数と総合ポイントの関係

総合ポイントはレビュー数とブックマークの数によって計算がなされる値であり,「小説家になろう」ではこの指標で作品をランキング化する.

図3からこれらの2つの指標は概ね比例関係にあるが, ユニーク閲覧数が多いにもかかわらず総合ポイントがあまり伸びていないものもやや見受けられ, 読まれるまでのフェーズと読んで評価を行うフェーズの存在が確認できる.

さらに, 年度ごとの閲覧数が100,1000,10000以下の作品の比率を表3に示す. 表3によれば, 投稿時期が現在に近づくと共に閲覧数の低い作品の割合が下がっている. この点から, 読者数の増加が作者数の増加よりも多いか, 一人当たりの読書数が増えていることがわかる.

表3 年度ごとの閲覧数が一定以下の作品の比率

	2017	2018	2019	2020	2021
10 以下	0.542	0.550	0.545	0.523	0.465
100 以下	0.868	0.875	0.846	0.804	0.738
1000 以下	0.966	0.966	0.955	0.926	0.885

4.2 アンケートによる調査と分析

ユニーク閲覧数を用いた客観的分析を行う事前調査として,「小説家になろう」に読者ユーザーとして参加していると自認している300名に対し, 作品を選ぶうえで大事にしている要素を, タイトル・ジャンル・あらすじ・キーワード・総合ポイントの5項目に対して順序付けし, 1番目に重要だとした項目の理由を記載してもらうアンケートをクラウドソーシングにて行った. この結果の票数をまとめたものを表4に示す.

なお, この5つの項目は小説家になろうに登録されている作品を, 読むための公式サイト「小説を読もう」[6]の作品のリスト表示において確認できる情報である.

表4 作品選びに関するアンケート評価の票数

	あらすじ	タイトル	ジャンル	総合ポイント	キーワード
1 位	88	84	91	21	16
2 位	100	58	97	23	22
3 位	71	71	58	35	65
4 位	30	38	34	110	88
5 位	11	49	20	111	109

表4のアンケートの結果から, 読者の多くはあらすじ・タイトル・ジャンルを作品を選ぶ手掛かりとしており, 2番目に重要だと答えた回答数のばらつきから, あらすじ・ジャンルをタイトルよりも重要視していることがわかる.

一方で, 総合ポイント・キーワードは重視していないという結果が得られた. この総合ポイントを重視しているという意見が少ない点から, 本研究において予測モデルの作成するにあたり,「閲覧者がその作品を見かけた際の総合ポイント」を特徴量に採用しない場合でもある程度もっともらしい結果が得られることが期待できる.

また, 自由回答欄として, 上記の5つ以外にも大事だと思う要素を教えてくださいという問いに対して, 55人が「著者」(作者・作家等の表記ゆれを含む)であると回答した. 他には「文字数」(5人)や「主人公」(5人)などという回答もあった. 自由回答の中では「著者」という回答が突出して多く, 著者も十分に重要な特徴量であることが伺える.

しかし, このアンケートでは小説家になろう全体の読者を対象としており, 小説の種別が短編・長編のどちらかを指定していない. そこで, 小説の種別が短編の場合に作者が重視されるかを検証するために,「小説家になろうに短編を2017年~2021年の間に複数回投稿している」作者の2017年以降の初回作と最新作のユニーク閲覧数を比較した. その結果, 短編小説の複数回投稿者14,747人の2017年以降の初回作の総合ポイントの平均は475.68, 中央値は10である一方, 最新作の総合ポイントの平均は448.89, 中央値は8であった. 平均値・中央値ともに多少のポイントの増減はあるものの, 著者が重要視されると仮定した場合には, 複数回投稿者の作品のポイントが下がるとは考えにくい. そのため本研究では著者情報を分析対象から除外した.

5. ユニーク閲覧数の予測

本章では「読むかどうかを決める」ステップの分析をおこなう. そのため, まず機械学習の手法を用いてユニーク閲覧数を表すカテゴリラベルを目的変数として学習させる. 具体的には, 読む前に得られる情報を特徴量, 5クラスのユニーク閲覧数ラベルを目的変数として, 2種類の機械学習のモデルを用いて学習・予測を行った.

本研究では, 表データに対して学習・予測を行うLightGBM[8]と, シーケンスデータに対して学習・予測を行うBERT[9]を利用したLightGBMとは, 勾配ブースティングを利用する機械学習の予測モデルの一つであり, 予測精度と時間のトレードオフにおいて特に優れたモデルである. BERTは, 自然言語処理の多くのタスクにおいて高い精度を記録している深層学習をベースとした事前学習モデルである.

5.1 学習の実行

本研究で用いている閲覧数のデータセットは, 閲覧数が少ないほうに寄っている不均衡なデータセットである. そこで, 不均衡なデータのままテストデータを分割したのち, 残りのデータに対してundersamplingを行い, 学習データと検証データに分割した. この操作により, 学習データ・検証データ・テストデータの各ラベルごとのデータ数は表5の通りとなった.

このデータを用いて, 上述の表データの準備を行い, LightGBMを用いて5クラスのユニーク閲覧数ラベルに対して学習・推論を行った.

表5 データ種別ごとのサンプル数

閲覧数ラベル	0	1	2	3	4
学習データ	3868	3868	3868	3868	3868
検証データ	967	967	967	967	967
テストデータ	9761	5790	2242	1209	1348

5.2 実験結果

表データを用いて学習に利用する各種キーワードの種類を徐々に増やした際に, どのように精度の向上が行われるのかを図4に示す.

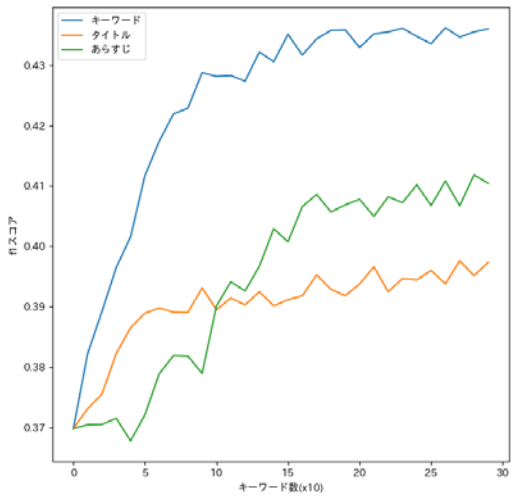

図4 キーワード種別ごとの利用数と精度の関係

図4によれば, 全体的にキーワードを用いた際の精度の向上が他に比べて大きいことがわかる. また, キーワード・タイトルキーワードに関しては, 利用する頻度上位単語が100個を超えるあたりまで大きく精度の向上が見られ, あらすじキーワードに関しては, 頻度上位50位までのものは精度向上に寄与しなかったが, その後精度の向上が見られる. これは, あらすじは比較的長い文章が多いため, 頻度が上位の単語は一般的に用いられる単語となってしまうため, あるあらすじの特徴をよく表すものとなっていなかったことが原因だと考えられる.

また, キーワード・タイトル・あらすじ全てを入力系列に含めて学習した際の, テストデータに対する各ラベルに対応する正答率を表6に示す. 併せて, キーワード(K)・タイトル(T)・あらすじ(S)を組み合わせて入力系列に含めて学習した際の, テストデータに対する各ラベルに対応するf1スコアの比較を図5に示す.

表6 全ての特徴量を入力に利用した際の精度

	precision	recall	f1-score	support
0	0.81	0.59	0.68	9409
1	0.42	0.51	0.46	5685
2	0.28	0.36	0.32	2220
3	0.26	0.48	0.34	1199
4	0.69	0.51	0.58	1344

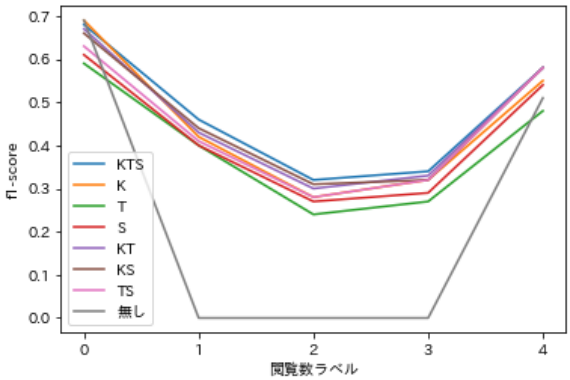

図5 特徴量を組み合わせて学習させた際の閲覧数ラベルごとのf1スコア

図5によれば，各種キーワードを含む場合とそうでない場合で，閲覧数が一番少ないグループと一番多いグループに対する予測精度は変化がなかったが，一方で閲覧数が中間のデータに対する予測精度に大きな乖離がみられた．

6. 考察

本研究では，小説のCGMにおける読者の段階的行動，つまり作品群から作品を見つけ出し，読んで評価をつけるまでのプロセスが段階的なものであるとの仮説のもと，投稿されている短編小説の閲覧数と，タイトル等の作品一覧ページから確認が可能な作品の表層の情報との関係性をみてきた．本章は，前述の内容を踏まえた上で，実験結果から小説のCGMにおける読者の行動について考察を行う．

6.1 初見から評価は段階的なプロセスなのか

ユニーク閲覧数と総合ポイントとの対応の図2を見てみると，

・概ね比例関係にある
・閲覧数は高いが総合ポイントが少ない作品が散見される
・閲覧数は低いが総合ポイントが高い作品は比較的少ない

ことがわかる．この関係性から，読者は自分の読書経験に従い，作品の一覧にて表示される情報をもとに作品を選び出す．そして概ね予想は当たる傾向にあるが，一部内容が期待とは異なった場合に評価を付けない，という段階的なプロセスに分かれていることが見て取れる．

6.2 読者は何の情報をもとに作品を選ぶのか

本研究において利用した2種類のモデルの両方における結果から，閲覧数の予測を行う上で精度に大きく寄与した特徴量はキーワードであることがわかる．この結果はタイトル・あらすじが大事であるというアンケート評価の結果に反しているため，キーワードがなぜ特徴量として重要なのかと，タイトル・あらすじはなぜ重要ではなさそうなのかについて考察を行う．

まず，キーワードがなぜ特徴量として重要なのかについて考察を行うべく，ユニーク閲覧数ラベルごとに，上位キーワード・タイトルキーワード・あらすじキーワードのそれぞれにおいてある作品に含まれる頻度上位300語の出現頻度と，出現頻度を作品に登録されているキーワードの数，あるいは形態素から抽出したキーワードの候補で割ったものを密度として算出した（表7~9）．

表7 閲覧数ラベルごとの上位キーワードの出現頻度の分布

閲覧数ラベル	0	1	2	3	4
データ数	48803	28950	11209	6044	6741
頻度の平均	3.39	4.24	4.72	5.10	5.46
頻度の分散	2.68	2.93	3.10	3.08	3.21
密度の平均	0.74	0.69	0.66	0.68	0.70
密度の分散	0.32	0.30	0.28	0.26	0.25

表8 閲覧数ラベルごとのタイトルキーワードの出現頻度の分布

閲覧数ラベル	0	1	2	3	4
データ数	48803	28950	11209	6044	6741
頻度の平均	1.05	1.60	2.41	3.12	3.67
頻度の分散	1.34	1.91	2.53	2.71	3.07
密度の平均	0.28	0.32	0.38	0.44	0.49
密度の分散	0.30	0.29	0.27	0.25	0.24

表9 閲覧数ラベルごとのあらすじキーワードの出現頻度の分布

閲覧数ラベル	0	1	2	3	4
データ数	48803	28950	11209	6044	6741
頻度の平均	7.97	9.67	11.39	13.07	14.61
頻度の分散	6.44	7.00	7.64	8.13	8.49
密度の平均	0.41	0.40	0.40	0.40	0.41
密度の分散	0.18	0.16	0.14	0.13	0.13

表7~9から，閲覧数ラベルが大きい，つまり閲覧数が多いグループほどある作品における頻度上位語の数は増える傾向にあり，出現頻度ではあらすじを用いた時が一番多い．しかし，各特徴量ごとの候補となるキーワードの数で割った密度に注目すると，上位キーワードを用いた際の数値がタイトル・あらすじを用いた際の数値に対して大きく差をつけている．ここから，著者が自身の作品に対してラベル付けするキーワードは，読者に対して訴求可能な情報として密度が高いことがわかる．これは作品に関するキーワードが，「R15」等の作品の内容を表さない抽象的なものであってもよい点や，タイトルやあらすじと違って人気作と同様のタグをつけても問題になりにくい点などが原因となっている可能性が考えられる．

続いて，タイトルに関して図7から，閲覧数ラベルが大きくなるほどタイトルキーワードの出現頻度・密度の両方が大きくなっているが，タイトル自体の長さを調べ

てみると，閲覧数ラベルが大きくなると同時に長くなっていることが確認できた（表10）．

表10 閲覧数ラベルごとのタイトルの長さの分布

閲覧数ラベル	0	1	2	3	4
データ数	48803	28950	11209	6044	6741
平均	10.59	13.97	18.87	22.62	24.98
分散	8.70	11.74	15.83	16.99	19.82

ここで，2017年の6月以降に投稿された作品の，2017年6月からの経過月数とタイトルの長さの平均の関係性を図6に示す．これによれば2020年の10月頃の前後で急激にタイトルの長さが長くなる期間があり，それ以降は同じ水準で推移している．図6と表8，表10の結果から，タイトルの長さ及び頻度上位の単語を用いる割合が増えたことで，タイトル自体がコモディティ化しており，それによってタイトルがそれを構成する単語のレベルでは読書行動の決定的な要因にはなりえないことが考えられる．

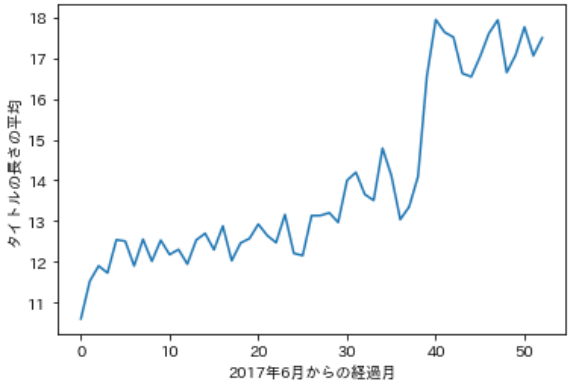

図6 タイトルの長さの平均値の推移

最後にあらすじについて，表9によれば頻度上位語があらすじの中に出現する頻度の平均は増えているが，密度の平均は閲覧数ラベルによらず一定である点から，あらすじ自体の長さの増加割合とほとんど一致している．

図4の結果では，出現頻度が多くなるにつれて徐々に分類性能が向上し，タイトルを用いた場合の精度を最終的に上回る点から，前述の「タイトルがコモディティ化している」という可能性が事実の場合，読者は読むかどうかをタイトルよりもあらすじを用いて差異を確認し，最終判断を下すというステップが確認できる可能性がある．

6.3 閲覧数が低い作品の要因の分析

本研究では，LightGBM，BERTの両方のモデルにおいて閲覧数ラベルが0のものと4のもの，つまり最も閲覧数の少ないグループと最も閲覧数の多いグループに対するf1スコアが高く，他のグループに対して識別が容易であった．さらに，閲覧数ラベルが0のデータに対するprecisionが特に高い値となっている点から，閲覧数ラベルが0のグループに所属する強い要因が存在することが考えられる．

ここで，特徴量が目的関数の結果の減少に対してどれだけ寄与したか，つまりある特徴量が閲覧数ラベルの推測にどれだけ影響を及ぼすかを示すfeature importanceを出力したところ，ジャンルの情報が大きな影響を持っていたことがわかった．そのため，ジャンルとそのジャンルに属する作品の閲覧数の平均を表11に示す．

表11によれば，「異世界恋愛」ジャンルが他のジャンルと比べて特筆して人気であることがわかったため，この「異世界恋愛」ジャンルを除いたジャンルにて表データを用いた予測を行った．その結果を表12に示す．

表12によると，閲覧数ラベルが4，つまり閲覧数の最も多いグループに対する予測精度が大きく下がっているが，閲覧数ラベルが0のデータに対する推論精度には特筆すべき変化はない．

表11 ジャンルと閲覧数の関係

小ジャンル〔大ジャンル〕	データ数	平均値	中央値
異世界〔恋愛〕	11262	25567	6227
現実世界〔恋愛〕	16271	2794	253
ハイファンタジー〔ファンタジー〕	9557	3483	345
ローファンタジー〔ファンタジー〕	6187	733	156
純文学〔文芸〕	8898	962	152
ヒューマンドラマ〔文芸〕	20364	1107	131
歴史〔文芸〕	1613	962	240
推理〔文芸〕	1435	1337	226
ホラー〔文芸〕	10254	658	191
アクション〔文芸〕	1147	517	162
コメディー〔文芸〕	9558	1977	204
VR ゲーム〔SF〕	367	2155	307
宇宙〔SF〕	770	759	192
空想科学〔SF〕	3298	438	150
パニック〔SF〕	766	763	199

続いて，さらに各種キーワードも含めない場合の学習・推論の結果を表13に示す．表12，表13の結果の，閲覧数ラベルが0の作品に対する予測精度に変化があまり見られない点から，そもそも閲覧数の低いグループに属する作品は，ある特定のキーワードが入っているかどうか等は問題ではなく，それ以前の問題を抱えていることがわかる．

表12 異世界恋愛ジャンルを除いた学習・推論の結果

閲覧数ラベル	precision	recall	f1-score	データ数
0	0.75	0.64	0.69	9674
1	0.42	0.35	0.38	5429
2	0.18	0.23	0.20	1795
3	0.10	0.20	0.14	746
4	0.17	0.59	0.27	453

表13 異世界恋愛ジャンル・キーワードを除いた学習・推論の結果

閲覧数ラベル	precision	recall	f1-score	データ数
0	0.72	0.58	0.64	9674
1	0.40	0.26	0.31	5429
2	0.15	0.21	0.17	1795
3	0.10	0.23	0.14	746
4	0.09	0.53	0.16	453

そこで，具体的な単語の情報を使わず，タイトルの長さ・あらすじの長さ・本文の長さ・投稿年・ジャンルを用いて同様の条件で実験を行い，その結果を機械学習モデルの説明性を示すための手法であるSHAP[11]を用いて，特徴量ごとに各クラスとして推測を行う寄与度を算出した（図7）．図7は，表13からもわかる通り，閲覧数ラベルが0以外のラベルについては推測の精度が低いため信頼性は薄いが，総合して推測に寄与する特徴量としてはタイトルの長さが一番重要視される点と，あらすじの長さが推測に寄与しない点から，読者は最低限の読みたい気持ちの検証を，タイトルによって行っていると考えられる．

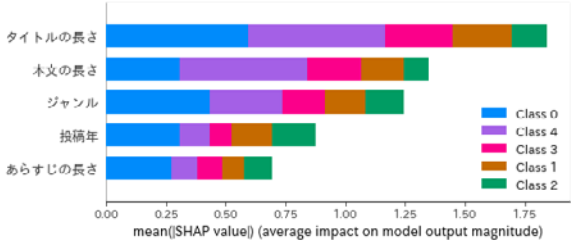

図7 特徴量ごとの推測に対する寄与度

7．結論

本研究では，読書をして評価を行う行動が「読むかどうかを決める」「読んでどうかを評価する」という二段階のプロセスからなることを仮定し，特に前者の読むかどうかに寄与する要因を分析するべく，作品の表層の情報と閲覧数の関係性を考察した．

その結果，閲覧数と評価を表すポイントの関係性からCGMにおける読者の，評価をつけるまでの読書行動は段階的である点，読者が大事だと認識しているタイトルやあらすじよりも著者が自由に編集できるキーワードが閲覧数関与している点，閲覧数が極端に低い作品は，最低限含むべき情報がタイトルに入っていない可能性がある点が明らかとなった．

参考文献

[1] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 19, 1–10. https://doi.org/10.1145/3411763.3450391
[2] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic Story Generation: Challenges and Attempts. In Proceedings of the Third Workshop on Narrative Understanding, pages 72–83, Virtual. Association for Computational Linguistics.
[3] 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.
[4]小坂直輝，小林哲則，林良彦，”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”,2022.
[5] “小説家になろう”. https://syosetu.com/, (参照 2022-8-31).
[6] “小説を読もう”. https://syosetu.com/, (参照 2022-8-31).
[7] “KASASAGI”. https://kasasagi.hinaproject.com. (参照2022-8-31).
[8] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang,Wei Chen, Weidong Ma, Qiwei Ye1, Tie-Yan Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 2017
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018
[10] “MeCab”, https://taku910.github.io/mecab/, (参照2022-8-31).
[11] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.