オンライン小説の傾向分析のための創作の連鎖関係の分析

小川 明水[1,a)]　伊東 栄典[2,b)]

概要：動画・イラスト・音楽・小説などのコンテンツを投稿・閲覧する CGM サイト (Consumer Generated Media) が人気である。CGM サイト内で人気を得たコンテンツが，ネットや実世界に広がるようになっている。例えば人気を得たオンライン小説が，漫画やアニメになり，多くの閲覧者を得ている。内容が多様で，かつ品質が均質でないオンライン小説群から，人気小説の傾向（流行の傾向）を抽出したり，将来人気となる小説を見つけることが出来れば，商機に繋がる。オンライン小説では，ある人気小説に触発された作家が，その人気小説のオマージュ小説を投稿することで，流行が始まることが有る。本研究では，ある人気小説に触発されて次々と生み出される小説群の繋がりを創作の連鎖と呼ぶ。本研究では，オンライン小説の将来流行予測や人気小説発見を行う前提として，小説の連鎖関係に基づくオンライン小説の傾向を分析する。

キーワード：利用者投稿型メディア, オンライン小説, 傾向分析, 類似小説, 創作の連鎖

Yoji Ogawa[1,a)]　Eisuke Ito[2,b)]

Abstract: Recently, CGM (Consumer Generated Media) become very popular in Japan. In case of online CGM novels, some popular contents are not only popular on the Internet, but also real worl. In online novels, some novels on a specific topic may be newly posted, simultaneously but asynchronously. Because, a writer, who is inspired by a popular novel, may post a homage novel of the previous popular novel. In this research, we focus on homage relation between novels, and we try to extract a similar novel group based on homage chain. In this study, we define the homage chain relationship using bookmarks, feature words of novel metadata, and text similarity in novel contents. In this paper, We report our proposing method, and small size experiments.

Keywords: CGM, Online novels, trend analysis, similarity, homage chain

1. はじめに

近年，利用者がコンテンツを制作・提供する CGM (Consumer Generated Media) が人気である。動画では Youtub やニコニコ動画，小説では「小説家になろう」やカクヨム，イラストでは pixiv などがある。本研究は小説 CGM である「小説家になろう」に着目する。

CGM 小説サイトでは誰でも小説を投稿できる。編集者の選別が無いため，作者の興味や関心に依拠する小説が投稿される。その結果，投稿小説のジャンルに偏りがあり，また小説の品質にばらつきが在る。この多様性の中で，読者はそのときに流行しているジャンルの小説を閲覧しがちである。作者も読まれやすい流行の小説を書きたがる傾向がある。

既存小説群から過去の流行を分析し，それにより将来の流行を予想できれば商機に繋がる。流行分析のために類似小説の抽出を行う。本研究では創作の連鎖関係に基づく類似小説群の抽出を試みる。創作の連鎖関係については第 3 節で述べる.

2. 対象データ

本研究では小説 CGM の「小説家になろう」に投稿された小説を分析する。ここではサイトの概要と，データ収集について述べる。

2.1 「小説家になろう」

「小説家になろう」？はヒナプロジェクト社が提供す

[1] 九州大学工学部電気情報工学科
[2] 九州大学情報基盤研究開発センター
[a)] ogawa.yoji.018@s.kyushu-u.ac.jp
[b)] ito.eisuke.523@m.kyushu-u.ac.jp

る小説 CGM である．2021 年 1 月 13 日，掲載小説数は 784,500，登録者数は 1,976,111 人である。Wikipedia の記事? によると，2019 年 4 月時点で月間アクセス数は約 20 億である。サイトに投稿されて公開された小説は誰でも読むことが出来る。サイトに利用者登録すると，自分の小説の投稿，他者の小説のブックマークが出来る。このサイトの小説は「なろう小説」と呼ばれており，人気の「なろう小説」は紙の書籍として出版されたり，マンガやアニメの原作になることもある。

2.2 小説メタデータの収集

分析するために，小説のメタデータ，本文（２話と３話のみ），読者のブックマークを収集した。収集した期間は 2020 年の 11 月である。

メタデータは，タイトル・作者名・キーワード・あらすじ・投稿日を含む。メタデータ取得用の Web API が公開されている。この API は「なろう API」と名付けられている。API に小説の識別子である Ncode を送ると，Ncode に対応する小説のメタデータを取得できる。表 1 に収集したメタデータの概要を示す。

表 1 収集メタデータ概要

項目	内容
期間	2004 年 4 月 20 日〜2020 年 12 月 01 日
形式	json 形式
データ件数 (小説数)	774,568
最終収集日	2020 年 12 月 01 日

図 1 に「小説家になろう」への月ごとの小説投稿数を示す。図 1 のグラフは収集データを分析して得たものである。新規小説の投稿数を見ると，短期的な減少はあるものの長期的には単調増加を続けている。2020 年には毎月約 1 万小説が新規投稿されている。

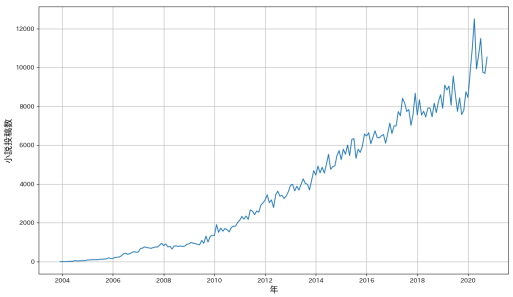

図 1 月毎の新規小説投稿数

2.3 小説本文の収集

後述する小説の類似度を算出するために，小説の本文を収集した。連載小説は複数の「話」から成る。そこで本文の第２話と３話のみ収集した。収集プログラムは Python 言語を用いて作成した。表 2 に収集した本文データの概要を示す。

第１話は集めていない。第１話には登場人物紹介や地名案内など，小説内容と異なる文が含まれることが多い。そのため文章の類似度算出に第１話は使いにくい。第４話以降を集めなかった理由は，アクセス制限のためである。１つの「話」をダウンロードすると，数秒間，新たな「話」をダウンロードができない。そのため２話と３話のみに制限した。同じ理由で，第 4 節の実験対象とした「歴史」ジャンルの小説に限定した。

表 2 収集本文データ概要

項目	内容
最終収集日	2020 年 12 月 01 日
対象小説数	48,869
話数 (小説数 x2)	97,738
形式	plain text

2.4 読者ブックマークの収集

最後にブックマーク収集について述べる。「小説家になろう」に登録した利用者は，読者としてお気に入り小説をブックマークできる。多くの読者ブックマークは公開されている。そこで，プログラムを作成してブックマークを収集した。ブックマーク収集は２つの Python プログラムで実現した。１つ目のプログラムは，全利用者のブックマーク（１ページ目）を収集する。1 ページ目を見ると，その利用者のブックマーク登録小説数が分かる。2 つ目のプログラムで，特定利用者の全ブックマークを収集する。

本研究は類似小説の抽出を目的とする。多くの読者が一緒にブックマークする小説は類似度が高いと想定できる。これは本段の特定部分に同一ジャンルの本を置くことに相当する。本棚の本が多い読者ほど，より目利きの読者であろう。同様にブックマークに登録する小説数が多い読者ほど目利きの読者であろう。そこで，ブックマーク登録小節数が 100 以上の読者だけ，ブックマークを収集した。

表 3 に収集したブックマークデータの概要を示す。

表 3 収集ブックマークデータ概要

項目	内容
最終収集日	2020 年 10 月 01 日
対象利用者数	169,699
のべ小節数	406,799,767
形式	HTML

3. 創作の連鎖関係

小説に限らず音楽や映画も含め，コンテンツの作成者は，既存コンテンツに影響を受けることが多い。例えば映画では，ジョージ・ルーカス監督の「スターウォーズ」や宮崎駿監督の「もののけ姫」は黒澤明監督の「七人の侍」に影響を受けているとの記事がある。またジェームズ・キャメロン監督の「AVATAR」は「スターウォーズ」や「もののけ姫」に影響されているらしい。他にも「オマージュ」や「リスペクト」と記すことで影響を公言する場合も多い。このような関係を，本研究では「創作の連鎖関係」と呼ぶ。

「小説家になろう」でも，既存の小説に影響を受けることが多い。そこで，創作の連鎖関係を４つの値を用いて定義することにした。

3.1 ブックマーク順序を考慮した共起数

先に述べたように，本研究は類似小説の抽出を目的とする。多くの読者が一緒にブックマークする小説は類似度が高いと想定できる。そこで，多数の読者ブックマークに共起出現する小説は類似度があるとする。ただし，順序関係も考慮する。

利用者２人が小説 a, b, c, x を以下の順序でブックマークしているとする。

```
u1  : < a, b, c>
u2  : < c, a, x, b>
```

普通の共起ペアであれば，(a,b), (a,c), (b,c) が２つのブックマークで共起出現したとする。本研究では順序を考慮するため２つのブックマークに共起したのは (a,b) のみとする。

順序制限を使う理由は，創作の連鎖を考慮するためである。小説 a の影響を受けて小説 b が作られたとする。この場合，b は a より後に投稿される。a と b の第１話投稿日時を a_1, b_1 とすると，は $a_1 < b_1$ となる。多くの読者が a と b を読み，かつブックマーク登録する場合，a は b より前に登録されることが多いだろう。逆順で登録する読者も居るだろう。しかし，多くの読者のブックマークを数え上げる場合，逆順にした読者は考慮しなくて良いと思われる。

3.2 小説メタデータに出現する特徴的な単語

「小説家になろう」では，全ての小説についてメタデータが存在する。メタデータには，小説タイトル，作者名，あらすじ，キーワードが含まれる。これらの項目には，小説の特徴を表す単語が含まれることが多い。読者の多くは，ある小説を読むか読まないかを決める際，小説を紹介するメタデータの文章を参考にする。読者は，自分の趣味に合う小説のうち，面白そうな小説を選ぶ。作者も自分の小説を読んで欲しいため，読者の興味を引きやすい単語入れた紹介文を作成する。

小説メタデータから抽出した特徴的な単語が，小説 a と b で類似する場合，a と b の２つは内容も類似する可能性が高い。そこで，小説メタデータの中で，「タイトル，あらすじ，キーワード」の文から，単語を抽出する。タイトルとあらすじには形態素解析ツールを適用して単語を抽出する。キーワードは，もとから単語の集合なので，そのまま使う。形態素解析で抽出する単語のうち，副詞は用いない。また，「小説家になろう」で出現頻度の高すぎる単語は不要語として除外する。

3.3 小説本文の類似度

本研究で着目する「創作の連鎖」では，影響を与えたコンテンツと，影響を受けたコンテンツの双方で，コンテンツ内容が似ている可能性がある。小説の場合，小説本文が内容になるため，本文の類似度も考慮する。

小説本文として，第２話と第３話だけを収集した。第１話には登場人物紹介や地名案内など，小説内容と異なる文が含まれることが多いため，集めていない。

小説本文（２話と３話）の類似度を計算する際，文書の分散表現（ベクトル化）を利用できる。単語の分散表現である Word2Vec を用いた文書のベクトル化や，Doc2Vec による文書のベクトル化手法が提案され，ツールも提供されている。

これらのツールを用いて Word2Vec による単語の分散表現の算出と，Doc2Vec による文書の分散表現算出する。

3.4 期間の制限

創作の連鎖は，比較的短期間に発生するように思われる。例として「ログ・ホライズン」と「オーバロード」の２つを考える。どちらも人気小説で，かつ後に書籍化・漫画化・アニメ化されている。この２作は両方とも多数プレイヤーが参加するオンラインゲーム世界に人間が転移される小説で，かつどちらも 2010 年に第１話が投稿されている。つまりこの２つは特定の小説[*1]に影響を受けて作成された可能性がある。

他にも「無職転生，この素晴らしい世界に祝福を，Re:ゼロから始める異世界生活」の３作品も類似している。いずれも書籍化・漫画化・アニメ化された人気作品である。３作とも 2012 年に第１話が投稿された，異世界に転生する小説である。この３つも，ある小説に影響を受けたのかもしれない。あるいは３つの投稿時間順に，影響を及ぼしたのかもしれない。

創作の連鎖関係を機械的に抽出する場合，第１話の投稿日時の時間差を制限する。ブックマークでの共起，特徴語の重なり，本文の類似度から，小説 a が b に影響している可能性が高いと計算される場合を考える。このとき，a と

[*1] 小説「ソードアート・オンライン」かもしれない

b の投稿時間差を一定期間以内に制限する。具体的には 1 年以内とする。a と b の第 1 話投稿日時を a_1, b_1 とすると，第 1 話の投稿時間差は $b_1 - a_1$ を 1 年以内とする。この制限は，創作の連鎖関係の候補を減らすために導入する。

4. 実験および評価

本論文で提案する創作の連鎖関係に基づく類似小説グループ抽出を実験する。「小説家になろう」の全小説を対象にした実験では多すぎる。そこで「歴史」ジャンルの小説で，人気順に 50 位までの小説 50 個を選び，それらを起点とする創作の連鎖関係を抽出した。

抽出条件を変化させ，3 つのグループ作成方法を比較する。3 つの作成方法を表 4 に示す。

表 4 3 つの方法

連鎖の条件	方式 A	方式 B	方式 C
1. ブクマ共起数	500 以上	500 以上	500 以上
2. 特徴語の重なり	あり	**なし**	あり
3. 本文の類似度	Doc2Vec	Doc2Vec	**Word2Vec**
4. 投稿間隔	1 年以内	1 年以内	1 年以内

4.1 適合率

まず，3 つの方式について，適合率 (precision) で評価する。適合率は，抽出器が「創作の連鎖関係である」と判断したもののうち，真に連鎖関係であるものの割合である [1]。機械的に抽出された 50 個グループ x 3 方式＝ 150 グループを全て調査するのは困難であった。そこで，各方式から 10 グループを抜き出し，そのグループに選ばれた小説が「創作の連鎖関係」で繋がるものかを人手でチェックした。その結果を表 5 に示す。

表 5 3 つの方法の適合率

	方式 A	方式 B	方式 C
適合率	52.77	39.69	26.00

適合率では方式 A が最も高い。方式 A の B の違いは，方式 A では「2. 小説メタデータの特徴語の重なり」を使い，方式 B では使わないことである。この結果から，特徴語の重なりは使う方が良いことが分かる。方式 A の C の違いは，方式 A が「3. 小説本文の類似度」の算出に Doc2Vec を使い，方式 C は Word2Vec を使うことである。この結果から，本文の類似度算出には Doc2Vec が良いことが分かる。

ただし，最も成績の良い方式 A でも適合率は約 53%であり，それほど高い値とは言えない。創作の連鎖関係による類似小説抽出を実現するには，パラメータなどの調整が必要である。

4.1.1 小説数と連鎖関係数

最後に，方式 A,B,C で，創作の連鎖関係で結ばれたグループのノード数（小節数）と枝数（連鎖関係）を調べる。図 2 にノード数と枝数での散布図を示す。図をみると，どの方式でもノード数と枝数は比例している。また，各方式で散布に大きな違いはない。

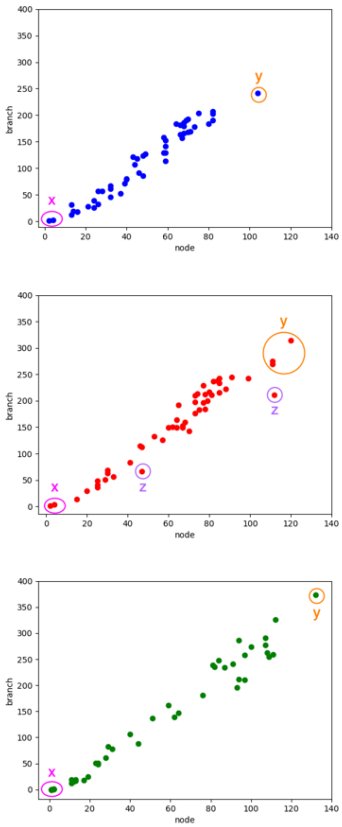

図 2 グループのノード数（小説数）と枝数（連鎖関係の数）

5. おわりに

我々は小説 CGM サイトの流行分析を目指している。本論では，流行分析のために，創作の連鎖関係に基づく類似小説グループの抽出を試みた。創作の連鎖関係を，ブックマーク共起数，小説メタデータにおける特徴語の重なり，小説本文の類似度，第 1 話の投稿日の 4 つの数値を用いて定義した。条件を変えた 3 つの方式を実装し，小説家になろうの「歴史」ジャンル小説に適用した。最も成績の良い方式 A でも適合率は約 53%で高い値とは言えない。創作の連鎖関係による類似小説抽出を実現するには，パラメータなどの調整が必要である。今後はパラメータの調整や，本文の増量などを検討したい。

参考文献

[1] 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版 (2002).