

利用者投稿型小説サイトにおける キーワードの多様性動向分析

本田 優也^{1,a)} 伊東 栄典^{2,b)}

概要: 近年、動画では YouTube やニコニコ動画が、小説では小説家になろうといった利用者投稿型の CGM (Consumer Generated Media) サイトが人気である。CGM サイトには毎日多数のコンテンツが投稿されており、また膨大な利用者がコンテンツを閲覧している。現在、CGM サイトへの投稿コンテンツの画一化が懸念されている。既に見たことのあるようなコンテンツや派生コンテンツの増加が感じられる。我々は今回、「小説家になろう」を対象に、オンライン小説の多様性動向を分析した。小説に付与されたキーワードについて、単語のエントロピーと類似度に着目して分析した。分析の結果、類似度の総和の上昇を観測でき、多様性の減少を定量的に評価できた。

キーワード: 利用者投稿型サイト, CGM (consumer generated media), 多様性, エントロピー, 類似度

Keyword diversity trend analysis on consumer generated novel site

YUYA HONDA^{1,a)} EISUKE ITO^{2,b)}

Abstract: Recent years, CGM (Consumer Generated Media) sites, such as YouTube and nicovideo for movies, syosetu.com for novel stories, become very popular. A lot of contents are posted to a CGM site every day, and also a large number of users are enjoying posted contents. At present, decreasing diversity of contents are mentioned by some bloggers. Some posted new content may be similar with previous posted contents. In this paper, the authors proposed two quantitative metrics of contents diversity, and applied them to the contents in syosetu.com. They focused the keywords given to the novel, and calculated entropy and similarity of keywords. As the results, they observed increase of similarity, and it shows decrease of diversity of contents.

Keywords: CGM (consumer generated media), online novel, diversity, entropy, similarity

1. はじめに

近年、YouTube やニコニコ動画、小説家になろうなどの利用者投稿型共有サービスが人気である。これらのサイトは CGM (Consumer Generated Media) とも呼ばれる。サービス開始から数年経過した CGM サイトは、社会に大きな影響を与えるメディアに成長している。CGM サイト

に毎日多数の動画や小説が投稿されており、また膨大な利用者がそれらを閲覧している。動画や小説だけではなく、静止画、マンガ、写真、レシピの CGM サイトもある。

主な小説投稿サイトとして、小説家になろう、E★エブリスタ、pixiv 小説、comico、魔法のiらんどがある [1]。本研究では、小説の CGM サイトである「小説家になろう」を対象とする。「小説家になろう」の利用者は増加しており、また投稿小説数も多い。人気小説は紙の書籍としても出版されている。また人気小説は、紙の書籍以外にも、マンガやアニメになる場合もある [2]。

我々は既に、ニコニコ動画を対象に、コンテンツの多様

¹ 九州大学大学院ライブラリサイエンス専攻
Dept. of Library Science, Kyushu University
² 九州大学情報基盤研究開発センター
Research Institute for IT, Kyushu University
^{a)} honda.yuya.128@s.kyushu-u.ac.jp
^{b)} ito.eisuke.523@m.kyushu-u.ac.jp

性動向について調査してきた [3][4]。また、「小説家になろう」について、小説数やリンク数などの統計分布 [5] や、リンク構造に基づく小説推薦 [6] について研究してきた。

「小説家になろう」の小説が人気になり紙の書籍として販売され始めた頃から、ネット小説の多様性が失われているのではないかという意見が出てきている。ニコニコ動画を運営するドワンゴ社の川上量生氏も、cakes の記事 [7] で「小説家になろう」における小説ランキング上位の小説が画一化傾向を指摘している。また Fukatsu 氏のブログ記事 [8] においても、「小説家になろう」に投稿される小説の画一化に対して考察している。実際に「小説家になろう」における累計ランキング上位 100 位 [9] を見ると、ほとんどがファンタジーな異世界に転生する小説であると分かる。コンテンツの多様性が減少して画一化が進むと、文化的多様性が失われる恐れがある。特定環境に特化し過ぎて多様性を失った文化からは、新たなイノベーションが発生しにくいと思われる。

三宅は「小説家になろう」で小説の多様性が失われている理由を考えている [10]。「小説家になろう」では登録すれば誰でも気軽に小説を投稿できる。現状や歴史に基づく小説を書く場合、調査に基づく知識と、それを記述する能力が必要となる。対してファンタジー作品は作者が設定した世界であり、読み手に自身の世界を説明するだけで良く、またその説明も難しいものではない。ゲームやマンガ等を体験した読者は、ファンタジー世界の世界観を共有しやすい。この理由からファンタジーのジャンルを作者は作りがちなのではないかと考えられる。

本研究では、コンテンツの多様性を定量的に分析する。多様性を定量的に示すため、情報エントロピーと cos 類似度に基づく尺度を提案する。これら二つの尺度を、小説に付与されたキーワードを対象に算出する。月毎の小説群についてキーワード欄の単語を抽出し、単語の多様性を算出した。二つの尺度による数値の推移を見ることで、コンテンツの多様性を分析する。

2. 対象サイト：小説家になろう

小説家になろうは、株式会社ヒナプロジェクトが提供する小説投稿サイトである。利用者登録すれば、無料で小説をサイトで公開できる。2004 年のサイト開設当初は個人サイトとしての運営されていた。その後のアクセス増加により、2008 年からグループによる運営に移行し、2010 年に正式に法人化した。Wikipedia[11]によると、2014 年 12 月時点で、アクセス数は月間約 9 億 5000 万 PV、ユニークユーザー数は 400 万人である。また 2015 年 03 月 15 日朝、登録者数が 55 万人を突破した。

小説家になろうにおける小説は「なろう小説」と呼ばれている。「なろう小説」の一部人気が出たものに関しては小説やマンガ、アニメになることもある。実際にマンガや

アニメとなった作品のいくつかを表 1 示す。

表 1 マンガ、アニメとなったなろう小説の例

作者名	小説名
長月達平	Re:ゼロから始める異世界生活
暁なつめ	この素晴らしい世界に祝福を！
丸山くがね	オーバーロード
大森藤ノ	ダンジョンに会いを求めるのは間違っているだろうか
佐島勤	魔法科高校の劣等生
橙乃ままれ	ログ・ホライズン
理不尽な孫の手	無職転生-異世界行ったら本気出す-

表 1 の作品は全てファンタジーなものである。なろう小説はファンタジー作品に人気が出やすく、マンガやアニメなどになりやすいように思われる。作者たちは比較的人気が出やすく、人気が出た後の展開もあるファンタジー作品を作ることが多くなると予想される。これもまた小説家になろうにおいて多様性が失いやすくする要因の一つでないかと考える。

「小説家になろう」における小説のメタデータに含まれる項目を表 2 に示す。

表 2 小説情報に含まれる項目

項目	説明
Ncode	小説コード (小説の識別子)
Title	小説の題名
Story	小説のあらすじ
Writer	作者
Keyword	キーワード
Genre	ジャンル
Userid	作者 ID

3. 小説データ収集

小説家になろうにおける小説データは、YAML 形式で提供されている。YAML 形式は構造化データを表現するフォーマットである。

小説家になろうが提供する「なろう API」[12]を用いて、YAML 形式ファイルの小説情報を収集するプログラムを Ruby 言語で作成した。小説毎に割り当てられている識別子 Ncode を指定すれば、その Ncode の小説情報を収集できる。しかし Ncode の割当てルールは自明ではない。一方、作者 ID は数値である。そこで作者 ID の数値を 1 から 45 万まで増しつつ、作者が投稿した小説の情報を収集した。作者 ID は約 40 万程度であることを事前に調べていたため、作者 ID の上限値を 45 万とした。45 万個のデータを効率よく収集するため、仮想マシン 8 台で分散して収集した。

収集した作者 ID 毎のデータに対して、その中から小説が登録されているものだけを用いるため、作者 ID 毎のデー

タより「allcount」値が1以上であるもののみをさらに抽出した。「allcount」とは作者情報に掲載される小説の投稿数である。その結果「allcount」が1以上であるデータが81449個であった。つまり、小説を一つ以上投稿している利用者は81,449人である。

最後に、集めた作者毎のデータから、小説識別子 Ncode 毎のデータ、つまり各小説のメタデータを作成した。データ収集、整形で得たデータの概要を表3に示す。

表3 収集データ概要

項目	内容
期間	2004年4月～2015年10月
形式	YAML形式
データ件数(小説数)	232,096
時点	2015年12月時点

4. 収集データの解析

収集した「小説家になろう」の小説情報メタデータを用いて、月ごとの小説投稿数、ジャンルの割合などを調査した。

4.1 収各月の小説投稿数

各月の小説投稿数を図4に示す。図4より2004年4月から2014年5月までの間、概ね右肩上がりに投稿小説数が増えている。その後小説投稿数が下がり始め2015年の小説の投稿数は月に2,500個前後となっている。

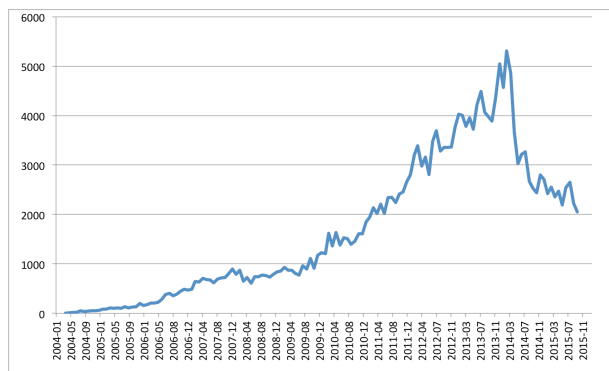


図1 各月の新規小説投稿数

4.2 小説投稿数に対する作者数

各作者の小説投稿数を調べた。投稿小説数の多い上位10人を表4に示す。図2には、投稿小説数と、その小説数を投稿した作者数の散布図を示す。投稿数の少ない散布図はだいたい直線になっているため、べき分布に近い分布である。

表4 小説投稿数の多い作者 (top 10)

順位	投稿数	作者 ID	作者名
1	1,327	26407	尚文産商堂
2	916	26055	神村律子
3	566	14644	竹仲法順
4	530	232918	蒼井深可
5	468	34969	百 (難しい童話)
6	421	21242	日下部良介
7	419	299670	さゆみ
8	383	153402	とーよー
9	359	369410	Clover の三ツ葉の方 (二百 or 四百文字)
10	328	147677	尖角

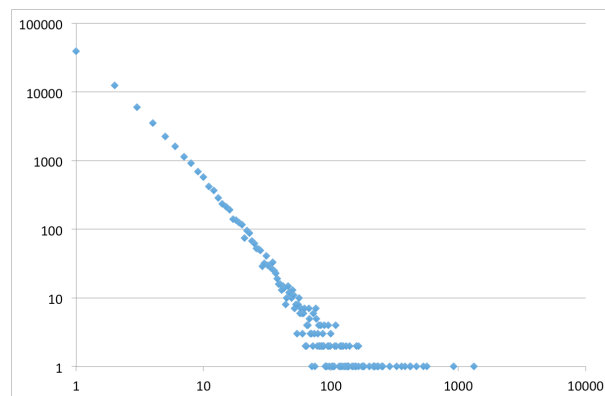


図2 小説投稿数-作者数 (両軸対数尺度)

4.3 ジャンル毎の割合

小説家になろうの小説には一部を除いて、作者が任意につけるジャンルが設定されている。このジャンルは全16種類あった。なお、2016年3月にジャンルは変更されており、本稿の分析は変更前のジャンルに基づく。

なお、ジャンルが指定がされていない小説もあり、その小説を0. ノージャンルとした。全体の小説に対するジャンル毎の小説の割合を円グラフで表したものを図3に示す。図3を見るとファンタジーと恋愛が大きな割合を占めていることが分かる。

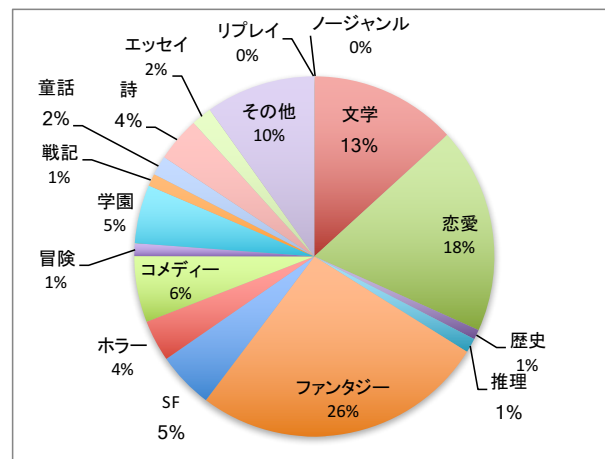


図3 ジャンルの割合

ジャンルの設定は作者が自由に設定できる。ジャンルが歴史である小説でも、歴史的時代に時間移動するファンタジー作品であることも多く、ジャンルの設定が曖昧な作品も多い。例として歴史ジャンルのランキング上位5番までを表6に示す[14]。

4.4 各月のキーワード数

各月に投稿された小説集合を対象に、キーワード欄に付与された単語について調査した。図4に各月のキーワード数を示す。グラフで青線は各月の一意なキーワード数を、赤線はキーワードの総数を示す。2014年まで一意なキーワード数、総数共に増加しておりその後減少している。増減は、各月の小説投稿数の増減と一致している。

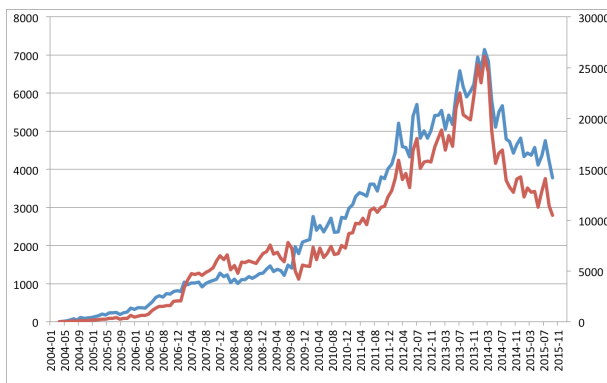


図4 各月のキーワード数 (青:一意な単語数, 赤:単語の延べ出現)

5. 情報エントロピーによる多様性動向分析

5.1 単語の情報エントロピー

シャノン¹は、自ら定義した情報エントロピーの考えを文書内の単語に適用し、英語文献における単語のエントロピーを推定している[13]。単語の情報エントロピーは式(1)で算出される。式(1)の $p(w)$ は単語 w の出現確率で、 W は全単語の集合である。 $H(W)$ の単位は bit/word である。

$$H(W) = - \sum_{w \in W} p(w) \log(p(w)) \quad (1)$$

5.2 単語情報エントロピーの動向

単語エントロピー $H(W)$ の値を、各月の投稿小説に付与されているキーワード欄の単語で算出した。図5の青線でエントロピーの推移を示す。図5の赤線は各月毎の小説投稿数である。

5.3 考察

図5を見ると、各月の小説投稿数は増加しているのに対し、単語エントロピーは8(bit/word)程度まで上昇したのち、一時減少しその後また10(bit/word)程度まで上昇して停滞している。シャノンの研究[13]でも、文書数を増やし

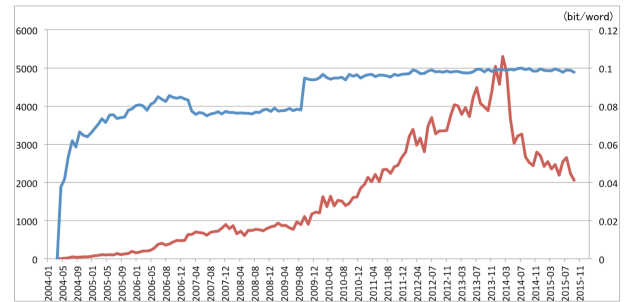


図5 各月の単語エントロピー (青線) と、投稿小説数 (赤線)

ても単語エントロピーは11bit/word程度より増えない事が分かっている。

図5で、単語エントロピーが急減・急増している期間は2007年4月～2009年10月である。この状況については6.5節で議論する。

6. cos 類似度による多様性動向分析

6.1 考え方

我々は、2つの小説感の距離もしくは類似度で、小説群の多様性増減を定量化できると考えた。まず何らかの数値ベクトルで各小説を表現する。図6に示す黒点1つは、小説1つに対応していおる。図の座標はベクトル値から決まる。もし図6の左図のように、点が散らばっている場合、2点間の距離は遠く、類似度は小さくなる。点が偏っている場合、2点間の距離は近く、類似度は大きくなる。

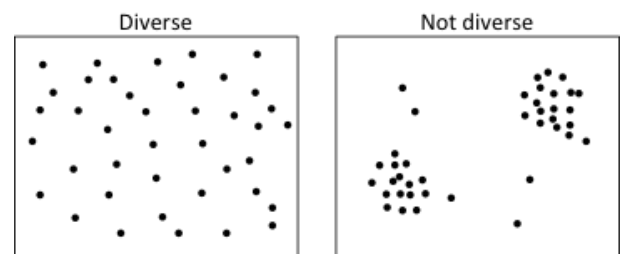


図6 多様性についてのモデル

二つの集合の距離、類似度を測定する方法として、ユークリッド距離、マンハッタン距離、cos 類似度、ピアソンの相関関係、Jaccard 係数、Dice 係数、Simpson 係数などが知られる[14]。集合間の類似性を「共通要素が多く、非共通要素が少ない」場合に大きいとすると、先に述べた手法のうち cos 類似度、Jaccard 係数、Dice 係数、Simpson 係数を集合の類似性の指標として扱うことができる。本研究では類似度の指標として最も用いられている cos 類似度を使うことにした。

6.2 cos 類似度

cos 類似度とは、2つのベクトル間の類似度を図る手法の一つである。文書をベクトルで表現すると、2つの文書の類似度を、文書ベクトルの cos 類似度で算出できる。cos

類似度は-1から1の値をとり、1に近いほど類似度が高い。

文書群を term-document matrix で表現する [15]。この行列を M とすると、行列の要素 $M(i, t)$ は、文書 i における単語 t の単語頻度である。二つの文書 i, j の $\cos$ 類似度は式 (2) で算出される。

$$\cos(i, j) = \frac{\sum_t M(i, t) * M(j, t)}{\sqrt{\sum_t M(i, t)^2} * \sqrt{\sum_t M(j, t)^2}} \quad (2)$$

「小説家になろう」では、キーワード欄には単語は高々1回しか出現しない。そのため $M(i, k)$ の値は0または1となる。

ある文書集合 D に含まれる全ての小説ペアについて、小説キーワードの $\cos$ 類似度を算出し、それらを足し合わせた値を算出した。この値を $SumCos$ とする。 $|D| = n$ の場合、ペア数は $nC_2 = n(n-1)/2$ 個となる。

$$SumCos(D) = \sum_{i=1}^{n-1} \sum_{j=i}^n \cos(i, j) \quad (3)$$

6.3 SumCos 値の推移

ある月 m に投稿が開始した小説集合を D_m とし、 $SumCos(D_m)$ を計算した。 D_m に含まれる小説が多い場合、小説ペアも多くなり、その結果 $SumCos$ 値も大きくなる。そのため小説数の影響を無くす必要がある。

最初に小説数を同数にした場合の $SumCos$ 値を算出した。各月の小説集合から、小説ランキング上位100位までの小説集合に限定した $SumCos$ 値を算出した。 $SumCos$ 値の推移を図7に示す。なお図7では、各月の小説投稿数が100個を超える2005年5月以降の推移を示す。

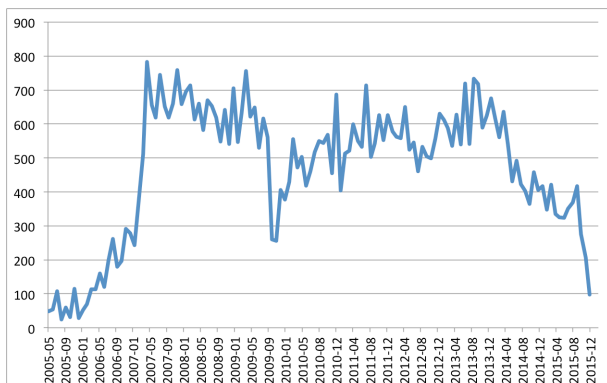


図7 SumCos 値 (上位100小説)

次に各月の全小説ペアの $SumCos$ を算出し、小説ペア数で割った平均値を算出した。平均 $SumCos$ 値の推移を図8に示す。

6.4 考察

図7に示した上位100小説における $SumCos$ 値を見ると、2007年頃に一度上昇したのち、2009年頃に減少し、そ

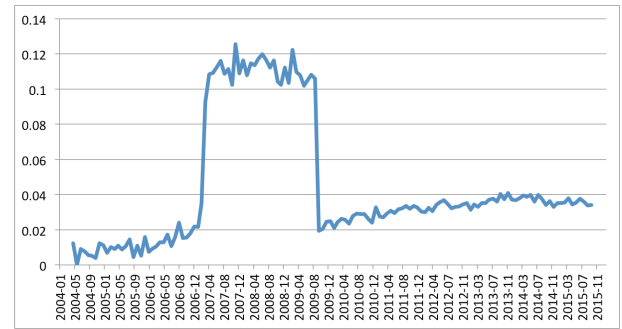


図8 SumCos 値 (平均値)

の後は増減を繰り返している。

図8に示した、平均 $SumCos$ 値の推移を見ると、2007年頃に大きく上昇し、その後2009年頃に大きく減少している。その後は緩やかに上昇している。

上位100個の $SumCos$ 値 (図7) では分かりにくいですが、平均 $SumCos$ 値 (図8) では、途中の急激な変化を外すと、増加傾向が続いていることが分かる。つまり、小説ペアにおいてキーワード欄に共通して出現する単語が増えたことがわかる。このことから小説のキーワードの多様性が減少傾向にあることを示している。

6.5 2007年4月～2009年9月についての考察

図8では、2007年3月から2007年4月に値が急増し、その後2009年9月から2009年10月に急減している。この期間について考察する。

図4を見ると、この期間は小説数の上昇より単語数の上昇が大きい。つまり、2007年4月から2009年9月の間、キーワード欄の単語が増加していたことを示している。これにより、図5に示す、キーワードの単語エントロピーも大きく変化している。また、同時に $SumCos$ 値も変化している。ただし、この期間にキーワード単語が増えた理由は分かっていない。

7. おわりに

本論文では、近年人気の小説CGMサイト「小説家になろう」を対象に、小説キーワードの多様性動向を調査した。Web APIを用いて、2015年12月時点での小説メタデータを全て収集した。収集したデータを、各小説毎のメタデータに整形した。

多様性動向を定量的に分析するために、 $\cos$ 類似度を用いる方法を提案した。投稿小説に付与されるキーワードについて、二つの小説のキーワードの $\cos$ 類似度を算出した。各月に投稿された小説群について、キーワード $\cos$ 類似度の総和である $SumCos$ 値を算出し、それを時系列でグラフ表示した。グラフから、各月の投稿小説における $SumCos$ 値の平均が徐々に増加していることが分かった。これにより、キーワード欄の単語に共通するものの増加、つまり

キーワードが類似する小説の増加を示せた。これは多様性の減少を定量的に示したと言える。

今後、ジャンルごとの多様性動向も調査したい。割合の高い（人気のある）ファンタジーや恋愛ジャンルと、割合の低い（人気がない）ジャンルで多様性動向が異なるかを調べたい。キーワード以外の小説タイトルやあらすじ文の単語についても分析を行いたい。多様性が減少しているのであれば、どこかの分野に偏りが出ていると考えられる。クラスタリング等の適用により、どの分野に偏っているのかを可視化したい。

謝辞 本研究は JSPS 科研費 15K00451 の助成を受けたものです。

参考文献

- [1] Wikipedia: Wikipedia - 小説投稿サイト, <https://ja.wikipedia.org/wiki/%E5%B0%8F%E8%AA%AC%E6%8A%95%E7%A8%BF%E3%82%B5%E3%82%A4%E3%83%88>.
- [2] NAVER まとめ: 「小説家になろう」のアニメ化作品まとめ, <https://matome.naver.jp/odai/2143054844927494601>.
- [3] 上畑恭平, 伊東栄典: タグの類似度に着目した利用者投稿サイト動画の多様性動向, Vol. 115, No. 381, pp. 83–88 (2015).
- [4] Kamihata, K. and Ito, E.: A quantitative contents diversity analysis on a consumer generated media site, *Proc. of AROB 21st 2016 (The Twenty-First International Symposium on Artificial Life and Robotics 2016)*, pp. 436–440 (2016).
- [5] Ito, E. and Shimizu, K.: Frequency and link analysis of online novels toward social contents ranking, *Proc. of SCA2012 (The 2nd International Conference on Social Computing and its Applications)*, IEEE, pp. 531–536 (2012).
- [6] Shimizu, K., Ito, E. and Hirokawa, S.: Predicting Future Ranking of Online Novels based on Collective Intelligence, *Proc. of ICDIPC2013 (The Third Int'l Conf on Digital Info. Processing and Communications)*, IEEE, pp. 263–274 (2013).
- [7] cakes, 川上量生: 川上量生の胸のうち, いま、好きなアニメをつくれるのはジブリくらい, <https://cakes.mu/posts/5036> (2014).
- [8] Takayuki, F.: コンテンツを最適化すると多様性は死ぬのか?, <http://fladdict.net/blog/2014/07/death-of-diversity.html> (2014).
- [9] ヒナプロジェクト社: 小説を読もう!, 累計ランキング BEST300, http://yomou.syosetu.com/rank/list/type/total_total/.
- [10] 三沢文也: かくいう私も青二才でね: なぜ、「小説家になろう」の作品群は魔王勇者モノばかりなのか?, <http://www.tm2501.com/entry/2015/02/15/202736>.
- [11] Wikipedia: 小説家になろう in Wikipedia, <https://ja.wikipedia.org/wiki/%E5%B0%8F%E8%AA%AC%E5%AE%B6%E3%81%AB%E3%81%AA%E3%82%8D%E3%81%86>.
- [12] なろうデベロッパー: なろう API, <http://dev.syosetu.com/man/api/>.
- [13] Shannon, C. E.: Prediction and Entropy of Printed English, *Bell System Technical Journal*, Vol. 30, No. 1, pp. 50–64 (1951).
- [14] データ分析・マイニングの世界 SAS: 類似度と

距離, <http://wikiwiki.jp/cattail/?%CE%E0%BB%F7%C5%D9%A4%C8%B5%F7%CE%A5>.

- [15] 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).