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

飯田，委哉
九州大学大学院システム情報科学府

伊東，栄典
九州大学情報基盤研究開発センター ： 准教授

佐嘉田，悠樹
九州大学工学部電気情報工学科

出版情報：火の国情報シンポジウム論文集. 2018, pp.1-7, 2018-03-01. 情報処理学会九州支部
バージョン：

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

飯田 委哉[1,a)]　伊東 栄典[2]　佐嘉田 悠樹[3]

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，Cos 類似度の総和を調べることで小説家になろうの画一化について定量的に評価をおこなった．次に，偏りのあるジャンルを特定するために，クラスタリングによるオンライン小説の多様性動向を調査する．各小説のあらすじを Doc2Vec を用いてベクトルで表現し，最後に，小説のベクトル群を X-means でクラスタリングする． クラスタ数やクラスタの要素数を分析することで，投稿小説の多様性動向の分析を行う．

キーワード：統計, ビッグデータ, クラスタリング,X-means,Doc2Vec, オンライン小説,CGM, 多様性分析

Analysis of diversity trend of online novel by clustering

Taro Joho[1,a)]　Eisuke Ito[2]　Yuki Sakata[3]

Abstract: In recent years, CGM (Consumer Generated Media) site, such as YouTube.com and Nicovideo.jp for movies, syosetu.com for novels, become very popular. Many contents are posted on those CGM sites everyday, and a huge number of users are browsing the contents. Now a day, some bloggers mentioned that similar contents are well posted to the CGM site. It is felt that contents and derivative contents which have already seen have increased. We proposed contents diversity metric, that is the sum of cosine similarities of contents. We applied our metric to contents of nicovideo.jp and syosetu.com, and found that contents diversity was decrease in both CGM. To identify bias of contents, we investigate the diversity trend of online novel using clustering. We extract synopsis part from each novels, and vectorize them by Doc2Vec. After that, we clustered the vectors using X-means clustering. We report our method and results of clustering.

Keywords: statistics, big data, clustering, X-means, Doc2Vec, online novels, CGM, diversity analysis

1. はじめに

利用者が動画や小説，画像などのコンテンツを投稿するサービス（CGM, Consumer Generated Media) が人気である．動画 CGM サイトである YouTube やニコニコ動画には毎日多数の動画が投稿されており，膨大な利用者が動画を閲覧している．近年，利用者の固定化や嗜好の画一化により，同ジャンルの動画の増加や，動画の多様性減少が指摘されている．多様性が減少し画一化が進むと文化的な活力も減り，サイト経営にも問題になる．

我々は CGM である「ニコニコ動画」と「小説家になろう」を対象に，コンテンツの多様性動向を分析してきた [1][2]．多様性動向を定量的に計測するための指標として，Cos 類似度の総和を提案した．ニコニコ動画における各月の投稿動画のメタデータについて Cos 類似度総和を算出した所，増加傾向化が見られた．これは類似コンテンツの増加傾向を示すため，コンテンツの画一化を定量的に示すと考えている．

本研究では Cos 類似度総和を「小説家になろう」にも適

1　九州大学大学院システム情報科学府
2　九州大学情報基盤研究開発センター
3　九州大学工学部電気情報工学科
a)　t.iida.630@s.kyushu-u.ac.jp

応し，コンテンツの画一化を定量的に評価した．コンテンツの画一化により，どのジャンルや分野の動画が増えているかを調査するため，本研究では小説のメタデータ群（文書群）にクラスタリングを適用する．小説群をクラスタリングするには，各小説をベクトルで表現する必要がある．各小説に付随するメタデータ，特にあらすじに含まれる単語を用いて，小説をベクトル化する．ベクトル化には，Doc2Vec で得た単語ベクトルを小説メタデータのあらすじに適用し，小説をベクトル化する．Doc2Vec で得た単語ベクトルを動画メタデータの単語に適用し，動画をベクトル化する．

各小説ベクトルに対して X-means によるクラスタリングを適用する．クラスタのサイズ等から小説集合の多様性について分析を行う．最後にクラスタリング結果への評価を Gini 係数を用いて行う．

本論文の構成を述べる．第 2 節では，小説家になろうおよびなろう API について述べる．第 3 節では，cos 類似度による多様性動向分析について述べる．第 4 節では，Doc2Vec を用いた小説のベクトル化および X-means による小説クラスタリングについて述べる．最後に第 5 節でまとめと今後の課題を述べる．

2. 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[3] 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 2018 年 1 月 31 日，登録者数が 1,185,453 人，掲載小説数は 542,291 作品である．

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	ダンジョンに出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 形式と JSON 形式で提供されている．小説家になろうが提供する「なろう API」[4] を用いて，JSON 型式ファイルの小説情報を収集するプログラムを python 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．Ncode は n で始まり数値 4 桁の後ろにアルファベットが辞書順に付与されている．例えば，n9999ab の次の小説は n0000ac となる．これを 2017 年 10 月 25 日から同年 11 月 15 日まで，当時最も新しく投稿された小説の Ncode を含む n0000a から n9999ej まで順にクローリングをおこなった．削除された小説のメタデータは得られないため，得られた小説数は 518967 件となった．また，小説を一つ以上投稿している利用者は 168430 人である．

この収集データを sqlite3 に格納し，適宜利用する形をとった．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〜2017 年 11 月
形式	json 形式
データ件数 (小説数)	518,967
時点	2017 年 11 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 月から現在までの間，概ね右肩上がりに投稿小説数が増えている．現在では 1 ヶ月あたり 10000 本を超える新作小説が投稿されている．

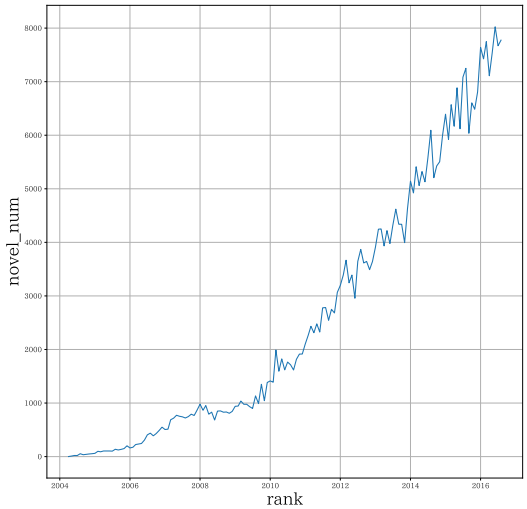

図 1 各月の新規小説投稿数

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，作者ごとの投稿小説数の散布図を示す．散布図は両対数軸で直線状になっているため，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	1,690	26407	尚文産商堂
2	835	540767	雪　つむじ
3	681	393690	zat
4	641	8909	西黄小路　　岳秋
5	633	21242	日下部良介
6	602	34969	百 (難しい童話)
7	569	14644	竹仲法順
8	564	13871	夏生
9	538	790405	袋小路　めいろ
10	517	26055	かみむら律子

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 種類あったが，2016 年 3 月にジャンルは再編されている．5 つの大ジャンルとそれに従う 20 のジャンルが設定されており，その他ノンジャンルを含めた全 21 ジャンルである．本稿の分析は変更後のジャンルに基づく．全体の小説に対

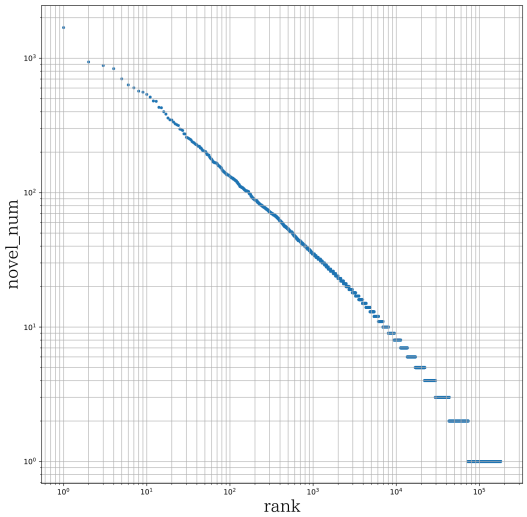

図 2 小説投稿数-作者数 (両軸対数尺度)

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．

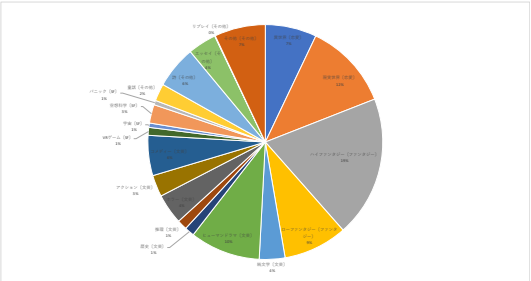

図 3 ジャンルの割合

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 6 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．一意なキーワード数，総数共に増加している．キーワード総数は 2016 年 3 月のランキング再編により，「異世界転移」もしくは「異世界転生」要素を含むものはキーワードにこれらを示すことが必須となったため急増したと思われる．

5. cos 類似度による多様性動向分析

5.1 考え方

我々は，２つの小説間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 4 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 4 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

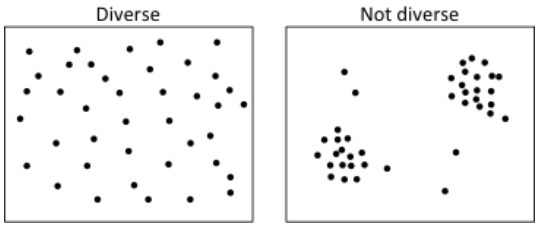

図 4 多様性についてのモデル

る場合，２点間の距離は近く，類似度は大きくなる．

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 [5]．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．

5.2 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 [6]．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 (1) で算出される．

$$cos(i,j) = \frac{\sum_t M(i,t) * M(j,t)}{\sqrt{\sum_t M(i,t)^2} * \sqrt{\sum_t M(j,t)^2}} \quad (1)$$

「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．

$$SumCos(D) = \sum_{i=1}^{n-1} \sum_{j=i}^{n} cos(i,j) \quad (2)$$

5.3 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．

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 5 に示す．

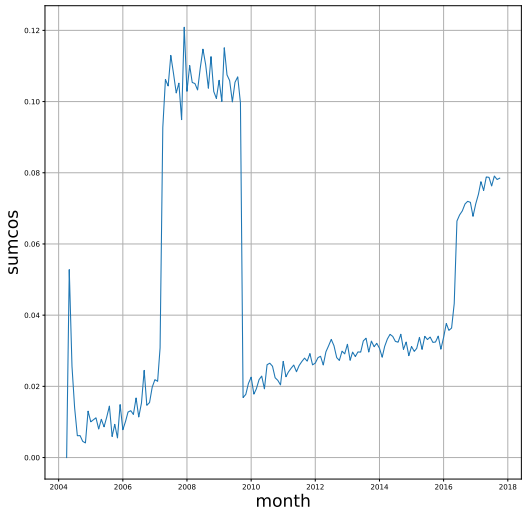

図 5 SumCos 値（平均値）

5.4 考察

図 5 に示した，平均 $SumCos$ 値の推移を見ると，2007 年頃 4 月頃に大きく上昇し，その後 2009 年 10 月頃に大きく減少している．その後は緩やかに上昇し，2016 年 6 月ごろから急増している．

平均 $SumCos$ 値 (図 5) では，2000 年代後半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．

5.5 2007 年 4 月〜2009 年 9 月についての考察

図 5 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 6 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．また，同時に $SumCos$ 値も変化している．

また 4 半期で登録小説数の多いキーワードを以下の表 5 に示す．現代 (モダン) というキーワードが 2006 年 4 月から最も多いキーワードとなり，2007 年 4 月からの四半期では前期にくらべて 3 倍となっている．そのまま，トップを維持し，SumCos が急激に下がる 2009 年 10 月でトップが恋愛と入れ替わっている．現代 (モダン) をキーワードに含む小説数の推移を図 7 に示す．2007 年 4 月 2009 年 10 月の間の増加が平均 Sumcos と同じ傾向となっていることからもこのキーワードを含む小説数の増加が要因であると推察される．

この期間に他に多いキーワードには高校生や恋愛が挙げられる．2006 年 ̃2008 年はケータイ小説が流行していた時期である．よって，それに触れていた人物が恋愛小説を小

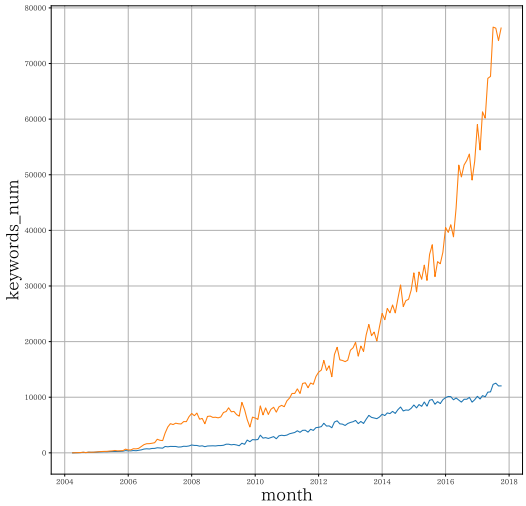

図 6 各月のキーワード数（青：一意な単語数，赤：単語の延べ出現）

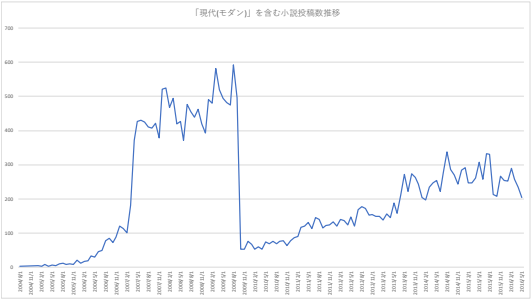

図 7 「現代 (モダン)」を含む小説数推移

説家になろうで投稿していたため，この時期のみ急激に SumCos が上昇していたと考えられる．

表 5 4 半期で登録小説数の多いキーワード

期間	キーワード	小説数
2005/10 12	恋愛	57
2006/1 3	恋愛	81
2006/4 6	現代 (モダン)	110
2006/7 9	現代 (モダン)	215
2006/10 12	現代 (モダン)	282
2007/1 3	現代 (モダン)	396
2007/4 6	現代 (モダン)	1229
2007/7 9	現代 (モダン)	1243
2007/10 12	現代 (モダン)	1321
2008/1 3	現代 (モダン)	1486
2008/4 6	現代 (モダン)	1217
2008/7 9	現代 (モダン)	1371
2008/10 12	現代 (モダン)	1275
2009/1 3	現代 (モダン)	1554
2009/4 6	現代 (モダン)	1496
2009/7 9	現代 (モダン)	1562
2009/10 12	恋愛	588

6. Doc2Vec を用いた小説のベクトル化

小説家になろうの投稿数の増加とともに，Sumcos が増加していることからサイト全体の多様性が減少していることがわかった．次にどのジャンル，テーマに偏っているかを調べるために小説のメタデータ群（文書群）をクラスタリングする．クラスタリングを適用するには，対象をベクトル化する必要がある．小説メタデータにはタイトル，説明文，キーワードが有る．うち，タイトルやキーワードは重要情報ではあるものの，文章や単語が少なく，かつ単語のゆらぎも有る．小説のあらすじを用いて各小説のベクトル化を行う．ベクトル化には Dec2Vec を利用した．

6.1 Word2Vec，Doc2Vec

Word2Vec は Tomas Mikolov らの開発した分散表現を生成する手法で，各単語を高次元のベクトルで表現する [7]．Word2Vec では，文章中に含まれる単語の出現数を利用する Countinuous Bag-of-Words モデルと，文章中に含まれる単語の並びから単語の出現確率を利用する Skip-gram モデルの両方の学習モデルを用いて，Hierarchical Softmax 及び Negative Sampling によって高速化を行っている．各単語を高次元ベクトルで表す手法「分散表現」では，単語のベクトルの加法・減法の結果が，単語の意味の加法・減法が成り立つ規則性が示されている．例えば vector('king') - vector('man') + vector('woman') が vector('queen') に近似する．

同様の手法を文章について使用したものに Doc2Vec?が存在する．Doc2Vec は文書の分散表現を生成できるため，文章をベクトル化できる．

6.2 小説家になろうあらすじベクトル化

Word2Vec や Doc2Vec を用いる場合，単語を適切なベクトルで表現するための学習データが必要である．図 8 にデータ処理の流れを示す．

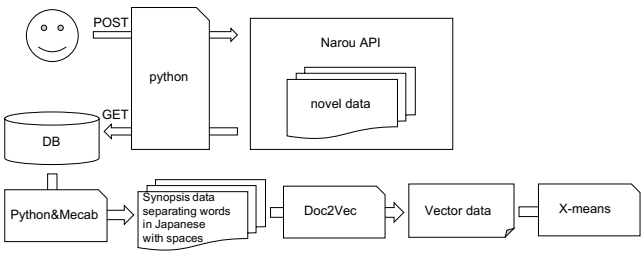

図 8 データ処理の流れ

まず，小説のメタデータから，小説のあらすじを月ごとに抽出した．小説のあらすじから改行を除いて一行の文章とし，Doc2Vec に適用する学習データ（コーパス）とした [9]．Python 用の自然言語処理及び機械学習モジュール群である gensim [10][11] に含まれる Doc2Vec を使い，学習用データから各あらすじの分散表現 (100 次元ベクトル) を生成した．

6.3 ベクトル化とクラスタ数決定指標の比較

Doc2Vec で得たあらすじメタデータ（文書）のベクトル群を，高速なクラスタリング手法である X-means 法でクラスタリングする．X-means のクラスタ数決定には AIC と BIC の２つを用いる．2x2=4 通りの組み合わせを比較し，最適な組み合わせを決める．図 9 にクラスタリングの概要を示す．

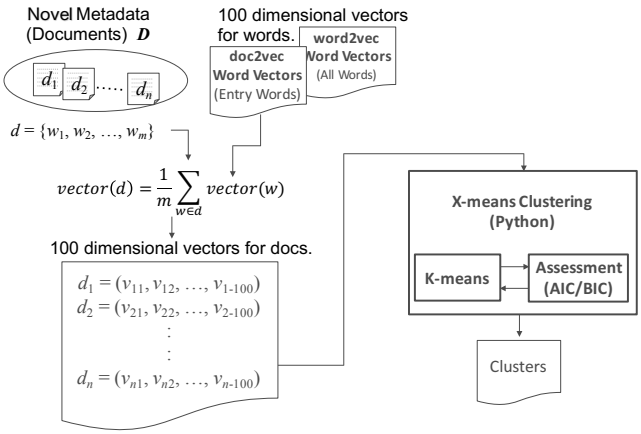

図 9 小説メタデータの X-means クラスタリング

6.3.1 X-means

K-means 法は，MacQueen，Lloyd，Forgy らが考案した非階層型クラスタリング手法である [12]．n 個のデータをデータ間の類似性 (距離) を尺度に，あらかじめ定めた K 個のクラスタに分類する．X-means 法は，Pelleg と Moore により考案された K-means 法を応用したクラスタリング手法で，K を自動推定するアルゴリズムである [13]．Pelleg と Moore の文献 [13] では，BIC を指標に用いてクラスタを評価し，指標が最も良い値となるクラスタ数 K を決定する．

石岡は，Pelleg らの手法に改良を加えた，BIC を用いたアルゴリズムを提案している [14], [15]．石岡の手法では，２分割の K-means クラスタリング手法を再帰的に適用する．クラスタに対し，分割まえと２分割後の BIC を比較し，２分割後の値が悪くなれば，分割せずに終了する．再帰的に行なうため，石岡の手法は停止が速い．ただし最適解でない場合もある．

6.3.2 AIC と BIC

AIC (Akaike's Information Criterion, 赤池情報量規準) は赤池弘次が考案した統計モデル評価規準である [16]．BIC (Bayesian information criterion, ベイズ情報量規準) は Schwarz が考案した統計モデル評価規準である [17]．AIC を式 (3) に，BIC を式 (4) に示す．

$$AIC = -2\ln L + 2k, \tag{3}$$

$$BIC = -2\ln L + k\ln n. \tag{4}$$

式 (3) および式 (4) で，L はモデルの尤度関数の最大値, k はモデルのパラメータ数, n はデータ数である．

AIC も BIC も値が最小となるモデルを選ぶことで良いモデル選択ができるとされている．AIC と BIC はいずれもよく用いられる手法であるものの罰則項が異なる．

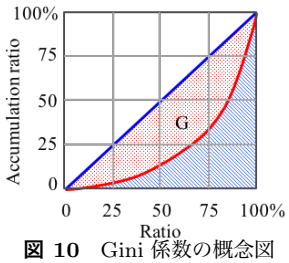

図 10 Gini 係数の概念図

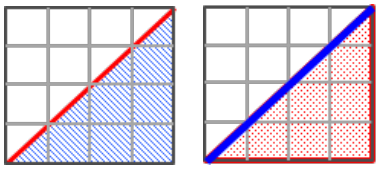

図 11 G=0(完全に均一) 及び G=1(完全に集中) の場合

6.4 本研究の X-means クラスタリング

本研究で用いた X-means におけるクラスタ数は，最初の局所解とした．K-means ($K \geq 2$) で分割したクラスタから AIC の値を算出する．次に (K+1)-means で分割したクラスタについて AIC の値を算出する．K の場合と K+1 の場合を比較し，K の場合の値が小さければ，その K が最適なクラスタ数として終了する．逆に K+1 の値が小さければ K+2 の場合と比較する．

これを月別投稿小説ごとに行い，クラスタ数の偏在性の傾向を評価する．またクラスタ内部についても評価するため，Gini 係数や Sumcos を適応する．

6.4.1 Gini 係数

Gini 係数は式 (5) で定義される値で，確率変数 F に対するローレンツ曲線 $L(F)$ と均等配分線によって囲まれる領域の面積と均等配分線より下の面積の比と定義される．均等配分線は分布が一様である場合のローレンツ曲線である．

$$Gini = \frac{1/2 - \int_0^1 L(F)dF}{1/2} = 1 - 2\int_0^1 L(F)dF \tag{5}$$

Gini 係数の概念を図 10 と 11 に示す．Gini 係数は 0 から 1 の値を取り，1 に近いほど不均等で，0 に近いほど均等であることを示す．Gini 係数は社会における所得分配の不平等さを測る指標として用いられることが多い．今回はクラスタの要素数の偏り調査に用いる．

7. おわりに

本研究では，オンライン小説投稿サイトである「小説家になろう」の多様性を定量的に評価することを目的として分析を行った．

まず小説間の類似性を示す，Cos 類似度の総和を求める

ことで多様性の減少に関して評価を行った．月別新規小説投稿数が増加に従い，平均 Sumcos 値も増加しているため，多様性の減少を定量的に確認することができた．

今後は投稿小説の偏りの特徴について定量的な評価を行う．投稿小説を月別に分割してクラスタリングを行い，クラスタ数の推移をみることで偏在性の確認を行う．また各クラスタについて Gini 係数や Sumcos などを用いることでも偏在性を調べたい．

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] Kyohei Kamihata and Eisuke Ito. A quantitative contents diversity analysis on a consumer generated media site. In *Proceedings of AROB 21st 2016 (The Twenty-First International Symposium on Artificial Life and Robotics 2016)*, pp. 436–440, 2016.

[2] Eisuke Ito and Yuya Honda. Keyword diversity trend of consumer generated novels. In *Proceedings of ICESS2017*, 2017.

[3] 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`.

[4] Narou-Developer. Narou api. `http://dev.syosetu.com/man/api/`.

[5] 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`.

[6] 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.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2 of *NIPS'13*, pp. 3111–3119, USA, 2013. Curran Associates Inc.

[8] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine*, pp. 1188–1196, 2014.

[9] 佐嘉田悠樹, 伊東栄典. Cgm 百科辞典を用いた利用者投稿動画クラスタリング. 平成 29 年度 電気・情報関係学会九州支部連合大会, pp. 544–545, 2017.

[10] gensim topic modeling for humans. `https://radimrehurek.com/gensim/`.

[11] Radim Řehůřek and Petr Sojka. Software framework for topic modeling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 05 2010.

[12] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297. University of California Press, 1967.

[13] Dau Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the 17th International Conf. on Machine Learning (ICML'00)*, pp. 727–734. Morgan Kaufmann, 2000.

[14] Tsunenori Ishioka. Extended k-means with an efficient estimation of the number of clusters. In *Lecture Notes in Computer Science*, Vol. 1983, pp. 17–22. Springer, May 2002.

[15] 石岡恒憲. X-means 法改良の一提案：k-means 法の逐次繰り返しとクラスターの再併合. 計算機統計学, Vol. 18, No. 1, pp. 3–13, 2006.

[16] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, pp. 267–281, 1973.

[17] Gideon Schwarz. Estimating the dimension of a model. In *The Annals of Statistics*, Vol. 6, pp. 461–464, 1978.