クラスタリングによる利用者投稿動画の偏在性分析

佐嘉田悠樹† 伊東 栄典††

† 九州大学工学部電気情報工学科 〒819–0395 福岡市西区元岡 744
†† 情報基盤研究開発センター 〒819–0395 福岡市西区元岡 744
E-mail: †y.sakata.222@s.kyushu-u.ac.jp, ††ito.eisuke.523@m.kyushu-u.ac.jp

あらまし 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。近年 CGM サイトへの投稿コンテンツの画一化への懸念がある。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。我々はニコニコ動画を対象に，動画の多様性動向について分析している。今回，動画のメタデータにクラスタリングを適用し，クラスタのサイズ等から動画集合の多様性について分析した。動画をクラスタリングするには，動画をベクトルで表現する必要がある。動画にはタイトル，説明文，視聴者が付与するタグが付随しているものの，これらの文字は重要情報ではあるものの，情報量が少なくまた単語のゆらぎも大きい。そこで，辞書であるニコニコ大百科等を利用して，動画ベクトル化する。ニコニコ大百科の説明ページから，Doc2Vec で見出し語を，Word2Vec で各単語をベクトル化した。単語のベクトルデータを用いて，各動画をベクトルで表現した。次に動画ベクトルの集合を分割 X-means でクラスラリングした。その後，各クラスタのサイズなどを分析することで，投稿動画の多様性を分析した，ベクトル化，クラスタリングの結果，および多様性動向の分析結果について報告する。

キーワード CGM, 多様性動向解析, ニコニコ動画, Word2Vec, クラスタリング, X-means

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 [1]。多様性が減少し画一化が進むと文化的な活力も減り，サイト運営にも問題になる。

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 [2] [3]。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。

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，本研究では動画のメタデータ群 (文書群）にクラスタリングを適用する。動画群をクラスタリングするには，各動画をベクトルで表現する必要がある。各動画に付随するメタデータに含まれる単語を用いて，動画をベクトル化する。

動画メタデータには，動画投稿者が付与するタイトルおよび説明文，投稿者と視聴者が付与するタグが有る。これらは重要情報ではあるものの，情報量が少なくまた単語のゆらぎも有る。そこで豊富な情報があり単語の解説を行っているインターネット百科事典の一つであるニコニコ大百科を援用する。ニコニコ大百科のページを取得し，記事内の文章に対しての Word2Vec 及び記事の見出し語に対しての Doc2Vec を用いて単語をベクトル化を行った。Word2Vec および 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。

各動画ベクトルに対して X-means によるクラスタリングを適用した。クラスタのサイズ等から動画集合の多様性について分析した。最後に Gini 係数を用いてクラスタリング結果への評価を行なった。ベクトル化，クラスタリングの結果，および多様性動向の分析結果について報告する。

本論文の構成を述べる。第 2 節では，ニコニコ動画およびニコニコデータセットについて述べる。第 3 節では，CGM 百科辞典であるニコニコ大百科を用いた動画のベクトル化について述べる。第 4 節では，X-means による動画クラスタリングについて説明する。第 5 節では，クラスタによる動画の多様性動向について分析を評価を行なう。最後に第 6 節でまとめと今後の課題を述べる。

2. ニコニコ動画とニコニコデータセット

2.1 ニコニコ動画

ニコニコ動画は，カドカワ社 (ドワンゴ社) が運営する動画共有サイトで，同社の会員登録制サービス niconico に含まれている。ニコニコ動画は「2 ちゃんねる (現:5 ちゃんねる)」や「ふたばちゃんねる」などのサイトを文化背景としたアニメ，ゲーム，ポップミュージックなどを中心とした所謂オタク文化の傾向を持つ。他の動画 CGM サイトとの違いとして，動画の再生画面に再生時間軸を指定したコメントを付与できるという特徴がある。このコメントは動画再生時に表示されるため，視聴者に擬似的な他者との閲覧体験共有の感覚を与える。他の特徴と

して，動画の視聴・投稿共に会員登録が必要であること，動画に対して視聴者がタグ付与できること，動画の人気ランキング機能などがある。表 1 に，ニコニコ動画の利用者数等の情報を示す。表 1 の値はカドカワ社の 2018 年 3 月期通期決算 [4] および同時期の決算短信 [5] に基づく。

表 1　ニコニコ動画 (2017 年 9 月時点)

項目	値
登録数	6,832 万人
プレミアム会員数	228 万人
男女比	男 67%, 女 33%
MAU (Monthly Active Users)	910 万人
サービス開始	2006 年 12 月
動画投稿サービス開始	2007 年 3 月

2.2　ニコニコデータセット

ニコニコデータセットは，ドワンゴ社と有限会社未来検索ブラジルから提供を受けて，国立情報学研究所が提供する「ニコニコ動画」の動画メタデータ等のデータセットである。国立情報学研究所が運営する情報学研究リポジトリ [6] で JSON 形式で提供されている。Ver.2 のデータセットは，約 1400 万件の動画メタデータと，その動画に投稿されたコメントデータを含む。動画メタデータにはタイトル，説明文，タグ，投稿日時，再生数，コメント数などのデータが含まれる。表 2 に，ニコニコデータセットのデータ件数等を示す。表 3 に，ニコニコデータセット中の 1 つの動画に対する動画メタデータに含まれる属性名について示す。

表 2　ニコニコデータセット

項目	値
動画数	14,269,919
単語総数	831,337,947
単語種類数	3,457,793
投稿期間	2007/3/6〜2016/8/31
容量	11.1 GB

表 3　動画メタデータに含まれる項目

No.	項目	説明
1	video_id	動画 ID
2	watch_num	再生回数
3	comment_num	コメント数
4	mylist_num	マイリスト数
5	title	動画タイトル
6	description	動画説明文
7	category	カテゴリタグ
8	tags	タグ
9	upload_time	投稿時間
10	length	再生時間

図 1 に，ニコニコ動画の動画再生数についての Rank-Frequency グラフを示す。表 4 に，動画メタデータの高頻出単語（上位 35 位）を示す。

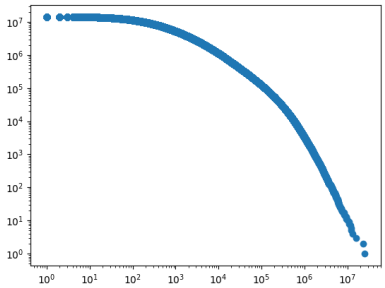

図 1　動画再生数の Rank-Frequency グラフ

表 4　ニコニコ動画のタイトル・タグ・説明文の単語登場数

順位	タイトル内	頻度	タグ	頻度	説明文内	頻度
1	実況	1,639,043	ゲーム	6,648,787	する	12,378,353
2	する	873,196	実況プレイ動画	2,006,908	動画	4,772,988
3	part	807,564	音楽	1,150,154	ない	4,472,229
4	実況プレイ	759,355	歌ってみた	818,679	いる	3,426,704
5	みる	708,075	エンターテイメント	625,757	次	3,004,698
6	歌ってみた	533,618	アニメ	579,621	前	2,938,824
7	Part1	335,547	投稿者コメント	507,306	なる	2,838,603
8	ない	304,341	リンク	462,900	マイリスト	2,691,947
9	動画	273,641	THE	430,299	ある	2,621,695
10	Part2	271,102	VOCALOID	428,822	様	2,302,883
11	The	230,065	実況プレイ	412,357	今回	2,005,581
12	目	229,490	東方	411,396	やる	1,215,014
13	版	217,158	IDOLM@STER	406,665	Part1	1,188,105
14	ゆっくり実況	208,889	動画	363,646	前回	1,142,278
15	Part3	193,759	ゆっくり実況プレイ	336,109	マイリス	1,141,503
16	プレイ	193,105	その他	318,843	見る	1,018,610
17	さん	180,406	もっと評価されるべき	294,784	こちら	990,481
18	Part	177,487	MikuMikuDance	286,919	投稿	972,056
19	歌う	155,405	プレイ動画	243,691	次回	935,430
20	から	149,625	ラジオ	228,952	もの	934,819
21	なる	147,785	実況	227,264	作る	890,896
22	Part4	146,368	演奏してみた	220,829	いく	844,532
23	目指す	144,866	スポーツ	212,600	できる	811,141
24	やる	144,859	初音ミク	206,660	他	805,463
25	MAD	144,363	BGM	206,632	まとめ	792,732
26	戦	141,929	MAD	191,017	コメント	784,221
27	プレイ動画	141,777	政治	187,712	Twitter	767,319
28	VS	136,732	Minecraft	174,840	しまう	746,796
29	ver	136,290	R-18	172,647	くる	731,160
30	MMD	134,518	シリーズ	166,666	歌う	719,862
31	ゆっくり	131,433	動物	160,392	blog	719,103
32	OP	129,306	ニコニコムービーメーカー	154,034	twitter	691,714
33	初音ミク	127,203	ボカロオリジナルを歌ってみた	153,228	いただく	647,443
34	THE	123,744	Part1	152,850	使用	635,840
35	Part5	117,244	PS3	150,351	コミュ	629,462

3.　ニコニコ大百科を用いた動画のベクトル化

本研究では動画のメタデータ群（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ためには，対象をベクトル化する必要がある。動画メタデータにはタイトル，説明文，タグが有る。これら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。そこでインターネット百科事典の一つであるニコニコ大百科を援用する。ここでは，ニコニコ大百科を用いて Word2Vec および Doc2Vec によるニコニコ大百科の単語または文書についてのベクトル化について述べる。ベクトル化により，略称や類義語などの単語間での表記の揺らぎを小さくすることができると期待される。Word2Vec，Doc2Vec によって得られたベクトルを用いて，動画のメタデータをベクトル化をする。

3.1　ニコニコ大百科

ニコニコ大百科は，大百科ニュース社（2017 年 7 月まで未来検索ブラジル）が運営するインターネット百科事典であり，会員登録制サービス niconico に含まれるサービスである。記事作

成・編集はニコニコ動画プレミアム会員に限定されている。同じ会員が所属しているためニコニコ動画と同じ文化に基づく内容の記事が多い。そのため，動画メタデータを親和性が高いと判断し，単語のベクトル化に用いることにした。ニコニコ大百科は他の CGM 百科事典同様にキーワード検索によって検索が可能である。また，検索語として特殊検索語「きみのすべてがみてみたい」を入れることによりニコニコ大百科内の全記事を表示できる。

3.2 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。Word2Vec では入力に文書を用いる。なお，空白区切りの西洋語と異なるため，日本語を扱う際には分かち書きを行う必要がある。

同様の手法を文章について使用したものに Doc2Vec [8] 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。Doc2Vec では入力に文書と各文書に関するタグを用いる。同様に日本語文書群に適用する際には文書内の各文を分かち書きする必要がある。

3.3 ニコニコ大百科を用いた単語ベクトル化

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。ニコニコ動画のメタデータに含まれる文章の分析には，ニコニコ動画に適した学習データが望ましいため，ニコニコ大百科を用いた。図 2 にデータ処理の流れを示す。

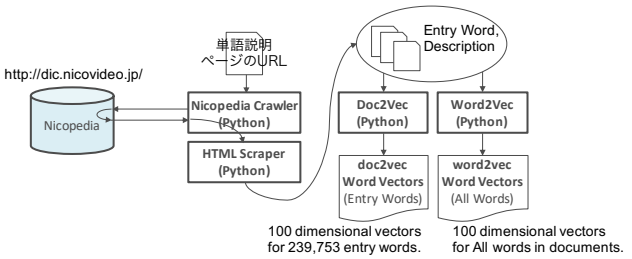

図 2　データ処理の流れ

「きみのすべてがみてみたい」を用いてニコニコ大百科の単語記事のページ 239,753 件を収集した。HTML で記述された記事のページから，広告やリンクや掲示板部分を除外して，単語を説明する文章部分のみをプログラムで抜き出した。見出し語およびその説明文を，Word2Vec 及び Doc2Vec に適用する学習データ（コーパス）とした [9]。見出し語はユーザー辞書に変換して日本語形態素解析エンジン MeCab [10] に適用する。

表 5　ニコニコ大百科使用データ

項目	値
収集ページ数	236,265
見出し語のページ数	176,018
単語総数	110,718,230
単語種類数	1,216,142
収集期間	2008/5/12〜2017/6/12
全ファイル容量	713 MB

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 [11] [12] に含まれる Word2Vec または Doc2Vec を使い，学習用データから単語の分散表現 (100 次元ベクトル) を生成した。Word2Vec では文書を入力として扱うため，見出し語と説明文をまとめて文書として用いた。Doc2Vec ではタグと文書を入力として扱うため，各記事の見出し語をタグとして，各説明文を文書として用いた。MeCab を用いて日本語辞書 mecab-ipadic-NEologd [13] [14] [15] およびユーザー辞書を用いて全単語を分かち書きを行って用いた。

3.4 動画メタデータのベクトル化

動画メタデータのベクトルは，式 (1) に示す，全単語のベクトル値の平均として算出した。

$$vector(d) = \frac{1}{m} \sum_{w \in d} vector(w). \qquad (1)$$

式 (1) で，w は文書 d に含まれる単語である。$vector(w)$ は，予め Word2Vec または Doc2Vec で算出した単語 w のベクトルである。また，m は文書 d の単語数である。これを各動画に対して求める。

4. ベクトル化とクラスタ数決定指標の組み合わせの最適化

式 (1) で得た動画メタデータ（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手法を比較し決定する。動画メタデータのベクトル化に用いる単語のベクトル化には Word2Vec または Doc2Vec を用いて生成した分散表現の 2 つを用いる。X-means のクラスタ数決定には AIC と BIC の 2 つを用いる。これら 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 3 にクラスタリングの概要を示す。

4.1 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 [16]。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 [17]。Pelleg と Moore の文献 [17] 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 [18], [19]。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

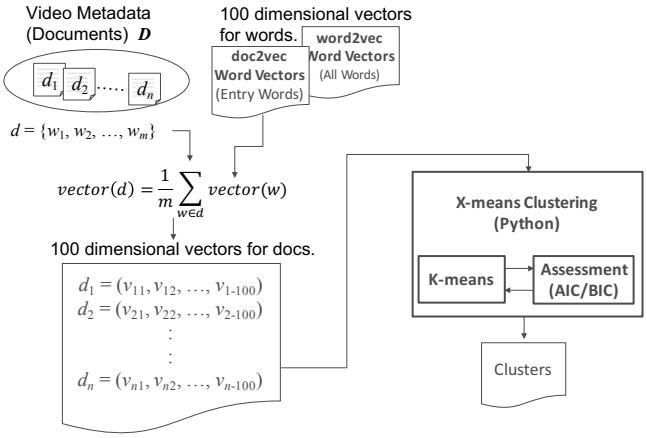

図 3　動画メタデータの X-means クラスタリング

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。

4.1.1　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 [20]。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 [21]。AIC を式 (2) に，BIC を式 (3) に示す。

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

$$BIC = -2\ln L + k\ln n. \tag{3}$$

式 (2) および式 (3) 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。

4.2　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 ($K \geqq 2$) で分割したクラスタから AIC 及び BIC 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 及び BIC を値を算出する。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。

4.3　比較ための動画メタデータ集合

ニコニコデータセットに含まれる全動画は 1400 万以上有るため，手法の比較には多すぎる。比較のためには均質な，同種類の動画を対象としたい。そこでボーカロイド楽曲の動画のみを対象とすることにした。ボーカロイド楽曲はニコニコ動画発祥の文化で，ボーカロイド楽曲動画は 2007 年から 2013 年に爆発的に増加し，2017 年現在も人気がある [22]。

Ryryo 氏が運営するボカロデータベース [23] から，ニコニコ動画に投稿されているボーカロイド楽曲の作者のうち，人気上位の 100 人を抽出した。また同サイトから，その作者が作成・投稿した全ボカーロイド楽曲動画を抽出した。抽出したボーカロイド楽曲動画は 3,702 個である。3,702 個のボーカロイド楽曲の動画メタデータをニコニコデータセットから抽出し，比較ための動画メタデータ集合とした。

4.4　比較結果

図 4 に，X-means のクラスタ数決定指標として BIC と AIC を，単語のベクトル化に Word2Vec を Doc2Vec を用いた場合の，クラスタ数を示す。

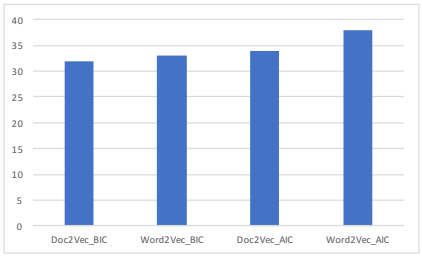

図 4　X-means クラスタ数 (Word2Vec/Doc2Vec, AIC/BIC)

図 4 では，AIC が BIC より，また Word2Vec が Doc2Vec よりクラスタ数が大きい。AIC は BIC と比較して罰則項がサンプル数 n の影響を受けないため第一項が大きくなり，そのため AIC の方が BIC より大きいクラスタ数になったと考えられる。Word2Vec は記事レベルの Doc2Vec と異なり，単語レベルで扱える情報が多くなり，クラスタが細分化されたと考えられる。

5.　クラスタリングによる偏在性分析

5.1　クラスタ数推移による偏在性分析

ニコニコ動画の多様性動向分析のため，クラスタ数推移による偏在性を調べる。前節の実験で，最もクラスタ数が大きくなったのは Word2Vec と AIC X-means の組み合わせであった。各月の投稿動画集合に対して Word2Vec のベクトル化と，AIC X-means のクラスタリングを適用した。

推移分析の周期は月ごとにした。ニコニコデータセットの動画メタデータのうち，投稿日が 2007 年 3 月 1 日〜2016 年 8 月 31 日までのデータを用いる。それらを投稿日を 1ヶ月毎 (1 日 0 時 0 分 0 秒以上から翌月の 1 日 0 時 0 分 0 秒未満まで) に区切って利用した。

図 5 に月ごとの動画投稿数（左軸）と，AIC X-means によるクラスタ数（右軸）を示す。図 5 を見ると，動画投稿数が増減と，クラスタリング数には相関が見られない。

5.2　Gini 係数での偏在性分析

ニコニコ動画の動画を要素とするクラスタについて，クラスタ数に変化は無くとも，一部のクラスタに大部分の動画が属することになれば，動画全体は偏在していることになる。そこで，Gini 係数を用いて各クラスタの要素数の偏りを分析する。

5.2.1　Gini 係数

Gini 係数は式 (4) 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。

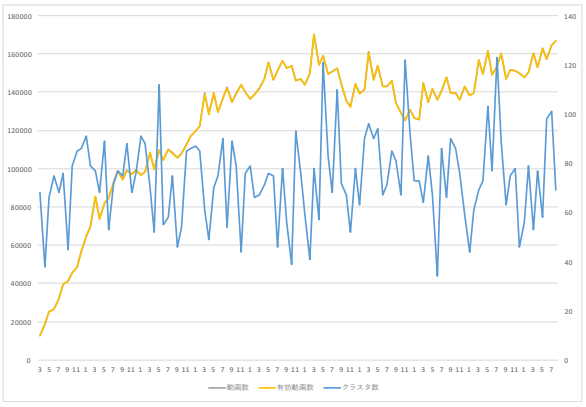

図 5　月ごとの動画投稿数と AIC X-means クラスタ数

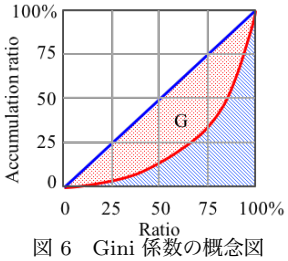

図 6　Gini 係数の概念図

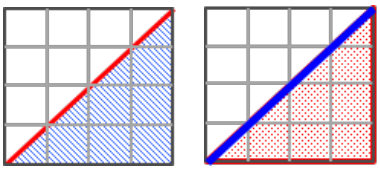

図 7　G=0(完全に均一) 及び G=1(完全に集中) の場合

$$Gini = \frac{1/2 - \int_0^1 L(F)dF}{1/2} = 1 - 2\int_0^1 L(F)dF \qquad (4)$$

Gini 係数の概念を図 6 と 7 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。

5.3　K 固定での K-means クラスタの Gini 係数推移

まず最初に K-means と Gini 係数の関係性を確認するため，クラスタ数 K を固定して K-means クラスタリングを適用した結果のクラスタについて，Gini 係数の推移を調べた。対象は図 5 と同じ動画メタデータ群である。2007/3/6〜2016/8/31 の間に投稿された動画メタデータ 14,269,919 件を，各月の投稿動画の部分集合（114 個）に分けたものである。

動画メタデータのベクトル化には，式 (1) を用いる。ニコニコ大百科の記事群から Word2Vec で得た単語ベクトルを用いる。K=64 と K=1024 に固定して K-means クラスタリングを適用した。出来たクラスタの要素数（動画数）から，Gini 係数を求めた。図 8 に K=64 の場合の Gini 係数推移を，図 9 に K=1024 の場合の Gini 係数推移を示す。

図 8 を見ると，64-means では時期による変化が見られない。図 9 の 1024-means では，2012 年まで Gini 係数が少し減少傾向にある。

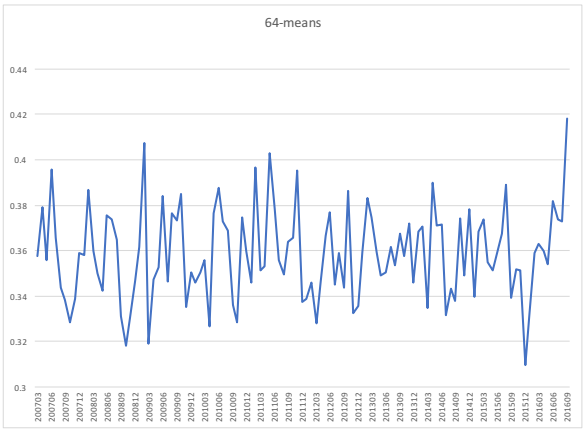

図 8　64-means Gini

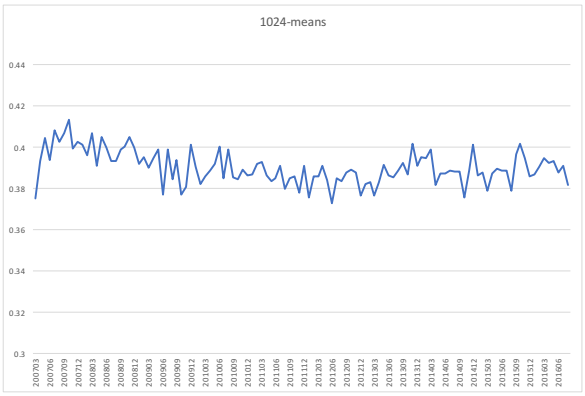

図 9　1024-means Gini

64-means では時期による変化が見られないのに対し，1024-means では特定の時期で見られたのは，1024-means と比較して 64-means は振動が大きいため観測しにくいということが考えられる。適当なクラスタ数を選択したクラスタリングでは振動が小さくなり，Gini 係数の変化を観測しやすい可能性がある。

5.4　X-means クラスタの Gini 係数推移

最後に，図 5 に示した X-means クラスタリング結果のクラスタで，クラスタの要素数についての Gini 係数を算出した。Gini 係数の推移を図 10 に示す。図 10 で，左軸（オレンジ線）は各月の動画投稿数で，右軸（青線）は Gini 係数の値である。

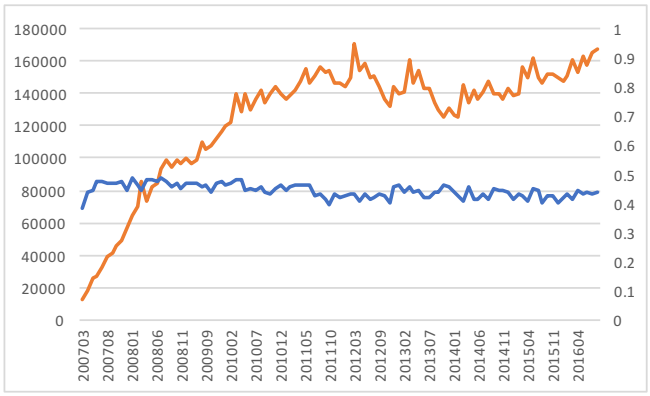

図 10　X-means クラスタの Gini 係数推移

図 10 を見ると Gini 係数に大きな変化が見られない。これは，クラスタ毎の要素数に偏りが少ない事を示している。これは，X-means のクラスタ数決定に用いた AIC では，クラスタ内のデータ数の比率を求めているためであろう。

6. おわりに

本研究ではニコニコ動画を対象に，クラスタリングによる利用者投稿動画の偏在性を分析した。サービス開始から 10 年経過したニコニコ動画では，投稿動画の多様性減少が懸念されている。多様性が減少しているのかを分析するため，動画メタデータ群へクラスタリングを適用し，クラスタの傾向を見ることで偏在性を分析した。

動画のメタデータには文章が少なく，またタイトル，タグ，説明文に含まれる単語にはゆらぎがある。ゆらぎを解消するため，ニコニコ大百科の記事を用いた。ニコニコ大百科の記事を Web から収集し，見出し語と説明文のみを抽出した。抽出した見出し語と説明文から Word2Vec 及び Doc2Vec で単語をベクトル化した。動画メタデータをベクトル化するため，Word2Vec 及び Doc2Vec で得た単語ベクトルを用いた。

動画メタデータのベクトルを用いて，動画を X-means クラスタリングした。単語ベクトル化の Word2Vec と Doc2Vec，X-means のクラスタ数数決定指標の AIC と BIC について比較した。ボーカロイド楽曲で比較した所，Word2Vec と AIC の組み合わせがクラスタ数が多くなった。

次に Word2Vec による単語ベクトル化と，AIC X-means クラスタリングによる，動画の偏在性分析を行った。全動画を投稿日で月ごとの部分集合に分割し，月ごとの動画集合を X-means でクラスタリングした。クラスタ数の推移を見た所，投稿動画数とクラスタ数の相関は無かった。クラスタのサイズ（動画数）の偏在性を分析するため，Gini 係数を用いた。K を固定した K-means（K=64 と 1024）と，X-means クラスタリング結果について，Gini 係数を算出し月ごとの推移を見た。その結果，明らか偏在性は確認出来なかった。今回のようなクラスタの大きさの割合から偏りを測るのは難しいと考えられる。

今後は以下を検討する。まず，クラスタリング手法を Ward 法などの階層化クラスタリングにした場合を比較したい。動画の偏りを見るには，単語の出現頻度も検討したい。今回の分析は，投稿動画のメタデータを分析しているため，動画作成者（投稿者）の視点からの調査になる。動画閲覧者側の視点での，偏在性調査も行いたい。どの利用者がどの動画を何回閲覧しているのかが分かれば，動画閲覧側からの偏在性分析も可能であろうと思われる。

謝　辞

本研究は JSPS 科研費 15K00451 の助成を受けたものです。

文　献

[1] 川上量生, “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.

[2] K. Kamihata and E. 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), 2016, pp. 436–440.

[3] E. Ito and Y. Honda, “Keyword diversity trend of consumer generated novels,” in Proceedings of ICESS2017, 2017.

[4] カドカワ株式会社, “2018 年 3 月期 通期決算,” http://pdf.irpocket.com/C9468/PoNw/gR6P/lj6p.pdf, p. 20, 03 2017.

[5] カドカワ株式会社, “平成 30 年 3 月期 第 2 四半期決算短信,” http://pdf.irpocket.com/C9468/Rt66/bMV7/JBbs.pdf, p. 3, March 2017.

[6] NII, “Infomatics data repository,” http://www.nii.ac.jp/dsc/idr/index.html.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proceedings of the 26th International Conference on Neural Information Processing Systems, ser. NIPS’13, vol. 2. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[8] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in Proceedings of the 31st International Conference on Machine, 2014, pp. 1188–1196.

[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,” in 平成 29 年度 電気・情報関係学会九州支部連合大会, 2017, pp. 544–545.

[10] 工藤拓 ・ 山本薫 ・ 松本裕治, “Conditional random fields を用いた日本語形態素解析,” 情報処理学会研究報告自然言語処理（NL）, vol. 2004, no. 47, may 2004, pp. 89–96. [Online]. Available: https://ci.nii.ac.jp/naid/110002911717/

[11] “gensim topic modeling for humans,” https://radimrehurek.com/gensim/.

[12] R. Řehůřek and P. Sojka, “Software framework for topic modeling with large corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 05 2010, pp. 45–50.

[13] 佐藤敏紀 ・ 橋本泰一 ・ 奥村学, “単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討,” in 言語処理学会第 23 回年次大会 (NLP2017). 言語処理学会, 2017, pp. NLP2017–B6–1.

[14] 佐藤敏紀 ・ 橋本泰一 ・ 奥村学, “単語分かち書き用辞書生成システム neologd の運用 — 文書分類を例にして —,” in 自然言語処理研究会研究報告. 情報処理学会, 2016, pp. NL–229–15.

[15] S. Toshinori, “Neologism dictionary based on the language resources on the web for mecab,” 2015. [Online]. Available: https://github.com/neologd/mecab-ipadic-neologd

[16] 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. University of California Press, 1967, pp. 281–297.

[17] D. Pelleg and A. 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). Morgan Kaufmann, 2000, pp. 727–734.

[18] T. Ishioka, “Extended k-means with an efficient estimation of the number of clusters,” in Lecture Notes in Computer Science, vol. 1983. Springer, May 2002, pp. 17–22.

[19] 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, 2006, pp. 3–13.

[20] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” 2nd International Symposium on Information Theory, 1973, pp. 267–281.

[21] G. Schwarz, “Estimating the dimension of a model,” in The Annals of Statistics, vol. 6, no. 2, 1978, pp. 461–464.

[22] 柴那典, 初音ミクはなぜ世界を変えたのか？. 太田出版, April 2014.

[23] Ryryo, “ボカロデータベース,” http://nicodb.jp/v/.