回帰分析によるCGM動画再生回数推定

柴田　知親†　　伊東　栄典††

† 九州大学工学部電気情報工学科　〒 819–0395 福岡県福岡市西区元岡 744
†† 九州大学情報基盤研究開発センター　〒 819–0395 福岡県福岡市西区元岡 744
E-mail: †t.shibata.123@s.kyushu-u.ac.jp, ††ito.eisuke.523@m.kyushu-u.ac.jp

あらまし　近年，ネットワークおよびモバイル端末の普及に伴い，Youtube やニコニコ動画といった，利用者投稿型動画共有サービスが人気である．動画以外にも，小説家になろう，pixiv，comico などに代表される，小説やイラスト，漫画などの投稿サイトも人気である．これらのサイトは，CGM（Consumer Generated Media）とも呼ばれる．我々は，CGM の一つであるニコニコ動画を対象に，動画のランキングや推薦を研究してきた．他の CGM と同様に，ニコニコ動画では極少数の動画は大量に再生されるのに対し，ほとんどの動画は再生回数が少ないというロングテールの分布をする．利用者が何らかの選択基準により動画選択した結果が再生回数に現れていると考えられる．そこで，CGM における利用者のコンテンツ選択モデルを確立したいと考えている．本研究では，コンテンツ選択モデル確立のために，動画の再生回数を回帰分析で推定する．最初に線形回帰分析で再生回数の推定した．その結果，多様な動画が含まれていると推定精度が低くなった．次に Deep Learning を適用し，Multi Layer Perceptron による再生回数の非線型回帰分析を行った．対象の動画集合の多様性を減らすため，クラスタリングで動画集合を分割した．分割した動画集合について，Deep Learning で再生回数を推定した．これらの手法および結果について報告する．

キーワード　CGM，ニコニコ動画，回帰分析，Deep Learning，MLP

1. はじめに

近年，スマートフォンの普及などにより，Youtube やニコニコ動画などの利用者投稿型動画共有サービスや，小説家になろう，pixiv，comico などに代表される，小説やイラスト，漫画などの投稿サイトが人気である．これらのサイトは CGM（Consumer Generated Media）とも呼ばれる．利用者によって投稿されたコンテンツがアニメ化されたり，楽曲がカラオケで配信されたり，その投稿者がメジャーデビューを果たすなど，CGM サイトは社会に大きな影響を与えるメディアへと成長している．我々はこれまでに，CGM の一つであるニコニコ動画を対象として，動画のランキングや推薦，多様性動向などを研究してきた [1] [2].

次に，CGM における利用者のコンテンツ選択モデルの確立を検討する．本研究では，その第一歩として，ニコニコ動画における動画再生回数を推定する回帰分析を行う．他の CGM と同様に，ニコニコ動画では極少数の動画は大量に再生されるのに対し，ほとんどの動画は再生回数が少ないというロングテールの分布をする．利用者が何らかの選択基準により動画を選択した結果が再生回数に現れていると考えられる．動画再生回数の推定を行うことで，どのような動画が人気を得るかがわかる．動画を評価し，投稿者にフィードバックを行うなどのサービスへ応用することができる．他の CGM コンテンツについても同様の研究が行われており，(株) ドワンゴによるニコニコ静画の閲覧数・お気に入り数予測でも，公開前にイラストにスコアを付けるなど，投稿支援サービスへの利用が考えられている [3].

まず，ニコニコデータセットを用いてカテゴリ毎に線形回帰分析で再生回数を推定した．次に，推定精度の高いカテゴリについてより詳細に分析する [4]．その結果，分析する動画集合の多様性が高いと推定精度が低くなった．次に，クラスタリングにより動画集合を分割し，多様性を減らした．動画集合について MLP（Multi Layer Perceptron）による推定を行い，線形回帰分析の結果と比較する．これらの手法および結果について報告する．

本文の構成を述べる．2. 節では，ニコニコ動画および本研究で使用したニコニコデータセットについて説明する．3. 節では，線形回帰分析についての概要および結果を述べる．4. 節では，3. 節の結果から，より詳細に線形回帰分析をした結果とその考察について述べる．5. 節では，動画集合の分割に用いた階層的クラスタリングの詳細および結果について述べる．6. 節では，MLP の構築，および 5. 節で定義した各クラスタへ適用した際の結果と考察について述べる．

2. ニコニコデータセット

2.1 ニコニコ動画

ニコニコ動画は株式会社ドワンゴが運営する動画共有サイトである．ユーザー登録することにより，動画投稿及び動画視聴を行うことができる．2006 年 12 月より開設されており，持株会社であるカドカワ株式会社の 2017 年 3 月期通期決算説明資料 [5] によれば，2017 年 3 月において，会員数が 6430 万人，うち有料会員数が 243 万人であり，日本の代表的な動画共有サイトの一つとなっている．

利用者は自分の制作した動画を投稿する一方で，その動画に対し動画を再生する，コメントをつける，マイリストと呼ばれ

るサイト内ブックマークに登録する，動画に関連するタグをつけるなどの行為を視聴者側の立場から行うことができる．なかでもコメント機能は，動画の再生時間軸上に対し，動画にオーバーレイする形で投稿できるというニコニコ動画独自のものである．また，タグについては，ニコニコ大百科という CGM 百科事典を利用して，利用者がタグについて説明する記事を自由に作成することができる．これらの独特なシステムにより，ニコニコ動画では，他の動画共有サイトには無い，独自の文化を築いて来た．

2.2 ニコニコ大百科

ニコニコ大百科は，利用者が自由に記述できる Wikipedia と類似した CGM 百科事典である．運営会社は niconico 全体の運営会社ドワンゴではなく，(株) 大百科ニュース社である．大百科ニュース社はドワンゴと同様，カドカワ (株) の子会社である．かつては (有) 未来検索ブラジルが運営していたが，2017 年 7 月 3 日にニコニコ大百科事業が大百科ニュース社に承継された．2008 年 5 月より開設されており，2018 年 1 月時点で 55 万件以上の記事が作成されている．

記事の編集はニコニコ動画のアカウントを持っている人に限られる．また，ニコニコ動画をはじめ niconico の各サービスとの連携を前提としているため，ニコニコ動画で使われるタグや動画投稿者，ニコニコ生放送でユーザー生放送を行っている人物などについての記事が多数存在している．動画におけるタグのように，記事にはそれに関連する動画へのリンクが記載されていることが多く，ニコニコ動画とニコニコ大百科には密接な関係があると言える．

本研究では，多数存在するニコニコ大百科の記事から，動画タグ（単語）の説明ページ 239,753 件を収集し，単語を説明する文章部分のみを抽出して [6]，動画のクラスタリングに使用した．表 1 に，用いたデータの数を示す．

表 1　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

2.3 ニコニコデータセット

ニコニコデータセットは国立情報学研究所が (株) ドワンゴおよび (有) 未来検索ブラジルから提供を受けて研究者に提供しているデータセットである．このデータセットには，ニコニコ動画コメント等データと，ニコニコ大百科データが有る．本研究では，前者に含まれる動画メタデータを利用している．動画メタデータは，ニコニコ動画サービス開始当初から 2016 年 8 月末までに投稿された動画について，2016 年 12 月に集計を行ったデータである．データ件数（動画件数）は 1,400 万件以上である．また，各動画のメタデータは表 3 に示す属性を含む．

表 2　ニコニコ動画コメント等データ概要

		Number
	集計期間	2007/3/6 – 2016/8/31
コメントデータ	総コメント数	3,465,469,245
	zip 圧縮サイズ	64GB
	展開後サイズ	200GB
動画メタデータ	総動画数	14,269,919
	zip 圧縮サイズ	3.6GB
	展開後サイズ	5.8GB

表 3　動画メタデータの内容

属性名	内容
video_id	動画 ID
title	動画タイトル
description	動画説明文
watch_num	再生数
comment_num	コメント数
mylist_num	マイリスト登録数
category	動画カテゴリ
tags	タグの列
upload_time	投稿時間 (UNIX 時間・秒)
file_type	動画フォーマット
length	動画再生長 (秒数)
size_high	高画質動画のファイルサイズ (byte)
size_low	低画質動画のファイルサイズ (byte)

3. 線形回帰分析

本研究では，動画再生回数の推定に複数の手法を適用した．まず最初に，線形重回帰分析を用いた．重回帰分析は，2 つ以上の説明変数から 1 つの目的変数を推定するための回帰分析法である [7]．

3.1 手　　法

説明変数として動画メタデータから，コメント数，マイリスト登録数，投稿時間，動画再生長の 4 つのデータを用いた．これらのデータは，利用者が動画を検索してその結果を表示する際，ソートの基準として用いられている．そのため，動画の選択に影響すると考えた．また，動画を連続閲覧する際，ある動画から他の動画へのリンクを辿る場合が多い．そのため，リンクも動画の再生回数に影響すると考えた．そこで，動画説明文中に動画 ID で表記される文字列をリンクとして抽出した．抽出結果から，他動画へのリンク数 out_degree と，他の動画からのリンク数 in_degree を算出した．リンク数の一例を図 1 に示す．

表 4 に示す 6 つを線形重回帰分析の説明変数とした．また，図 2 に，線形回帰分析のモデルを示す．

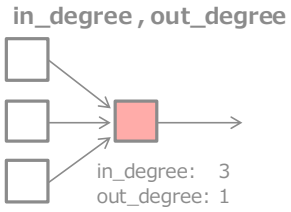

図 1　リンク数の例

データセット内の全動画と，各カテゴリ動画の部分集合それ

表 4　説 明 変 数

コメント数，マイリスト登録数，投稿時間，動画再生長，in_degree，out_degree

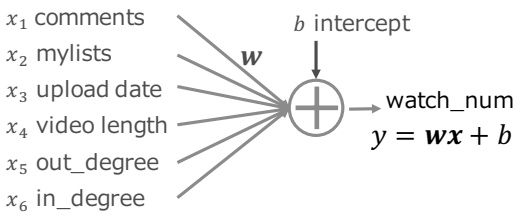

図 2　線形回帰分析のモデル

ぞれに図 2 のモデルを適用し [8]，推定精度を比較した．推定精度の指標には，線形回帰分析の決定係数 R^2 を用いた．決定係数は，式 (1) で表される．式 (1) で，y_i は各実測値，f_i は回帰分析によるそれらの推定値，μ は実測値の平均である．

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \mu)^2} \tag{1}$$

決定係数の値は 0 から 1 の範囲で表される．常に平均値を出力する予測モデルに比べて，相対的にどのくらい性能が良いのかを表し，1 に近いほど良い性能であることを示す [9]．

3.2　結　　果

表 5 に，全動画と各カテゴリ動画それぞれの決定係数を示す．「VOCALOID，ニコニコインディーズ，歌ってみた」などの音楽関連カテゴリは，全体に比べて決定係数が高い．一方「自然，ゲーム，政治」などのカテゴリでは決定係数が 0.6 未満となり，推定し難い．

4.　VOCALOID 動画の分析

柴那典氏が著書「初音ミクはなぜ世界を変えたのか？」[10] で述べているように，ボーカロイド楽曲はニコニコ動画発祥の文化である．3.2 節の表 5 で示したように，VOCALOID カテゴリ動画は決定係数が最高であるため分析対象としても興味深い．この分野の動画を詳細に分析する．

4.1　VOCALOID 動画の分割

投稿時間を用いて VOCALOID カテゴリ動画の日毎の動画投稿数を図 3 にグラフで示す．図 3 では，2009 年 10 月末から投稿数が爆発的に増大している．これは 2009 年 10 月 29 日のニコニコ動画アップデート時に「VOCALOID」が公式カテゴリとして追加されたためであろう．

実際に図 4 では，2009 年 10 月 29 日以前にボーカロイド楽曲の投稿は増加していた．しかし，2009 年 10 月 29 日より前の投稿動画は，投稿者が後で変更しない限り category 属性には VOCALOID 以外の値が含まれている．一方 tags には投稿者や視聴者が自由に単語を登録できる．ボーカロイド楽曲動画では，tags 属性に VOCALOID を含む場合が多い．

そこで，category が VOCALOID の動画を部分集合 1，tags に VOCALOID を含む動画を部分集合 2 とする．両者の積集合の割合は大きい．次に両部分集合を，ニコニコ動画のアップデート時期と動画投稿数の増減具合から，表 6 の時期に分割し，それぞれの傾向を調べる．以降，各部分集合を 1，1A，...，1D，2，2A,...，2D と表記する．10 個の部分集合それぞれに線形重回帰分析を適用した．

表 5　カテゴリごとの線形回帰分析の決定係数

カテゴリ	決定係数
全体	0.601
VOCALOID	0.952
ニコニコインディーズ	0.930
歌ってみた	0.927
音楽	0.914
踊ってみた	0.860
演奏してみた	0.846
東方	0.831
旅行	0.821
描いてみた	0.809
料理	0.768
動物	0.764
実況プレイ動画	0.756
作ってみた	0.729
ニコニコ技術部	0.708
アイドルマスター	0.708
ラジオ	0.704
ニコニコ手芸部	0.696
エンターテイメント	0.695
アニメ	0.695
スポーツ	0.682
車載動画	0.664
ニコニコ動画講座	0.649
例のアレ	0.644
その他	0.628
歴史	0.615
日記	0.605
R-18	0.605
科学	0.570
政治	0.569
ゲーム	0.565
自然	0.484

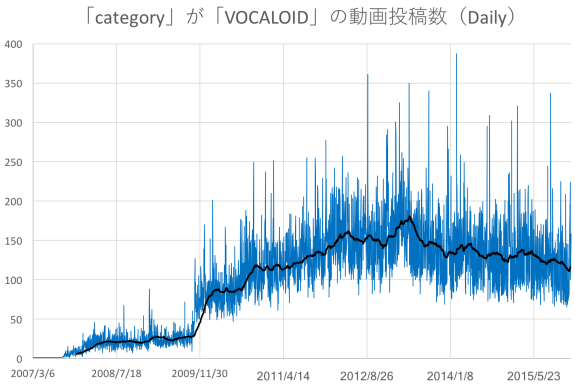

図 3　category が VOCALOID の動画投稿数

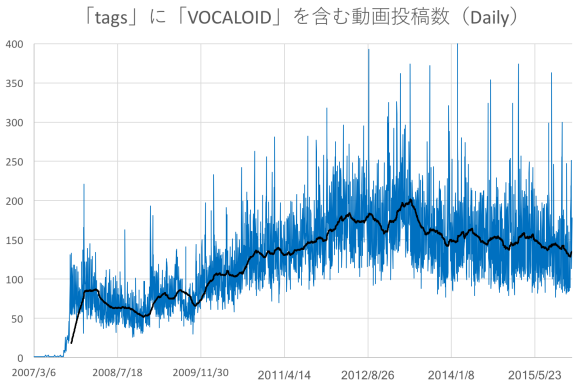

図 4　tags に VOCALOID を含む動画投稿数

表 6　VOCALOID 動画投稿数の増減に伴う期間の分割

分類名	期間	傾向
無印	2007/10 – 2016/08	(全体)
A	2007/10 – 2009/10	(公式カテゴリなし)
B	2009/11 – 2012/04	(投稿数増加)
C	2012/05 – 2013/10	(投稿数停滞)
D	2013/11 – 2016/08	(投稿数減少)

4.2　結　　果

今回は各決定係数に大きな差が見られないため，各動画の再生回数の実測値を y_i，推定値を f_i として，誤差 Δ_i，誤差率 ε_i を以下の式 (2)，(3) ように定義し，誤差率の平均値を各部分集合における推定精度の指標とした．

$$\Delta_i = |y_i - f_i| \tag{2}$$

$$\varepsilon_i = \frac{|y_i - f_i|}{y_i} \tag{3}$$

表 7 に結果を示す．また，図 5 は，各説明変数が目的変数へ与える影響の度合いを，標準化偏回帰係数の割合として算出したものである．さらに，推定精度が最高である 1D と，最低である 2C，同期間で平均誤差率の差が大きい 1C について，それぞれ図 6 から図 8 に，縦軸を実測値，横軸を推定値として両対数グラフで描画した．

表 7　各分類における推定値の平均誤差率

部分集合	動画数	平均誤差	平均誤差率
1	336371	2130.92	0.9324
1A	16571	6983.64	1.9883
1B	105669	2940.50	1.6998
1C	84127	2096.11	1.4915
1D	130004	1120.52	0.7826
2	416164	2183.24	1.2712
2A	53498	4491.68	1.6725
2B	121630	2758.61	1.6385
2C	94296	2063.78	2.4758
2D	146740	1140.42	1.0980

4.3　考　　察

表 7 の平均誤差率を見ると，部分集合 1D が最も推定精度が高く，部分集合 2C が最も推定精度が低い．また，1C と 2C では積集合の割合が大きいにも関わらず，平均誤差率の差が大きい．VOCALOID カテゴリでない動画の誤差率が大きいと考えられる．さらに，図 6 から図 8 を見ると，実際の再生回数が小さいほど推定値の誤差率が大きな動画が多い．線形回帰分析ではパラメータの推定に最小二乗法を用いるため，実測値と推定値を両対数グラフにプロットするとこのような広がり方をする．しかし，低再生回数の部分に着目すると，全体の広がり以上に推定値が大きく外れている動画が存在する．平均誤差率の差が大きい図 8 と図 7 を比べると明らかである．これらのことから，動画集合の多様性が高いほど推定し難いと考えられる．

図 5 では，マイリスト数が再生数推定に最も影響していることが分かる．VOCALOID 動画に限らず，視聴者はお気に入り動画をマイリストへ登録している．ボーカロイド楽曲動画は複数回再生を前提とする視聴者が多く，マイリスト登録率が他カテゴリより高くなるため，影響度が高いと思われる．表 8 に，再生回数に対するマイリスト登録数の割合をカテゴリ別に示す．表 8 を見ると，実際に，VOCALOID を含め，音楽関連カテゴリの動画はマイリスト登録率が高いことがわかる．表 5 でそれらの決定係数の値が高いことからも，マイリストから複数回再生された結果，再生回数とマイリスト登録数の相関性がより高くなったと考えられる．

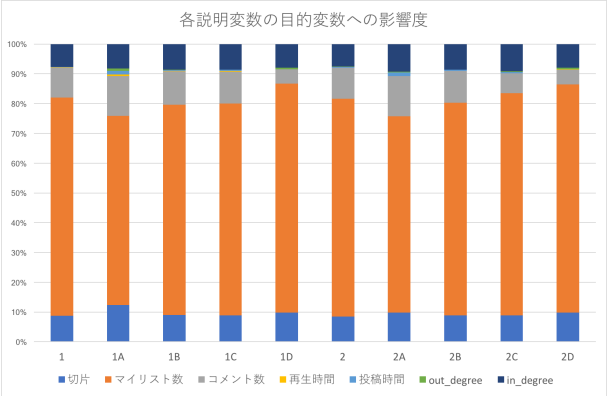

図 5　各説明変数の目的変数への影響度

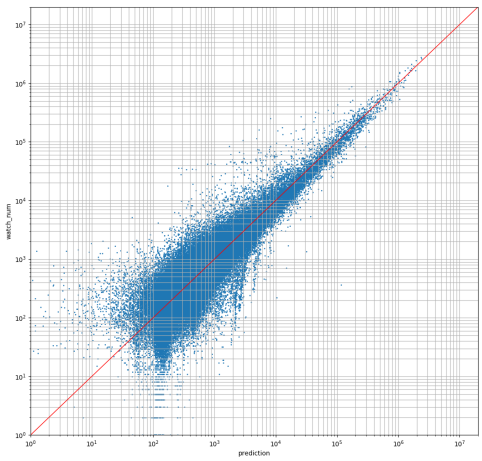

図 6　実測値と推定値の関係 (1D)

表 8　カテゴリごとの平均マイリスト登録率

カテゴリ	平均マイリスト登録率
全体	1.33%
VOCALOID	2.35%
ニコニコインディーズ	2.18%
音楽	1.94%
描いてみた	1.91%
歌ってみた	1.83%
その他	1.65%
ニコニコ動画講座	1.64%
踊ってみた	1.47%
アイドルマスター	1.41%
実況プレイ動画	1.29%
ニコニコ手芸部	1.23%
演奏してみた	1.20%
ラジオ	1.17%
アニメ	1.11%
旅行	1.10%
東方	1.10%
スポーツ	1.09%
エンターテイメント	1.06%
ゲーム	1.03%
歴史	0.99%
作ってみた	0.97%
日記	0.94%
自然	0.93%
動物	0.87%
料理	0.85%
ニコニコ技術部	0.82%
例のアレ	0.81%
科学	0.74%
車載動画	0.72%
R-18	0.71%
政治	0.67%

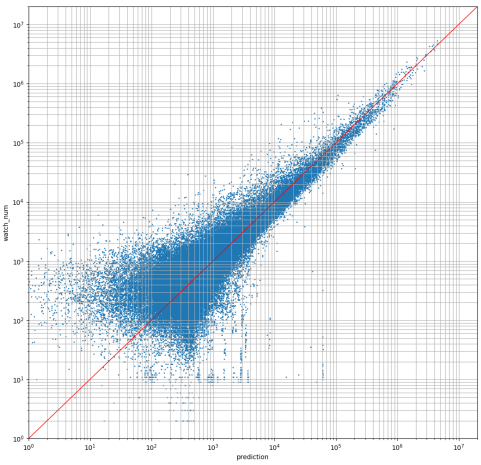

図 7　実測値と推定値の関係 (2C)

5.　階層的クラスタリング

4. 節の考察から，動画の多様性を考慮すると，特定の種類の動画に特化した部分集合を定義し，それぞれで回帰分析を行うことで，より高い推定精度を得られると考えた．本節では，特定の種類の動画から成る部分集合を求めるためのクラスタリングについて述べる．

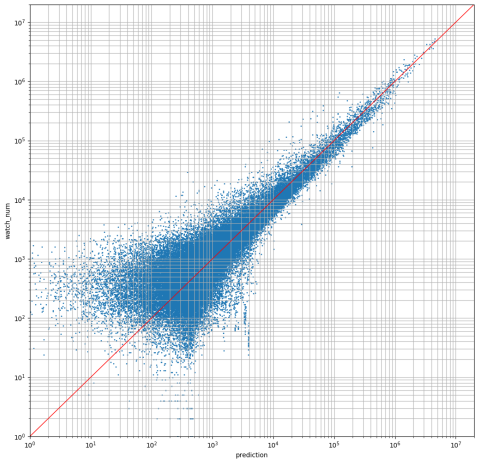

図 8　実測値と推定値の関係 (1C)

5.1　ニコニコ大百科を用いた動画のベクトル化

クラスタリングを適用するには，クラスタリング対象を何らかの数値ベクトルで表現する必要がある．ここでは，動画を説明するテキスト（メタデータ）を用いて動画をベクトル化する．

動画の特徴を示すものとして，視聴者がその動画に自由に登録できるタグがある．ニコニコ動画のタグは最大で 10 個である．また，ゆらぎも多い．例えば，MikuMikuDance を MMD と書くものがある．ゆらぎを吸収するため，Word2Vec を用いる．また，タグ自体を説明するものとしてニコニコ大百科がある．ニコニコ大百科を学習データとして，Word2Vec によるタグのベクトル化をした．

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] に含まれる Word2Vec 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 [6]．Word2Vec には，表 1 の全記事中に含まれる全単語を用いて分散表現を求めた．図 9 にデータ処理の流れを示す．

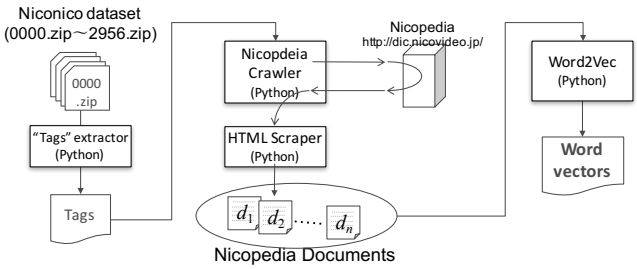

図 9　データ処理の流れ

動画のベクトルは，式 4 に示す，各タグのベクトル値の平均として算出した．

$$vector(v) = \frac{1}{m}\sum_{t \in v} vector(t). \tag{4}$$

式 4 で，t は動画 v に含まれるタグである．$vector(t)$ は，予め Word2Vec で算出したタグ t のベクトルである．また，m は動画 v が持つタグの数である．

5.2 Ward 法によるクラスタリング

ニコニコ動画のタグは最大で 10 個登録することができ，動画全体を各カテゴリに分類するのと同様に，より詳細な情報を動画に付与出来る．カテゴリを含め，特定の種類の動画により特化したタグへと，階層的に動画をクラスタリング出来ると考えた．

各動画の 100 次元ベクトルを利用して，部分集合 1A，...，1D，2A，...，2D それぞれについて Ward 法による階層的クラスタリングを行った [12]．Ward 法は階層法の中でもっとも精度が高い方法である [13]．クラスタ数が 2 つの時の距離の 15%を閾値として，クラスタ間の距離が閾値を超えるまでクラスタを結合した．

5.3 結　　果

表 9 に，部分集合 1A，...，1D，2A，...，2D それぞれのクラスタリング結果を示す．

表 9　各部分集合におけるクラスタ数

部分集合	クラスタ数	部分集合	クラスタ数
1A	56	2A	67
1B	43	2B	51
1C	38	2C	47
1D	44	2D	45

次に，各クラスタを表す語を定義する．クラスタ c_j 内の動画の各タグを t_i，c_j におけるタグ t_i を持つ動画の数を $n_{i,j}$ とし，クラスタの総数を $|C|$，$|\{c : c \ni t_i\}|$ をタグ t_i を含む動画が存在するクラスタの数とする．以下の式 (5) から (7) で表される $tfidf_{i,j}$ の値を各クラスタの各タグごとに算出し，各クラスタで $tfidf_{i,j}$ 上位 5 つのタグをそのクラスタを表す語とした．

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{5}$$

$$idf_i = log\frac{|C|}{|\{c : c \ni t_i\}|}, \tag{6}$$

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i. \tag{7}$$

線形回帰分析で誤差率が最も大きい部分集合 2C について，表 10 に，各クラスタの特徴と，クラスタ毎の平均誤差率を示す．クラスタ 8，15，35，36 を見ると，VOCALOID とは無関係なタグが上位に来ている．また，クラスタ 18 は他のボーカロイド楽曲動画をランキング形式でまとめた動画が多く，24 や 34 はユーザー生放送の録画を動画として投稿したものが多い．これらはボーカロイド楽曲動画と比べると再生時間が長いなど，特徴が大きく異なる．先の考察で述べた通り，動画集合の多様性が高いほど推定し難いと言える．

6. MLP を用いた回帰分析

6.1 MLP の構築

5. 節で定義した部分集合それぞれに回帰分析を適用する．線形回帰分析では再生回数が低いほど誤差率が大きくなるため，全体の平均誤差率が大きくなる．そこで，式 (8) の関数 L を誤差関数とする MLP を定義し，再生回数のスケールが異なる場合でも誤差率が同程度になるよう学習した [14]．

表 10　各クラスタの特徴 (2C)

クラスタ	tfidf 上位 5 つのタグ	データ数	平均誤差率
0	ニコニコカラオケ DB, ボカロカラオケ DB, 能く在る輪廻と猫の嘶, 君はいなせなガール, KTKT	3403	1.6364
1	ニコニコカラオケ DB, ニコカラ, ボカロカラオケ DB, アコギ, おさむらいさん	2501	0.8822
2	ニコニコカラオケ DB, ニコカラ, ボカロカラオケ DB, ぽからまとめ, UTAU	2840	1.4671
3	MikuMikuDance, 東方 MMD, MMD, Tda 式 Append ミク, アイドルマスター	2894	0.6726
4	MikuMikuDance, MMD, Tda 式 Append ミク, 亞北ネル, MMD ドラマ	2532	0.6842
5	MMD, MikuMikuDance, らぶ式ミク, Lat 式ミク, Tda 式 Append ミク	3634	0.7908
6	VOCALOID → UTAU カバー曲, UTAU 音源配布所リンク, UTAU, UTAU 海外組, UTAU カバー曲	2455	1.3067
7	VOCALOID → UTAU カバー曲, UTAU, UTAU オリジナル曲, 波音リツ, 波音リツキレ音源	3916	0.9224
8	UTAU, 中二病でも恋がしたい！, 鼻 8P, パブ MEIKO, ななみ	527	12.3965
9	HANASU, UTAU, トークロイド, 健音テイ, 重音テト	384	1.2175
10	UTAU オリジナル曲, テトオリジナル曲, UTAU, UTAUROCK, 重音テト	1404	0.9546
11	ミクオリジナル曲, 期待の新人, もっと評価されるべき, VOCAROCK, VOCALOID 処女作	1221	0.4434
12	VOCAROCK, ミクオリジナル曲, 初音ミク, VOCALOID 処女作, リンオリジナル曲	1265	0.7892
13	ミクオリジナル曲, 初音ミク, 初音ミクオリジナル曲, VOCALOID 処女作, VOCALOID オリジナル曲	2301	1.1864
14	ミクオリジナル曲, 初音ミク, ききいるミクうた, かわいいミクうた, 初音ミク Append	4315	1.1770
15	中二病でも恋がしたい！, 鼻 8P, パブ MEIKO, ななみ, とこしえ	1529	11.3750
16	がくっぽいど, 神威がくぽ, がくぽオリジナル曲, がくっぽいど誕生祭 2012, がくこ	1508	1.0588
17	VOCALOID カバー曲, T.M.Revolution, KAITO, ボカロクラシカ, VOCALOID → VOCALOID カバー	1462	1.3324
18	日刊 VOCALOID ランキング, VOCALOID ランキング, YAMAHA, UTAU ランキング, 鏡音ランキング	1045	2.3762
19	リンオリジナル曲, KAITO, VOCALOID → VOCALOID カバー, 鏡音レン, レンオリジナル曲	2268	1.4114
20	トークロイド, 漫才, ボカロ実況, ボカロラジオ, バトロセンタブロデューサ	1509	1.3188
21	GUMI オリジナル曲, メグッポイド, V3GUMI, GUMI, VOCAROCK	2025	0.9700
22	GUMI オリジナル曲, V3GUMI, VOCALOID3, メグッポイド, GUMI 新曲ランキング	1315	0.7823
23	GUMI オリジナル曲, V3GUMI, クロスフェード, メグッポイド, GUMI	2181	0.7766
24	ボカロオリジナルを歌ってみた, 歌ってみた, 合唱シリーズ, ハモリ, ユーザー生放送	611	2.1048
25	ニコカラ, inst, ユクエシレズ, 世界寿命と最後の一日, ジョバンニ	796	1.1514
26	VOCALOID → VOCALOID カバー, リンオリジナル曲, レンオリジナル曲, KAITO, 鏡音レン Append	2865	0.6481
27	ルカオリジナル曲, VOCAROCK, 巡音新曲ランキング, 英語ルカ, るからん	2206	0.7906
28	もっと評価されるべき, ミクオリジナル曲, ききいるミクうた, VOCALOID → VOCALOID カバー, かわいいミクうた	2635	0.4449
29	ミクオリジナル曲, VOCALOID 殿堂入り, ききいるミクうた, 初音ミク, かわいいミクうた	4056	0.6142
30	カゲロウプロジェクト, 歌ってみた, ボカロオリジナルを歌ってみた, VOCAMASH, 踊ってみた	2468	1.3952
31	作業用 BGM, ボカロオリジナルを演奏してみた, 演奏してみた, 楽譜, 弾いてみた	1437	1.6857
32	つんでれんこ, ニコニコ技術部, 作業用 BGM, 描いてみた, VOCALOID を描いてみた	1355	1.4538
33	ProjectDIVAArcade, 初音ミク-ProjectDIVA-, DIVA エディット動画, PS3, DIVA 自作譜面	3376	0.9743
34	VOCALOID 生放送, ユーザー生放送, 日記, ボカロ系生放送, VOCALOID カバー曲	339	306.5681
35	盗作, 剽窃, 検索妨害, 無断転載, 唐澤貴洋	483	2.5826
36	韓国語, VocaNicoNight, 音楽, 俺の妹がこんなに可愛いわけがない, ラジオ	3094	2.9243
37	駅名記憶向上委員会, 駅名, VOCAMASH, マッシュアップ, 環境 P	1712	0.9296
38	ゲーム, 太鼓さん次郎, ProjectDIVAArcade, 開封動画, ポケモンパン	1282	1.5110
39	HD, PDA, 高画質, フル HD, PS3	785	0.6418
40	ボカロクラシカ, バッハ, ボカロ聖歌隊, ムソルグスキー, バラキレフ	831	1.6302
41	VOCALOID カバー曲, 衣装 P, おっさんホイホイ, スナック初音, みくみくテレビ探偵団	4103	1.3564
42	CeVIO, ささらオリジナル曲, さとうささら, CeVIO カバー曲, CeVIO オリジナル曲	391	0.7783
43	ゆかりオリジナル曲, 結月ゆかり, VOICEROID, 結月祭, exVOICE	1676	0.9428
44	IA オリジナル曲, IA, ボカロ動画紹介動画, VOCALOID → VOCALOID カバー, GUMI オリジナル曲	2855	1.4584
45	海外組 VOCALOID, 洛天依, MIRIAM, 洛天依オリジナル曲, Oliver	887	1.1578
46	VOCALOID3, VOCALOID → VOCALOID カバー, ラピスオリジナル曲, ガチャッポイド, リュウト	3619	1.1937

$$L = \frac{1}{n}\sum_{i=1}^{n}\{(log(y_i+1) - log(f_i+1)\}^2 \tag{8}$$

部分集合 2 に対し，中間層の数を 1,2,...,7，各層のノード数を 4,8,12,24,48 としてグリッドサーチを行った．その結果，中間層 6 層，各層のノード数 24 が最良であった．図 10 にモデル

を示す．

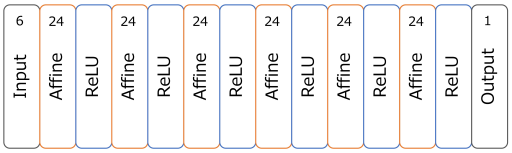

図 10　MLP

また，各 Affine 層に対し Dropout [15] を実装し，過学習を抑えた．Dropout は，ノードをランダムに消去しながら学習する手法である．Dropout を用いることで，学習用データと評価用データの精度の隔たりが小さくなる [16]．

6.2 手　　法

図 10 のモデルを 5. 節で定義した各クラスタに適用した．その際，各クラスタ内の動画を学習用と評価用のデータにランダムで 9 対 1 に分割した．パラメータの更新回数を 2000 回として学習し，評価用データに対して誤差関数 L が最小となる時点でのパラメータを保存した．学習済みモデルに対し学習用と評価用のデータを入力し，その推定結果を用いて各クラスタの平均誤差率を算出した．この操作を 10 回繰り返し，クラスタ毎に平均誤差率の平均値を算出した．また，部分集合 1A，...，1D，2A，...，2D それぞれにおいて，各クラスタの平均誤差率から，各部分集合の平均誤差率を算出した．

6.3 結　　果

表 11 に，部分集合 1A,...,1D,2A,...,2D の平均誤差率を示す．クラスタリングによる推定精度の差を見るため，クラスタリングを適用せずに上記手法で MLP を適用した際の結果も示す．また，線形回帰分析で平均誤差率が最低であった部分集合 2C における，クラスタ毎の平均誤差率を表 12 に，実測値と推定値の関係を図 11 に示す．

表 11 を見ると，1A 以外の部分集合ではクラスタリングにより推定精度が向上している．また，表 12 と表 10 を比較すると，各クラスタの平均誤差率は大幅に減少している．さらに，図 11 を見ても，線形回帰分析における低再生回数部分の広がりが抑えられていることが確認できる．

表 11　各部分集合における平均誤差率

	クラスタリングなし	クラスタリングあり
部分集合	平均誤差率	平均誤差率
1A	0.3488	0.3495
1B	0.4434	0.3452
1C	0.4450	0.3793
1D	0.4725	0.3912
2A	0.3938	0.3496
2B	0.4435	0.3508
2C	0.4559	0.3846
2D	0.4997	0.4030

表 12　各クラスタの平均誤差率 (2C)

クラスタ	データ数	平均誤差率	クラスタ	データ数	平均誤差率
0	3403	0.2636	24	611	0.4715
1	2501	0.2528	25	796	0.3246
2	2840	0.2469	26	2865	0.337
3	2894	0.4342	27	2206	0.281
4	2532	0.3362	28	2635	0.2659
5	3634	0.3833	29	4056	0.311
6	2455	0.3634	30	2468	0.594
7	3916	0.3394	31	1437	0.4841
8	527	0.1855	32	1355	0.4713
9	384	0.3312	33	3376	0.413
10	1404	0.3305	34	339	0.2378
11	1221	0.2296	35	483	0.6498
12	1265	0.2802	36	3094	0.5757
13	2301	0.2909	37	1712	0.491
14	4315	0.3578	38	1282	0.4606
15	1529	0.9034	39	785	0.2538
16	1508	0.34	40	831	0.3406
17	1462	0.4369	41	4103	0.459
18	1045	0.2224	42	391	0.3431
19	2268	0.4216	43	1676	0.3782
20	1509	0.3629	44	2855	0.5422
21	2025	0.3848	45	887	0.3542
22	1315	0.3477	46	3619	0.4481
23	2181	0.3464			

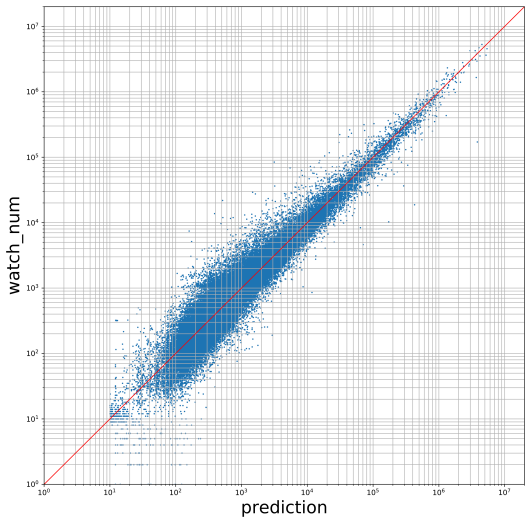

図 11　MLP による実測値と推定値の関係 (2C)

6.4 考　　察

表 12 を見ると，各クラスタで平均誤差率に差がある．また，データ数との相関性は低い．今回，動画集合の多様性を低くするため，タグを用いた階層的クラスタリングを適用した．しかし，平均誤差率の高いクラスタにおいては，依然として動画の多様性が高いと考えられる．

本研究では，閾値をデンドログラムの下位 15%に設定した．この閾値を下げることで，クラスタ数を増やし，より多様性の低い動画集合を定義できる可能性がある．また，デンドログラムの枝分かれの回数でクラスタを定義したり，非階層的手法を用いることで，改善を図れる可能性もある．学習データの数が少ないと Neural Network の性能が下がるという指摘もあるた

め [16]，最適なクラスタリングをする必要がある．

7. おわりに

本研究では，CGM における利用者のコンテンツ選択モデルを検討の第一歩として，ニコニコ動画における動画の再生回数を推定する回帰分析を行った．線形回帰分析の結果，音楽関連の動画では高い精度で推定できた．音楽関連の動画は，複数回再生することを前提とした利用者がマイリストに登録するため，再生回数とマイリスト数の相関性が他のカテゴリよりも高く，推定し易い．推定精度が最も高い VOCALOID 関連の動画について，投稿時期で動画集合を分割し，それぞれに線形回帰分析を適用した．その結果，動画集合の多様性が高いほど推定し難いことが分かった．動画集合の多様性を減らすために，Word2Vec を用いて動画のタグをベクトル化し，Ward 法による階層的クラスタリングをした．各クラスタに対して MLP による回帰分析を適用した結果，推定精度を大幅に向上できた．

タグ情報を用いたクラスタリングと MLP による回帰分析を行うことで，動画再生回数を高い精度で推定できた．各説明変数およびタグと再生回数の関係は明らかではなく，CGM における利用者のコンテンツ選択モデルの確立には到っていない．また，SNS でのシェアやサムネイル画像も動画選択に影響していると考えられる．本研究で用いた説明変数およびタグ情報による動画ベクトルに加え，SNS シェア数などのソーシャル的特徴，また，画像解析的側面から定量化したサムネイル画像の特徴を用いた，新たなモデルを確立する必要がある．

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] N. Murakami and E. Ito, "Emotional video ranking based on user comments," in *Proceedings of iiWAS2011*. ACM, December 2011, pp. 499–502.

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

[3] 株式会社ドワンゴ, "ニコニコ静画閲覧数・お気に入り数予測," https://nico-opendata.jp/ja/casestudy/view_count_prediction/index.html, 2015.

[4] 柴田知親，伊東栄典, "回帰分析を用いた動画再生回数の推定," 平成 29 年度　電気・情報関係学会九州支部連合大会, vol. 70, pp. 367–368, 2017.

[5] カドカワ株式会社, "2017 年 3 月期 通期決算説明資料," http://pdf.irpocket.com/C9468/wReJ/oisY/GePU.pdf, 2017.

[6] 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.

[7] 久保拓弥, データ解析のための統計モデリング入門一般化線形モデル 階層ベイズモデル MCMC.　岩波書店, 2012.

[8] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," *Proceedings of the 9th Python in Science Conference*, pp. 57 – 61, 2010.

[9] 有賀康顕，中山心太，西林孝, 仕事ではじめる機械学習.　株式会社オライリー・ジャパン, 2017.

[10] 柴那典, 初音ミクはなぜ世界を変えたのか？.　株式会社太田出版, 2014.

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

[12] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," http://www.scipy.org/, 2001–.

[13] 平井有三, はじめてのパターン認識.　森北出版株式会社, 2012.

[14] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," https://www.tensorflow.org/, 2015.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[16] 斎藤康毅, ゼロから作る Deep Learning —Python で学ぶディープラーニングの理論と実装.　株式会社オライリー・ジャパン, 2016.