回帰分析を用いた動画再生回数の推定

柴田, 知親
九州大学工学部電気情報工学科

伊東, 栄典
九州大学情報基盤研究開発センター

出版情報：電気関係学会九州支部連合大会講演論文集, pp.367-368, 2017-09-28. 電気・情報関係学会九州支部
バージョン：
権利関係：許諾番号：17SB0112

回帰分析を用いた動画再生回数の推定

柴田 知親 *　伊東 栄典 **
(九州大学 *工学部電気情報工学科 **情報基盤研究開発センター)
* t.shibata.130@s.kyushu-u.ac.jp, ** ito.eisuke.523@m.kyushu-u.ac.jp

1　はじめに

ネットワークおよびモバイル端末の普及に伴い, 動画，小説, 画像などのコンテンツ投稿サービス（CGM, Consumer Generated Media）が人気になった. 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 [1]。近年 CGM コンテンツの多様性減少の懸念が出ている. ニコニコ動画の投稿動画の多様性動向を調査した [2]. 次に我々は，CGM における利用者のコンテンツ選択モデルを検討する. その第一歩として，動画の再生回数を推定する回帰分析を行った.

2　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, 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。各動画のメタデータは表 1 に示す属性を含む.

表 1: ニコニコデータセット内のデータ

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　手法と方針

本研究では動画再生回数の推定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. 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.

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. そのためリンクが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.

表 2 に示す 6 つを線形重回帰分析の説明変数とした.

表 2: 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

4　線形回帰分析

4.1　概要

最初にデータセット内の全動画と，各カテゴリ動画の部分集合で線形回帰分析を適用した. 次に, 推定精度の高い動画集合について詳細に分析した. 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, 各実測値を y_i, 回帰分析によるそれらの推定値を f_i, 実測値の平均を μ とすれば, 以下の式で表される.

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

決定係数の値は 0 から 1 の範囲で表される. 1 に近いほど回帰式が実際のデータに当てはまることを意味する.

4.2　結果

表 3 に, 全動画と上位 5 つのカテゴリの決定係数を示す. 「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.

表 3: 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

5　VOCALOID 動画の詳細分析

ボーカロイド楽曲はニコニコ動画発祥の文化であり, かつ VOCALOID カテゴリ動画は表 3 で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.

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 1 にグラフで示す. 図 1 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.

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. しかし以前の投稿動画は, 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 4 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 個の部分集合に線形重回帰分析を適用した.

5.1　結果

今回は各決定係数に大きな差が見られないため, 各動画の推定値の誤差を | 実測値 – 推定値 |, 誤差率を | 実測値 – 推

表 4: 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)

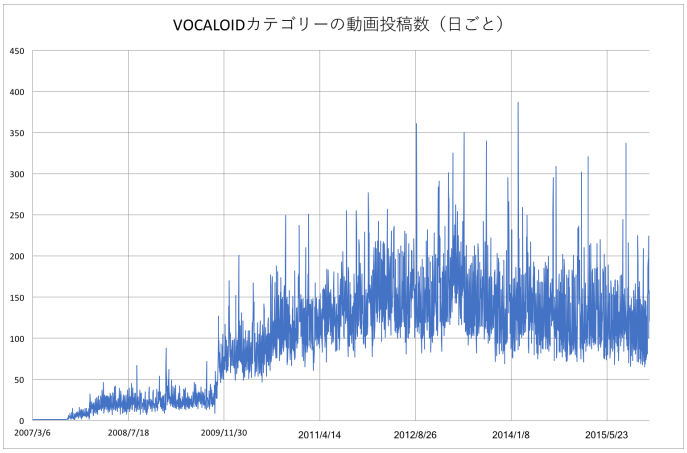

図 1: VOCALOID カテゴリの動画投稿数

定値 |/実測値 と定義し, 誤差率の平均値を各分類の推定し易さの指標とした. 表 5 に結果を示す. 図 2 は, 標準化偏回帰係数の値から, 各説明変数が目的変数へ与える影響の度合である. 図 3 は, 最も推定し易いと考えられる 1D について, 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ものである.

表 5: 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

5.2 考察

図 2 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で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ため, 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. この予想は, 音楽関連カテゴリの動画で決定係数の値が高い結果に繋がる.

図 3 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い. 低再生回数の動画は多数存在し, それらは品質もバラバラであるため推定が難しいのであろう.

6 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, 動画の再生回数を推定する回帰分析を行った. ニコニコ動画の音楽関連の動画では, 線形回帰分析で再生回数を高い精度で推定できた.

動画の多様性を考慮すると, 動画メタデータのうち本研究では使用していない文字列データを用いたクラスタリングにより, 推定精度の向上が期待できる. またサムネイル画像の影響も大きい。今後は文字データによるクラスタリングや, サムネイル画像も利用した分析, 特に非線形回帰分析を適用していきたい.

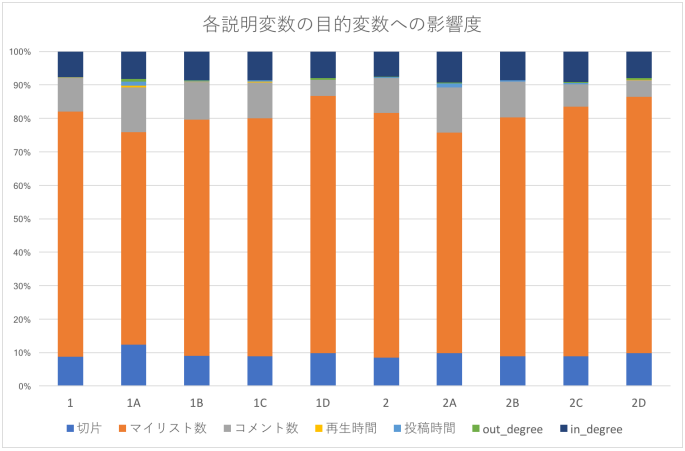

図 2: 各説明変数の目的変数への影響度

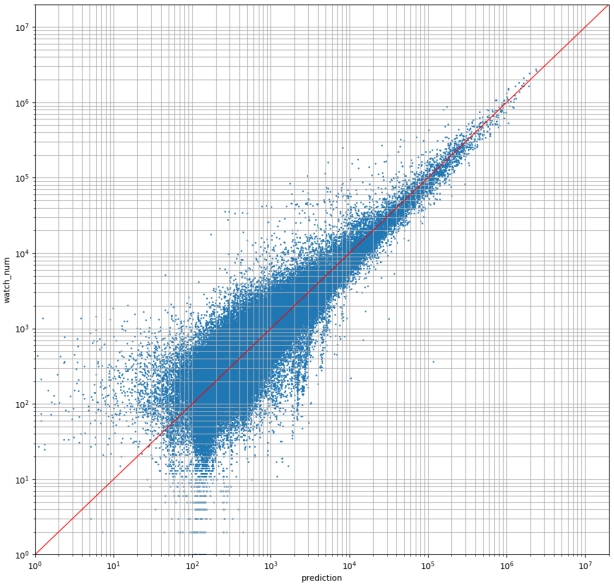

図 3: 1D における実測値と推定値の関係

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] Naomichi Murakami and Eisuke Ito, "Emotional video ranking based on user comments," Proc. of ACM iiWAS2011, pp.499-502, 2011.

[2] Kyohei Kamihata and Eisuke Ito, "A quantitative contents diversity analysis on a consumer generated media site," Proc. of AROB 21st 2016, pp.436-440, 2016.

[3] Seabold, Skipper, and Josef Perktold, "Statsmodels: Econometric and statistical modeling with python," Proc. of the 9th Python in Science conference, pp.57–61, 2010.