ニコニコ動画における動画タグの多様性分析

上畑 恭平　　伊東 栄典　　馬場 謙介

九州大学工学部，九州大学情報基盤研究開発センター，九州大学図書館

kyohei.kamihata.123@s.kyushu-u.ac.jp　ito.eisuke.523@m.kyushu-u.ac.jp

baba.kensuke.060@m.kyushu-u.ac.jp

1. はじめに

近年，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が投稿されており，また膨大な利用者が動画を閲覧している．動画以外にも，小説投稿サイトや写真共有サイトも人気である．

我々は，ニコニコ動画を対象に，視聴者投稿コメントの感情分析に基づく動画ランキング手法の研究してきた[1]. また他の CGM サイトとして，小説投稿サイト「小説家になろう (syosetu.com) 」を対象に，小説に付与されたタグの分析や[2]，お気に入り登録の構造解析に基づく小説ランキング手法を研究してきた[3].

現在，CGM サイトに投稿されるコンテンツの画一化が指摘されている．ニコニコ動画を運営するドワンゴ社川上量生氏へのインタビュー記事[4]では，再生回数上位の動画は，同一カテゴリの動画になりつつあるという傾向を指摘している．Fukatsu 氏のブログ記事[5]では，「小説家になろう」サイトに投稿される小説で，人気が上位になる小説が画一化することについて考察している．

コンテンツの多様性が減少し，画一化が進むと，文化的多様性が失われると思われる．ある特定の環境に特化し過ぎて多様性を失った文化からは，新たな文化的イノベーションが発生しにくいと思われる．

我々は，CGM サイトであるニコニコ動画を対象に，動画に付与されたタグの多様性を分析する事にした．本論文の構成を述べる．2 節では国立情報学研究所が提供するニコニコデータセットについて述べる. 3 節では動画集合における，様々な頻度解析について述べる. 4 節では，本論文の主題である多様性について定義する．5 節では，実データを用いた多様性の測定および時系列での動向を示し，その考察を行う．最後に 6 節でまとめと今後の課題を述べる．

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 2006 年 12 月 12 日にサービスを開始した，視聴者投稿型の動画配信サービスである．運営開始から 8 年経過した 2014 年 12 月末現在，1100 万件を超える動画が投稿されている．会員数も膨大で，wikipedia[6]によると 2013 年 6 月時点での一般会員のアカウント数は 3000 万を超えており，有料のプレミアム会員数も 200 万を超えている．

2.2. ニコニコデータセット

国立情報学研究所は，情報学研究リポジトリと名付けた，研究用のデータ集合を提供している．ドワンゴ社および未来検索ブラジル社は，国立情報学研究所に協力して研究者にニコニコデータセットを提供している[7]. 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．本研究では前者の動画コメント等データを利用している．前者のデータ数などの概要を表 1 に示す．

表 1　動画コメント等データ概要

項目	内容
期間	2007 年 3 月～2012 年 11 月
形式	JSON 形式
データ件数（動画数）	8,305,696
一意なタグ	5,328,341

ニコニコ動画コメント等データに含まれている項目の一部を表 2 に示す．

表 2　動画メタデータに含まれる項目

項目	説明
video_id	動画 ID
title	動画の題名
description	動画の説明文
upload_time	動画投稿日時
length	動画長
movie_type	動画のファイル形式
view_counter	閲覧回数（再生回数）
comment_counter	コメント数
mylist_counter	マイリスト登録数
tags	タグ

3. 動画の頻度解析

ニコニコデータセットの，動画メタデータを用いて，頻度などを調査した．全体の傾向の他に，ボーカロイド関連の音楽エンターテイメント動画の動向も調査した．これを省略のために，V 系動画と呼ぶ．

V 系動画は「エンタメ・音楽」カテゴリの中に含まれる，「音楽，歌ってみた，演奏してみた，踊ってみた，VOCALOID」の 5 つの単語のいずれかをタグに含む動画とする．V 系動画は，2007 年頃に発生し，2013 年頃までに大きな盛り上りを見せたニコニコ動画発のボーカロイド音楽文化 [8] に関する動画と考えている．

3.1.　各月の動画投稿数

各月の動画投稿数を図 1 に示す．

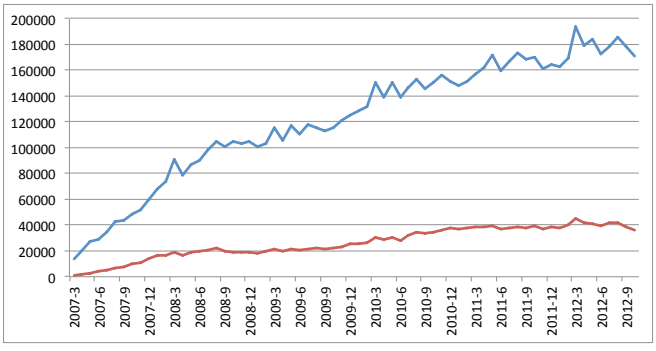

図 1 動画投稿数

図 1 の青線は全動画投稿数で，赤線は V 系動画の投稿数である．増減はあるものの，概ね右肩上がりに投稿動画数が増えている．2012 年の動画の投稿数は月 18 万個程度で，その V 系動画は 4 万個程度である．赤線の V 系動画は 2007 年 10 月以降，全体の約 20%を占めている．

3.2.　一意なタグ数

次にその月に投稿された動画集合を対象に，付与されたタグについて調査した．図 2 に各月の一意なタグ数を示す．

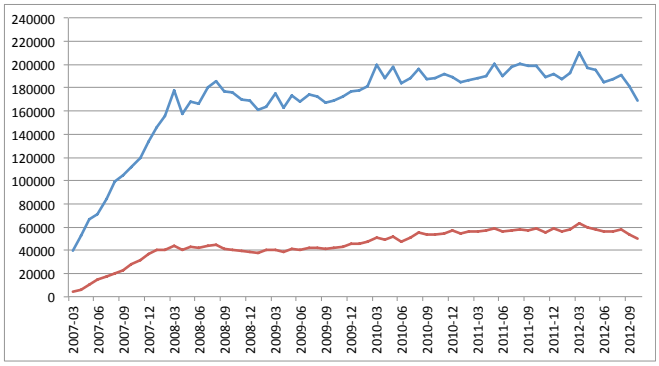

図 2　各月の一意なタグ数

図 2 の青線は全動画でのタグ数で，赤線は V 系動画のタグ数である．2008 年 3 月まで急激に増加し，その後は毎月 180 万個程度のタグ数になっている．赤線の V 系動画集合から抽出されるタグは，概ね全体の 25%を占めている．

3.3.　動画再生回数の順位-頻度

動画の再生回数を降順で並べたデータを作成した．そのデータに基づき，縦軸に再生回数，横軸に順位を取った散布図を図 3 に示す．なお，両軸とも対数尺度にしている．

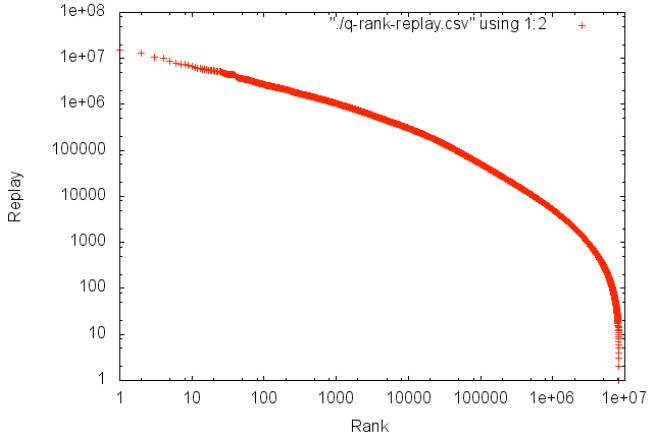

図 3　動画の順位-再生回数（対数尺度）

図 3 で分かるように，再生回数上位の動画の分布は直線に近い．両対数グラフで直線であるため，冪乗則（Power law）に近い分布をしている．しかしながら，再生回数の低い部分は，直線ではない．

次に，横軸に再生回数，縦軸にはその再生回数を持つ動画の数を散布図で描いた．この散布図を図 4 と図 5 に示す．

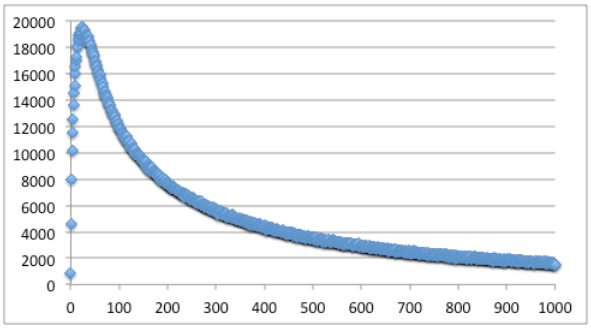

図 4　再生回数-動画数（1000 回以下）

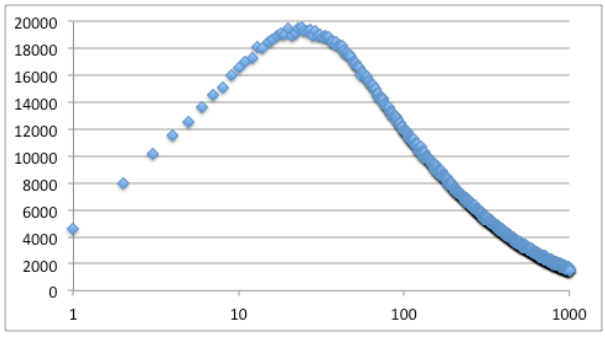

図 5　再生回数-動画数（1000 回以下・横軸対数尺度）

図 5 を見ると分かるように，横軸を対数尺度にすると，正規分布に近い曲線を描くことが分かる．このため，再生回数の分布は対数正規分布に近い分布であることが分かる．

3.4.　タグ頻度（出現回数）の順位-頻度

動画に付与されたタグについて，各タグの出現回数（頻度）を降順で並べたデータを作成した．そのデータに基づき，縦軸に頻度，横軸に順位を取った散布図を図 6 に示す．なお，両軸とも対数尺度にしている．

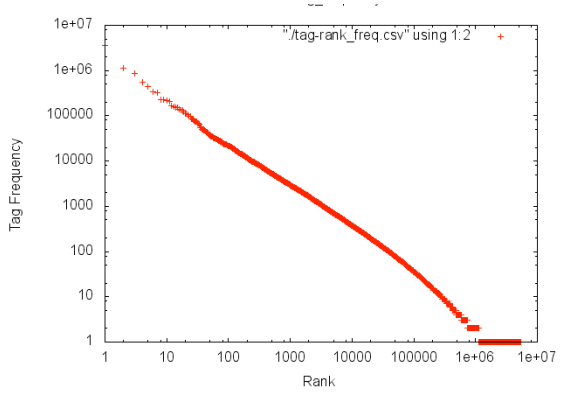

図 6　タグの順位-頻度（両対数尺度）

図 6 の分布は両対数尺度で直線を示している．このためタグの出現頻度は冪分布であることが分かる．小説などの自然言語文における単語の出現頻度分布は冪分布になる．動画のタグ群も自然な分布をしていると言える．

4. タグの多様性

先に，インタビュー[4]やブログ[5]で，CGM サイトへの投稿コンテンツの多様性減少への懸念が指摘されていることを述べた．筆者らの感覚としても多様性が減り，画一化が進んでいるように感じられる．本当に多様性が減少しているのかを判断するためには定量的な指標が必要である．

本研究ではコンテンツ多様性の度合いを数値で評価する指標を提案する．そのため，動画に付与されているメタデータを，特に動画に付与された単語を用いて多様性の度合いを数値化する．

4.1.　多様性についての考え方

コンテンツの多様性について考えるため，最初に極端な場合を考える．もしもコンテンツが完全に画一化されているならば，全てのコンテンツに付与されるタグも同じになる．コンテンツ数（文書数）を n，タグの単語 w の文書頻度を $df(w)$ とすると，全てのタグ w について $df(w)=n$ である．

逆に完全に多様であれば，全コンテンツに付与されるタグが異なるであろう．完全に多様な場合は，全てのタグ w について $df(w)=1$ である．

実際の動画では，カテゴリやジャンルを指定するタグを付与する．カテゴリタグは 30 個で有限であるため，これは多様にならない．また，図 6 で示したように，多くのタグは出現頻度が 1 である．頻度 5 以下のタグが殆どであるため，低頻度のタグだけを見て多様であるとすることは望ましくない．

4.2.　タグ多様性の定義

情報エントロピーの考えを用いて，コンテンツ集合（文書集合）に対するタグの多様性を定義する．その際，以下の記号を用いる．

D　：　文書集合，
n　：　文書数（$|D| = n$），
W　：　タグ集合，
$\mathrm{df}(w)$　：　タグ w の文書頻度．

情報エントロピーの考えた方を用いて，集合 D とタグ集合 W の多様度を定義する．

$$H(W) = -\sum_{w \in W} p(w)\log(p(w)), \quad 0 \le p(w) \le 1.$$

ここで $p(w)$ はタグ w の出現確率である．ニコニコ動画では，1 つの動画に 1 つのタグを複数回付与できない．そのため $p(w) = \mathrm{df}(w)/n$ になる．

5. タグの多様性動向

情報エントロピーをタグに適用したタグ多様性の度合いを，各月の投稿動画に付与されているタグデータで算出した．各月のタグの多様度の推移を図 7 と図 8 に示す．どちらの図でも，図 2 に示した一意なタグの数を表示している．

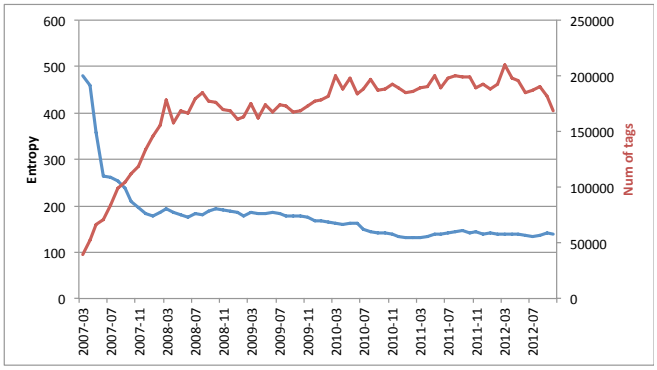

図 7　タグ多様度（青）と一意なタグ数（赤）

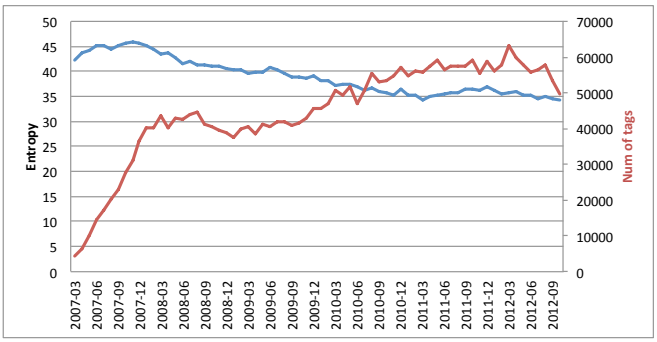

図 8　V 系動画限定：
タグ多様度（青）と一意なタグ数（赤）

図 7 と 8 を見ると，どちらでも一意なタグ数(赤線）は緩やかに増加しているのに対し，タグ多様度（エントロピー）は減少している．

6. おわりに

本論文では，近年人気の CGM サイト，ニコニコ動画を対象に，コンテンツの多様性の動向を調査した．投稿動画に付与されるタグについて，情報エントロピーの定義を援用して，毎月のタグ多様性を数値で表現した．

情報エントロピーを用いて，月ごとのタグ多様性を算出し，それを時系列で折れ線グラフ表示した．その結果，一意なタグ数（赤線）は緩やかに増加しているのに対し，タグ多様度（エントロピー）は減少していることが分かった．

今後は，タグ以外の題名や説明文を含めたエントロピーも算出したい．今回は動画に付与されるタグを独立と考えたが，実際には意味的な依存から，従属出現するタグが有ると思われる．これについても調査したい．

将来は電子コンテンツにおける利用閲覧モデルも考えたい．多様性喪失の原因として，端末の狭さがあると思われる．書店や図書館と異なり，PC 等では多数のコンテンツを一覧できない．また，コンテンツを試すには一つずつ閲覧するしかない．独力で多数を試すには時間が掛かるため，既知コンテンツに近いものを閲覧するのであろう．作者も，人気を得やすい分野のコンテンツを作りたがる傾向がある．利用者の閲覧モデルを作ることで，多様性喪失の原因が分かり，そこから多様性を保持する閲覧ソフトの開発ができると思われる．

文　献

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

[2] Eisuke Ito, Kazunori Shimizu: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012, pp.531-536, Nov. 2012.

[3] Kazunori Shimizu, Eisuke Ito, Sachio Hirokawa: Predicting Future Ranking of Online Novels based on Collective Intelligence, Proc. of ICDIPC2013, SDIWC, pp.261-272, 2013.

[4] Cakes, 川上量生: 川上量生の胸のうち, https://cakes.mu/posts/5036 (accessed at Dec.12, 2014).

[5] Takayuki Fukatsu, 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 (accessed at Dec.12, 2014).

[6] ニコニコ動画 (Dec.12,2014) in *Wikipedia: The Free Encyclopedia*. Retrieved from http://ja.wikipedia.org/wiki/%E3%83%8B%E3%82%B3%E3%83%8B%E3%82%B3%E5%8B%95%E7%94%BB

[7] 国立情報学研究所, ドワンゴ社: ニコニコデータセット: http://www.nii.ac.jp/cscenter/idr/nico/nico.html, (accessed at Dec.12, 2014).

[8] 柴那典, 初音ミクはなぜ世界を変えたのか？, Amazon Services International, 2014.