タイトルと説明文に着目した利用者投稿サイト動画の多様性分析

上畑, 恭平
九州大学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

出版情報：火の国情報シンポジウム論文集. 2016, pp.1-6, 2016-03-02. 情報処理学会九州支部
バージョン：

タイトルと説明文に着目した 利用者投稿サイト動画の多様性分析

上畑恭平[†1]　伊東栄典[†2]

近年，YouTube やニコニコ動画などの利用者投稿型動画サイトが人気である．これらの CGM (Consumer Generated Media) サイトには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．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．見たことのあるような動画や，派生動画が増えていると感じられる．我々はニコニコ動画を対象に，動画コンテンツの多様性動向を分析している．今回, 動画のタイトルと説明文を対象に，出現する単語の類似度に着目して分析した．

CGM movie contents diversity analysis based on words in title and description

KYOHEI KAMIHATA[†1]　EISUKE ITO[†2]

Recent years, CGM (Consumer Generated Media) sites, such as YouTube and nicovideo, become popular. CGM is contents delivery media, and some of them are able to give an influence on society. A lot of movies are posted to a CGM site every day, and also a large number of users are enjoying posted movies. At present, decreasing diversity of contents are indicated by some bloggers. Posted movies may be similar with previous posted movies. The authors are afraid that decreasing diversity of contents causes less energetic cultural activity. In this paper, the authors tried to measure diversity of contents in a CGM site based on words in title and description. They propose two quantitative metrics of contents diversity, and apply them to a CGM contents in nicovideo.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，動画に付与されたタグの多様性の定量的な評価[1][2]や, 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[3]. 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[4]，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[5].

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[6]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．

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イトルと説明文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，タイトルと説明文の多様性と情報エントロピーについて述べる．5 節では，cos 類似度について述べる．6 節では，実データを用いたタイトルと説明文の cos 類似度の測定及び時系列での動向を示し，考察を行う．最後に 7 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[7]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．

†1 九州大学大学院システム情報科学府
Department of ISEE, Kyushu University
†2 九州大学情報基盤研究開発センター
Research Institute for IT, Kyushu University

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[8]．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． また，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 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
一意なタグ数	5,328,341

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	動画に付与されたタグ

3. 動画の頻度分析

ニコニコデータセットの，動画メタデータを用いて，月ごとの動画投稿数，タイトルと説明文の単語数，頻度などを調査した．

3.1 形態素解析

ニコニコデータセットの，各動画メタデータに含まれるタイトルと説明文を Mecab で形態素解析を行った．Mecab での形態素解析は，mecab-ipadic-neologd をシステム辞書として使用した．mecab-ipadic-neologd は，Mecab の標準のシステム辞書では正しく分割できない新語・固有表現などを読み仮名と原型付きで約 208 万組採録しているシステム辞書である．形態素解析を行った後，各動画メタデータの単語集合の中から名詞のみを抽出した．

3.2 各月の動画投稿数

各月の動画投稿数を図 1 に示す．図 1 から，2007 年 3 月から 2012 年 11 月までの間，概ね右肩上がりに投稿動画数が増えていることが分かる．2012 年の動画の投稿数は月 18 万個程度である．

3.3 一意な単語(名詞)数

次にその月に投稿された動画集合を対象に，付与されたタイトルと説明文の単語について調査した．図 2 に各月の一意な単語(名詞)数を示す．タイトルと説明文の単語どちらも各月の動画投稿数に比例して増加している．

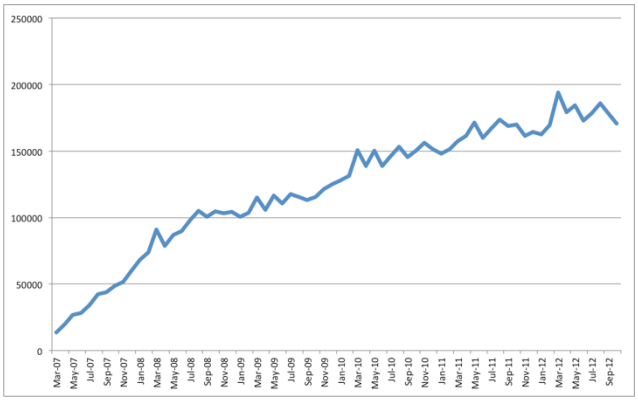

図 1 動画投稿数

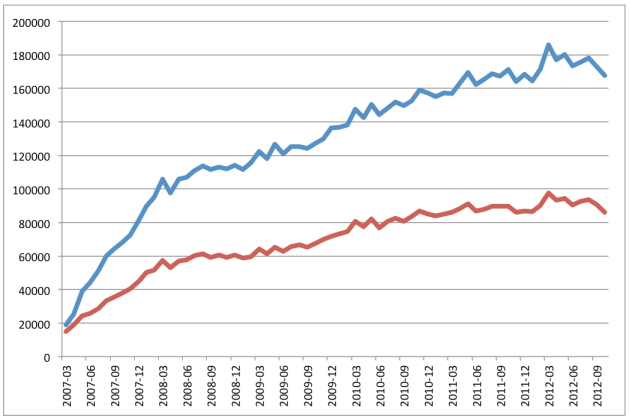

図 2　各月の一意な単語(名詞)数
(赤線：タイトル，青線：説明文)

3.4 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．

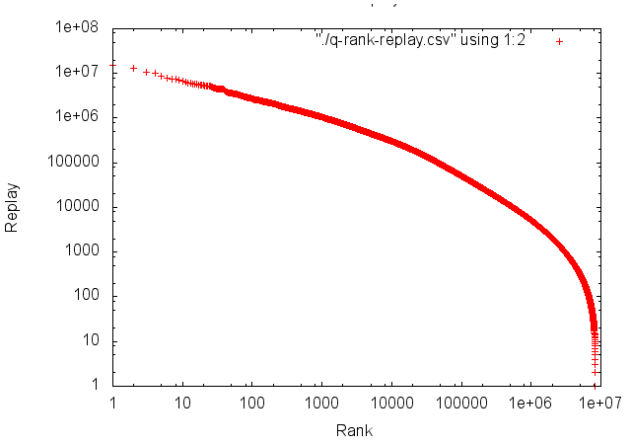

図 3　動画の順位-再生回数（対数尺度）

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 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.

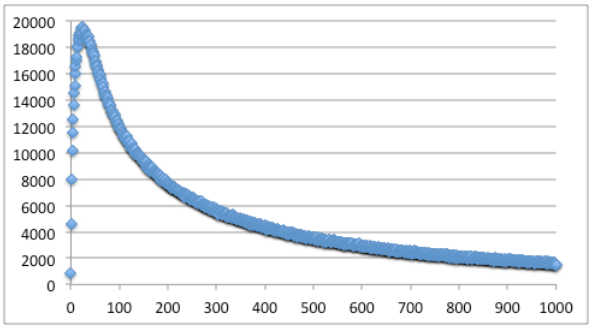

図 4　再生回数-動画数（再生回数 1000 回以下）

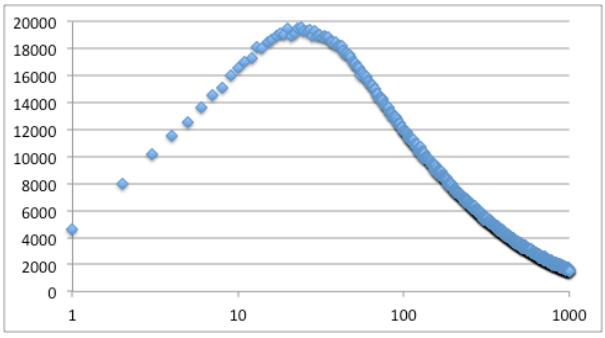

図 5　再生回数-動画数（再生回数 1000 回以下・横軸対数尺度）

3.5 単語(名詞)の頻度(出現回数)の順位-頻度

動画に付与されたタイトルと説明文について，各単語（名詞）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それぞれ図 6，図 7 に示す．なお，両軸とも対数尺度にしている.

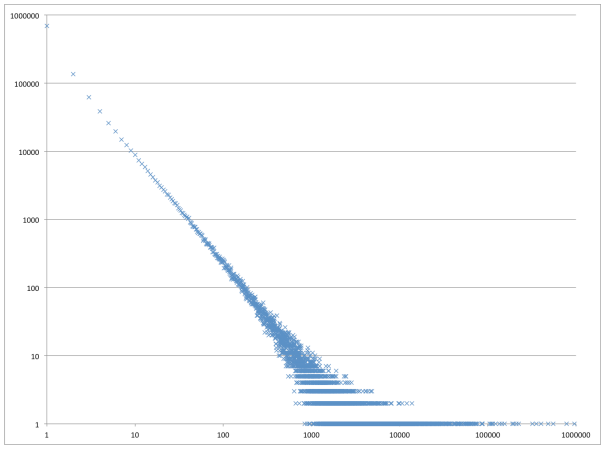

図 6　タイトルに出現する単語(名詞)の順位-頻度（両対数尺度）

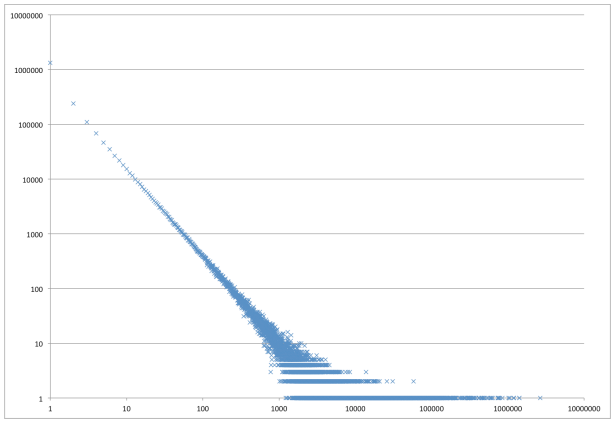

図 7　説明文に出現する単語(名詞)の順位-頻度（両対数尺度）

4. 情報エントロピーによる多様性分析

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タイトルと説明文の単語(名詞)を用いてそれぞれの多様性の度合いを数値化する.

4.1 単語の情報エントロピー

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）に対するタイトルと説明文の多様性を定義する．その際，以下の記号を用いる．

D	:	動画集合,
n	:	動画数 ($\|D\| = n$),
W	:	タイトルと説明文それぞれの単語(名詞)集合,
$df(w)$	:	単語 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 と単語集合 W の多様度を単語当たりの情報エントロピー $H(W)$ として定量化する．

$$H(W) = -\Sigma p(w)\log\big(p(w)\big),$$
$$p(w) = \frac{df(w)}{\sum df(w)}, \qquad 0 \leqq p(w) \leqq 1.$$

ここで $p(w)$ は単語 w の出現確率である．単語 w の出現確率は $p(w) = df(w)/\Sigma\, df(w)$ になる．

4.2 タイトル・説明文の多様性動向

情報エントロピーをタイトルと説明文のそれぞれの単語(名詞)集合に適用した．単語の多様性の度合いである $H(W)$ の値を，各月の投稿動画のタイトルと説明文のそれぞれの単語を用いて算出した．図 8，図 9 の青線は各月の $H(W)$の推移である．また，図 8，図 9 の赤線は，各月の動画のタイトルと説明文の単語集合における一意な単語(名詞)の数を示す．

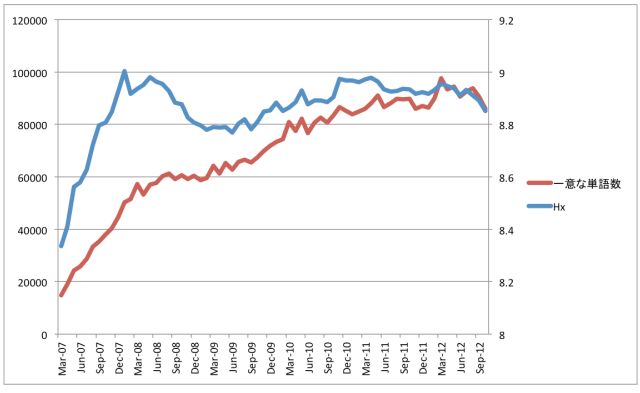

図 8　タイトルの単語の多用度（青線）と一意な単語(名詞)数（赤線）

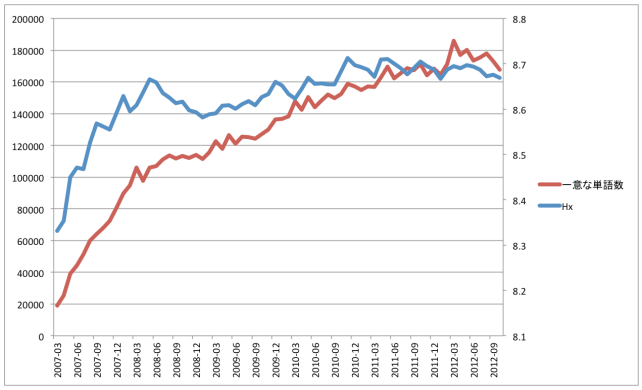

図 9　説明文の単語の多用度（青線）と一意な単語(名詞)数（赤線）

図 8，図 9 を見るとどちらのエントロピーも 2007 年 12 月までは増加しているが，そこから 2008 年 12 月までは減少している．その後，また緩やかに増加している．

5. 類似度による多様性分析

次に，cos 類似度による多様性分析について述べる．

5.1 考え方

図 10 は，類似度に基づく多様性分析の概念図である．図中の点をコンテンツ(動画)と考える．図 10 の左を多様，右を多様でないと考える．図 10 の左のようにコンテンツが多様性であるならば，各動画間の距離は長くなり，類似度は小さくなる．図 10 右のように多様性が減少しているならば，距離が長くなり，類似度は大きくなる．

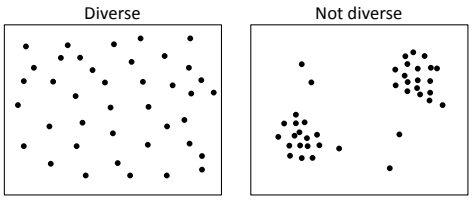

図 10　類似度に基づく多様性分析のモデル

2 点間の距離，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．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[9].

これら 4 つのうち，比較的計算が容易であり，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．コンテンツ集合の，全ペアの cos 類似度を計算し，それらの総和で多様性を定量化する．

5.2 cos 類似度について

cos 類似度とは，２つの文書間の類似度を測る手法の一つである．文書をベクトルとみなして，２つのベクトルの向きの近さを類似度の指標としたものが cos 類似度である．

一般に cos 類似度は-1 から 1 の値を取るが，

値が大きいほど２つの文書は似ていると言える．本研究では，各動画に付与されたタイトルと説明文の単語群をそれぞれ一つのベクトルとみなし，２つの単語群の全ての組み合わせについて cos 類似度を算出し，それらを足し合わせることで cos 類似度の総和を求める．対象とする文書数を同じにして，cos 類似度の総和を比較することで，文書の類似度が増加しているかを判断できる．

5.3 cos 類似度の定義

以降で用いる記号を，以下のように定義する．

D	:	文書集合,
n	:	文書数 ($\|D\| = n$),
W	:	タイトルと説明文それぞれの単語集合,
d_i	:	動画 i に付与された単語群ベクトル.

cos 類似度を算出する際の２つの単語群ベクトルを d_i, d_j とし．それぞれのベクトルが以下であるとする．

$$d_i = (a,a,b,c,d),$$
$$d_j = (a,c,e,e,f,g).$$

この時 d_i と d_j で次元が異なるので，次元を揃えたベクトル x を考える．

$$x = (a,b,c,d,e,f,g).$$

d_i と d_j から x の要素の数を表したベクト d'_i と d'_j を作成する．

$$d'_i = (2,1,1,1,0,0,0),$$
$$d'_j = (1,0,1,0,2,1,1).$$

この２つのベクトル d'_i と d'_j を用いて cos 類似度を算出する．

$$Cosine\ Similarity = \frac{d'_i \cdot d'_j}{\sqrt{|d'_i|} \cdot \sqrt{|d'_j|}}$$

文書集合 D の２つの単語群の組み合わせについて cos 類似度を算出し，それらを足し合わせて cos 類似度の総和を算出した．なお，組み合わせの数は $n(n\text{-}1)/2$ 個になる．

6. cos 類似度の総和の動向

動画に付与されたタイトルと説明文の単語群の cos 類似度の総和を月ごとに算出した．ここでは，文書数(動画数)は, $n = 1000$である. 図 11 では, 各月の再生回数上位 1000 個の動画のタイトルの単語を対象に cos 類似度の総和を算出した．図 12 では，各月の再生回数上位 1000 個の動画の説明文の単語を対象に cos 類似度の総和を算出した．

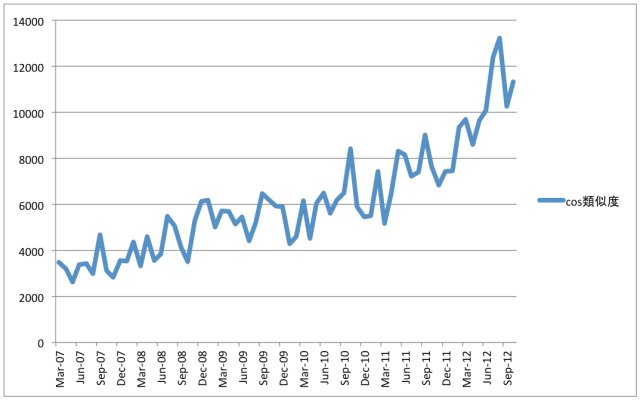

図 11　タイトルの cos 類似度の総和
(再生回数で上位 1000 個の動画)

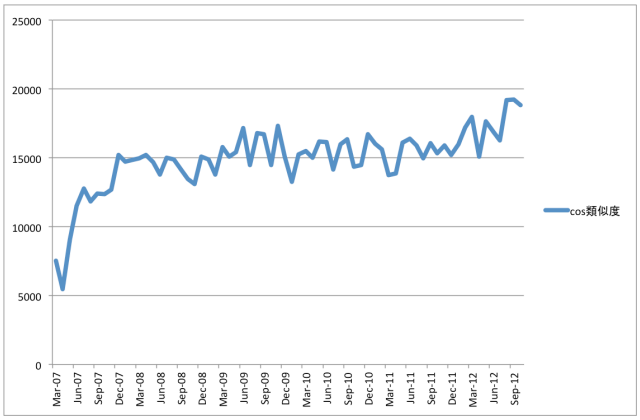

図 12　説明文の cos 類似度の総和
(再生回数で上位 1000 個の動画)

図 11 を見ると，各月のタイトルの cos 類似度の総和が穏やかに増加しているのがわかる．それに対して図 12 を見ると，2007 年の 12 月までは，各月の説明文の cos 類似度の総和は急激に増加しているが，それ以降はおおよそ 15000 程度の値をとっていることがわかる．

7. 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．

情報エントロピーの定義を援用して，毎月のタイトルと説明文のそれぞれの単語(名詞)の多様性を数値で表現した．情報エントロピーを用いて，月ごとの単語の多様性を算出し，それを時系列で折れ線グラフ表示した．その結果，一意な単語数（赤線）は緩やかに増加しているのに対し，エントロピーもおおよそ増加している．このことからは，タイトルと説明文それぞれの単語集合について，多様性が減少していると判断することはできない．

また，cos 類似度を用いて，投稿動画に付与されるタイトルと説明文についてそれぞれ毎月の単語集合の類似度を数値で表現した．cos 類似度を用いて，月ごとに単語群の cos 類似度の総和を算出し，それを時系列で折れ線グラフ表示した．その結果，再生回数上位 1000 個のタイトルの cos 類似度の総和は緩やかに増加している．このことから，共通要素が多くなり，非共通要素が少なくなっている２つの単語群の組み合わせが増加している．つまり，タイトルについて，類似している単語集合を持つ動画が増加している．それに対して，再生回数上位 1000 個の説明文の cos 類似度の総和は，2007 年 12 月までは急激に増加しているが，その後おおよそ一定の値をとっている．

cos 類似度から再生回数上位 1000 位の動画のタイトルの単語集合の多様性は徐々に減少しているが，説明文の単語集合の多様性は 2007 年 12 月以降あまり変化がないということがわかる．

今後は，単語群についてクラスタリングを行い，単語集合の偏りや全体の傾向を調査していきたい．また，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, そこから多様性を維持する閲覧手法を開発できると考えている．

謝辞　本研究は JSPS 科研費 15K00451 の助成を受けたものである．

参考文献

1) 上畑恭平, 伊東栄典: タグの類似度に着目した利用者投稿サイト動画の多様性分析, 信学技報, vol.115, no.381, AI2015-40, pp.83-88, 2015.

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

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

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

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

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

7) 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

8) 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).

9) Similarity and distance: http://wikiwiki.jp/cattail/?%CE%E0%BB%F7%C5%D9%A4%C8%B5%F7%CE%A5