A quantitative contents diversity analysis on a consumer generated media site

Kamihata, Kyohei
Graduate School of Information Science and Electrical Engineering, Kyushu University

Ito, Eisuke
Research Institute for Information Technology, Kyushu University

出版情報：21st International Symposium on Artificial Life and Robotics, pp.436-440, 2016-01-22. International Society of Artificial Life and Robotics
バージョン：
権利関係：

A quantitative contents diversity analysis on a consumer generated media site

Kyohei Kamihata[1], Eisuke Ito[2]

[1]Graduate school of ISEE, Kyushu University, Japan
[2]Research Institute for IT, Kyushu University, Japan

[1] kamihata.k.411@s.kyushu-u.ac.jp, [2] ito.eisuke.523@m.kyushu-u.ac.jp

Abstract: Recent years, CGM (Consumer Generated Media) services, such as YouTube and nicovideo, are growing into social contents communication media. A lot of movies are posted to a CGM site, and many users are viewing them every day. Some web news articles mentioned that the diversity of movie contents might decrease in CGM site. The authors believe that contents diversity is necessary to keep CGM site activity, and for cultural sustainability. The authors try to measure diversity of movie contents of a CGM site quantitatively. They propose two quantitative metrics of contents diversity, and apply them to a CGM contents in nicovideo.jp.

Keywords: CGM, Contents Diversity, Tag, Cosine Similarity, Information Entropy

1 INTRODUCTION

Recent years, CGM (Consumer Generated Media) services, such as YouTube and nicovideo, are growing into social contents communication media. A lot of movies are posted to a CGM site, and many users are viewing them every day. We have been focusing on the nicovideo.jp as the target of folksonomy based content searching [1]. We have studied not only movie CGM but also text novel CGM. We proposed a ranking technique based on the analysis of novel keywords [2], and the structure analysis of bipartite graph between user and contents [3].

Nowadays, some web news articles mentioned that the diversity of movie contents may decreases in CGM site. Mr. Kawakami, who is the president of Dwango Company, mentioned that page view popularity ranking might cause decreasing diversity of CGM contents [4]. We believe that contents diversity is necessary to keep CGM site activity, and for cultural sustainability.

We try to measure diversity of movie contents of a CGM site quantitatively. In this paper, we propose two quantitative metrics to measure contents diversity. We apply these metrics to a CGM contents in nicovideo.jp, and evaluate our proposed metrics.

The rest of this paper is organized as follows. In section 2, we describe the dataset of nicovideo.jp briefly. Section 3 shows basic statistical analyses. In section 4, we introduce an information entropy based quantitative metrics to measure contents diversity. Section 5 describes cosine similarity based contents diversity. The time series trend of the metrics indicates decreasing of contents diversity quantitatively. Finally, we conclude this paper in section 6.

2 NICONICO DATASET

2.1 Nicovideo.jp

Nicovideo.jp (*Nico-Nico Douga*, in Japanese) is a popular movie sharing CGM in Japan. Nicovideo started on December 12, 2006. At December 2014, more than 11 million movies are posted to the site. The number of registered members is over 30 million, and the premium membership account is over 2 million at June 2013 [5].

2.2 Niconico dataset

Dwango company, which operates nicovideo.jp, offered the niconico dataset for researchers. The dataset is provided by NII (National Institute of Informatics) of Japan [6]. Researcher can get the dataset from the IDR (informatics data repository) of NII. Some other dataset is also provided at IDR.

Niconico dataset includes not only movie meatadata, but also nicovideo comments for each movie, and the niconico encyclopedia data. In this paper, we only used movie metadata. The number of records is indicated in Table 1. Table 2 shows attributes of a record in the dataset.

Table 1 Outline of niconico dataset

Items	Contents
Term	Mar., 2007 – Nov. 2012
Data form	JSON
Records (movie metadata)	8,305,696
Unique tags	5,328,341

Table 2 Attributes of movie metadata in niconico dataset

Attribute	Description
video_id	ID of a movie.
title	Title of the movie.
description	Description of the movie given by creator (author).
upload_time	Upload time of the movie.
length	Length of the movie.
movie_type	File format of the movie.
view_counter	Total views of the movie.
comment_counter	Number of posted comments to the movie.
tags	Tags of the movie given by viewers.

3 FREQUENCY ANALYSIS

3.1 Monthly movie uploads

Figure 1 shows the number of movie uploads in each month. The number of movie uploading is increasing for the most period. During 2012, The number of uploads in a month is over 180,000.

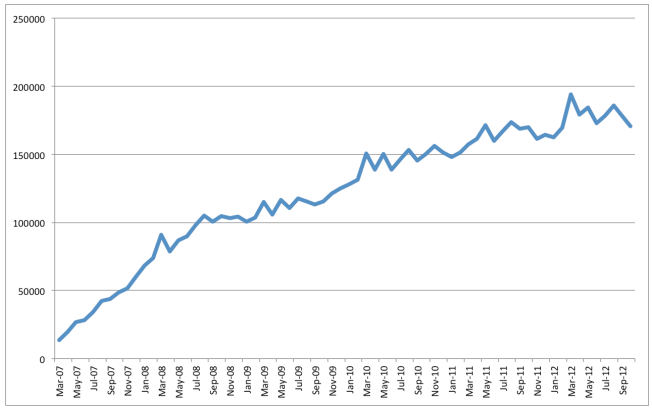

Figure 1 Monthly movie uploads

3.2. Number of unique tags

Viewer can give tags and comments to a movie, and the viewer can add the movie into his/her favorite list in nicovideo.jp. We extracted tags from niconico dataset, and count the number of unique tags in each month. Figure 2 shows the number of unique tags of each month.

The number of unique tags increased suddenly since service start until March 2008. After that, there is some fluctuation, the number of unique tags keep about 1.8 million.

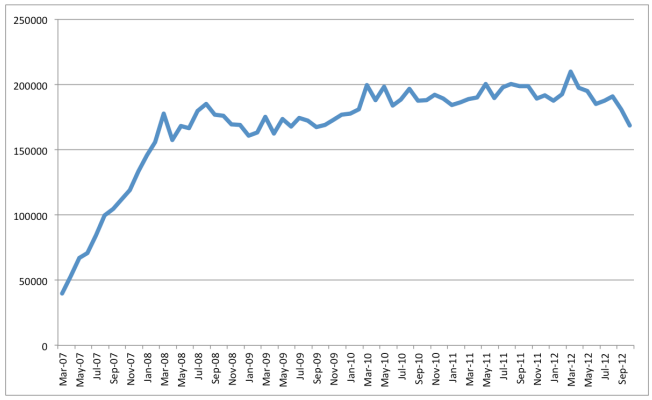

Figure 2 Monthly unique tags

3.3. Rank of the total view count of movies - frequency

Next, we investigate the distribution of view count of movies. Figure 3 shows a scatter of rank and view counts. A dot corresponds a movie, and horizontal axis is the rank , and the vertical axis is the view counts of the movie. Both axes are in log scale.

The top part (left upper part) looks like a straight line in figure 3. It may be distributed near power law because it is a straight line in both log scale graph. However the bottom part (right down part) is not straight line.

We plot another scatter chart in figure 4 and 5. In figure 4 and 5, horizontal axis is view count, and vertical axis is the number of movies, which correspond to the view count. The curve in figure 5 draws a normal distribution curve. Therefore, we find out that the view count distribution of nicovideo.jp movies is close to the lognormal distribution.

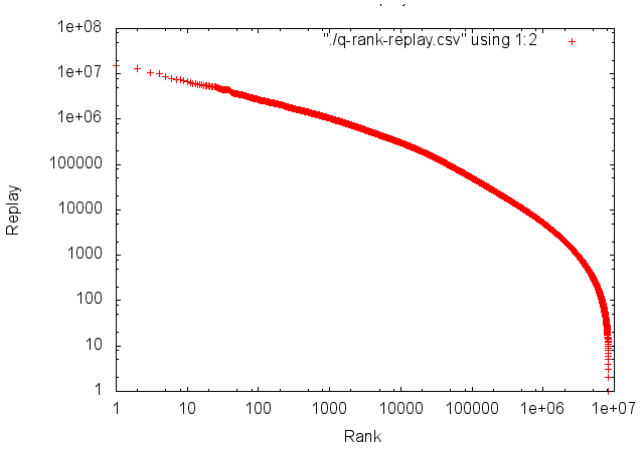

Figure 3 Movie's rank – frequency (view counts)

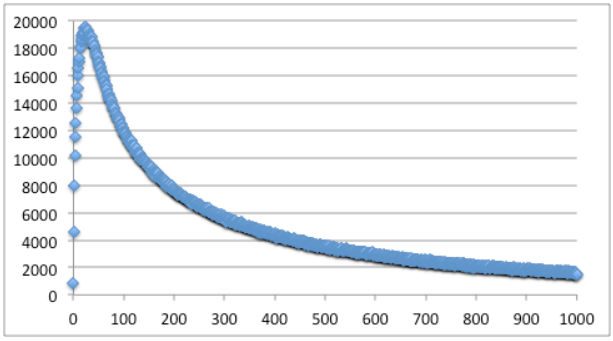

Figure 4 View count and corresponding movies. (View count < 1,000)

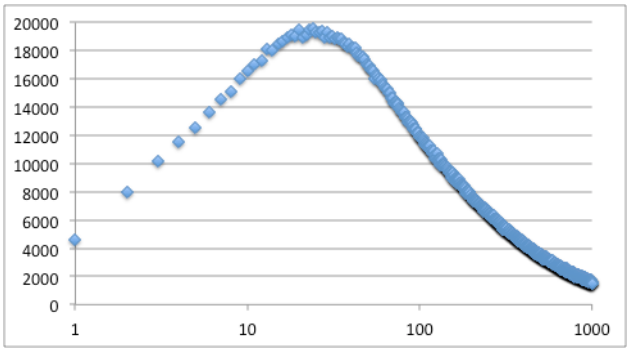

Figure 5 View count and corresponding movies.
(View count < 1,000 and log scale in horizontal axis)

3.4 Tag's rank - frequency

As mentioned in section 3.2, viewer can give tags and comments to a movie in nicovideo.jp. Figure 6 shows the tag part of nicovideo.jp movie viewer.

Figure 6 Tags of a movie (ID: sm15630734).

We extract tags from each movie metadata, and count frequency of each tag. Figure 7 shows rank frequency plot of tags. Both axes are in log scale.

Figure 7 illustrates a straight line in both log scale, therefore, the distribution of tag frequency follows the power-law distribution. We know that the word frequency in natural language documents follows the power-law distribution. Then, tags distribution is similar to natural language distribution.

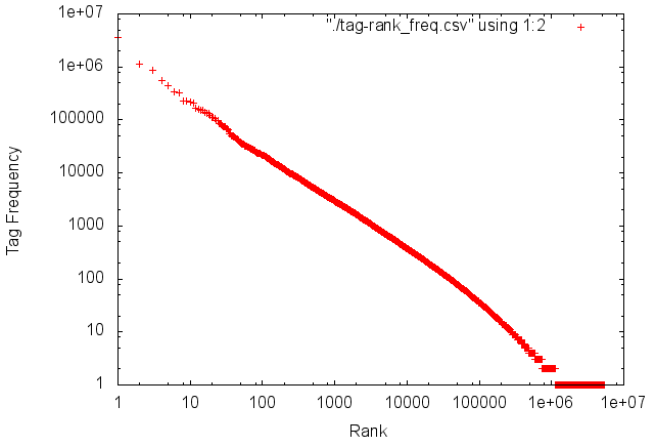

Figure 7 Tag's Frequency-Ranking. (both log scale)

4 ENTROPY BASED DIVERSITY

As we mentioned in section 1, an interview [4] pointed diversity decrease of contents in a CGM site. To investigate whether the diversity decreases or not, it is a quantitative metric is necessary. We use the following symbol for expressions.

D	:	A document (content) set,
n	:	The number of documents ($\|D\| = n$),
W	:	Tag set,
df(w)	:	Document frequency of tag w.

4.1 Basic idea of entropy based contents diversity

Before definition of quantitative metrics, let us consider two extreme cases. Let n be the number of contents (the number of documents), and be df(w) is the documentary frequency of word w. If contents are perfectly uniform, all contents are same. In this perfect uniform case, the same tags will be given to all contents, and then, df(w) will be n for all tag w.

Next, let us consider the opposite extreme case. If all contents were perfectly diverse, there is no similarity between any two contents. In this perfect diverse case, a tag will be given to only one content, and then, df(w) will be 1 for all tag w.

Actuary, df(w) of a tag w is between two extreme cases. Tags, which are used as genre or category, are frequently appeared, and df of those tags are high. Tags, which represent creator nickname or content name, are appeared a few times, then df of those tags are low.

As shown in figure 6, frequency of most tags is 1. It isn't good to decide only the stag, which is low frequency because frequency of tags, which are less than 5, is most.

4.2 Definition of tag entropy

Shannon estimated the entropy of real English documents [7]. He applied the information entropy to the words of document. The information entropy is calculated by the expression (1). The unit of $H(W)$ is bit/word, if the bottom of a logarithm is 2.

$$H(W) = -\sum_{w \in W} p(w)\log(p(w)) \qquad (1)$$

We applied it to calculate the entropy of tags. In expression (1), p(w) is the appearance probability of word w. In nicovideo.jp, One tag can be given at most one time for one movie. Therefore, p(w) of a tag w is df(w)/n.

4.3 Entropy trend of the tags

Let D_m be the movie set, and all movie in D_m is posted a specific month m in a year. We made the tag set W_m from D_m, and calculated $H(W_m)$ of tags in W_m according to (1). Blue line in figure 7 shows the monthly trend of tag entropy. Red line in figure 7 is the number of the unique tags in each month.

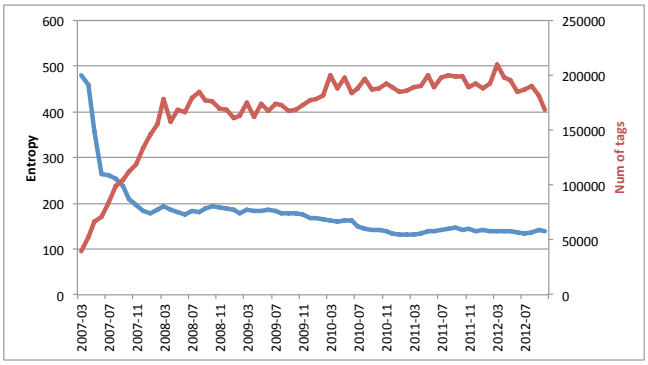

Figure 8 Entropy of tags (blue) and the number of the unique tags (red)

As shown in figure 1, the number of uploading movies increases, and the number unique tags gradually increasing. However, the entropy of tags decreases. This indicates that diversity of nicovideo contents may be decreased as time has passed.

5 SIMILARITY BASED DIVERSITY METRIC

Secondly, we propose a content diversity metrics using cosine similarity.

5.1 Basic idea of entropy based contents diversity

Figure 9 illustrates a model of contents diversity. A dot corresponds to a content. If contents are diverse, then distance between two contents will be long and similarity of the two contents will be small. On the other hand, if contents are not diverse, then distance of two contents will be close, and similarity of them are large.

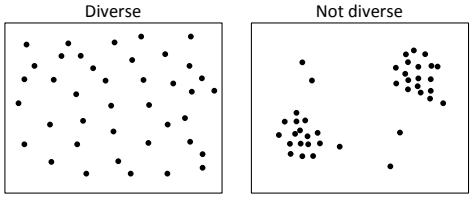

Figure 9 A model of diversity.

So, sum of all similarity of all document pairs may be a metrics diversity of contents set.

5.2 About the cosine similarity

There are some definitions for distances and similarity [8]. In this time, we use cosine similarity because it is basic similarity. The cosine similarity is one of the methods to measure the similarity between two vectors.

In the document-word matrix model, a document is expressed as a word vector. Figure 10 illustrates the idea of document matrix. Usually, M(i,j) is term frequency of word w_i in document d_j.

document

word	d_1	d_2	…	d_j	…	d_m
w_1	1	1		0		1
w_2	0	1		1		0
:						
w_i	1	1		$M(i,j)$		1
:						
w_n	1	0		0		1

Figure 10 Document word matrix.

The range of cosine is from -1 to 1, normally. In case of document-word matrix, the range of cosine similarity for every two documents or every two words, is 0 to 1. Because, all elements in the matrix are non-negative integer. Cosine similarity is calculated by expression (3).

$$\cos(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| * |d_j|} = \frac{\sum_{k=1}^{m} (M(i,k) * M(j,k))}{\sqrt{\sum_k (M(i,k))^2} * \sqrt{\sum_k (M(j,k))^2}} \quad (2)$$

In nicovideo.jp, one tag can be given at most one time for one movie. Then, an element $M(i,j)$ must be 0 or 1. Therefore, the inner product of d_i and d_j will be the number of tags, which are commonly appeared in both d_i and d_j. The value of $|d_i|$ will be the square root of the number of tags. Consequently, it is easy to calculate cosine similarity.

5.6 TREND OF SUMCOS

We calculated sum of cosine similarity (SumCos, for short) of all document pairs, according to the expression (3). For document set D, the number of pairs is $n(n-1)/2$, where n is the number of documents.

$$SumCos(D) = \sum_{i,j} (\cos(d_i, d_j)) \quad (3)$$

As shown in figure 1, the number of uploading documents (movies) is different. For normalization, we select 1,000 documents (n=1000) for each month.

Firstly, we select top 1,000 ranking movies by number of replay. Figure 11 illustrates SumCos of top 1,000 movies' tags. Secondly, we selected 1,000 movies randomly. Random selection was repeated 10 times, and calculated the average of SumCos. Figure 12 is the trend of random case.

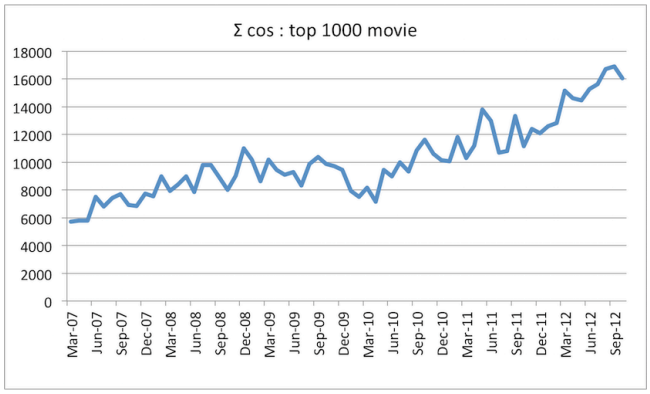

Figure 11. CosSim trend (Top 1000).

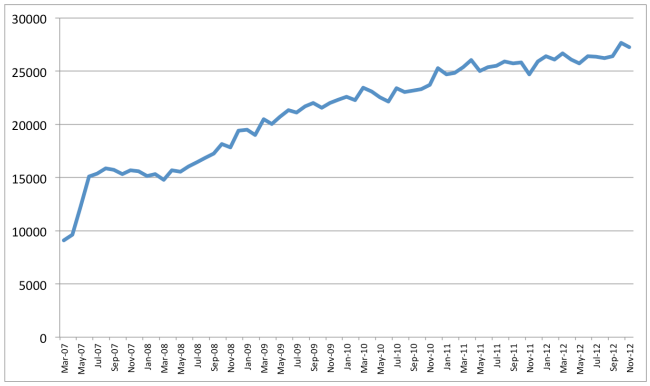

Figure 12. CosSim Trend
(Average of 10 times random 1,000 movie selection)

We find out that CosSim (the sum of the cosine similarity) gradually increases in both figure 11 and figure 12. SumCos trends indicate that diversity of nicovideo contents may be decreased as time has passed.

6 CONCLUSION

In recent years, some people mentioned decreasing of diversity of CGM contents. In this paper, we proposed two quantitative metrics for contents diversity. One is entropy-based diversity, and the other is sum of cosine similarity. We applied them to the set of metadata nicovideo.jp. The entropy of tags is decreasing every month, and sum of the cosine similarity is increased. These may be quantitative evidences of decreasing contents diversity in nicovideo.jp.

In the future, we would like to check diversity trend of other contents such as online novels and academic papers. We want to consider user's behavior model.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15K00451.

Reference

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

[2] Eisuke Ito, Kazunori Shimizu: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012, pp.531-536, 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, Nobuo Kawakami: Nobuo Kawakami's mind, https://cakes.mu/posts/5036 (accessed at Dec.12, 2014).

[5] Nicovideo (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

[6] National Institute of Informatics, DWANDO: NicoNico dataset: http://www.nii.ac.jp/cscenter/idr/nico/nico.html, (accessed at Oct.24, 2014).

[7] C.E. Shannon, Prediction and Entropy of Printed English, Bell System Technical Journal, Volume 30, Issue 1, pp. 50-64, 1951.

[8] Similarity and distance: http://wikiwiki.jp/cattail/?%CE%E0%BB%F7%C5%D9%A4%C8%B5%F7%CE%A5 (accessed at Oct.24, 2014)