Keyword diversity trend of consumer generated novels

Ito, Eisuke
Research Institute for Information Technology, Kyushu University

Honda, Yuya
Department of Library Science, Graduate School of Integrated Frontier Sciences, Kyushu UNiversity

出版情報：2017-07-31. 電気学会
バージョン：
権利関係：

Keyword diversity trend of consumer generated novels

Eisuke Ito
Research Institute for IT.
Yuya Honda
Grad. School of Library Science
Kyushu University
Motoka 744, Nishi-ku, Fukuoka, 819-0345, Japan.
{ito.eisuke.523@m, honda.yuya.128@s}.kyushu-u.ac.jp

ABSTRACT

Recent years, CGM (Consumer Generated Media), such as YouTube and nicovideo.jp for movies, syosetu.com for novel stories, become very popular. A lot of contents are posted to CGM sites every day, and also a large number of users are enjoying posted contents. At present, some articles mentioned decreasing diversity of contents. Some posted new content may be similar with previous posted contents. The authors are afraid that decreasing diversity of contents causes less energetic cultural activity. In this paper, the authors proposed two quantitative metrics of contents diversity, and applied them to the contents in syosetu.com. They focused the keywords which are given to the novel by the novel author, and calculated entropy and similarity of keywords. As the results, they observed increase of similarity, and it shows decrease of diversity of contents.

KEYWORDS

Big data analysis, CGM, word frequency, contents diversity, document term matrix, cos similarity.

1 INTRODUCTION

Recent years, CGM (Consumer Generated Media) services, such as YouTube and nicovideo.jp, are growing into social contents communication media. Not only movies but alto online novels are also popular. A lot of novels are posted to syosetu.com, and many users are reading and enjoying them every day. There are some major novel CGM sites such as syosetu.com, estar.jp, pixiv.jp and comico.jp. In this research, we focus on ”syosetu.com”, which is the most popular novel CGM site in Japan. There are more than 450 thousand online novels in syoteu.com as of Feb.2017. The number of contents and viewers are increasing. Most of online novels are written by amateur writers and might not be good quality. However, there are some high-quality novels and those popular high-quality novels are published as paper books. A few very popular novels may also become manga and animation.

We have been focused on syosetu.com as a research target of CGM contents search and recommendation. We proposed a ranking methods based on the analysis of novel keywords in [7], and reported the structure analysis of bipartite graph between user and contents in [2].

Nowadays, some blog articles mentioned that the diversity of novels may decreases in syotetu.com. Mr. Kawakami, who is the president of Dwango Company, mentioned that page view popularity ranking might cause decreasing diversity of CGM contents [3]. We believe that contents diversity is necessary to keep CGM site activity, and for cultural sustainability.

In this paper, we propose two quantitative metrics of contents diversity. One is entropy based metrics, and the other one is similarity based metrics. Especially, we focused on words in keyword field. Keywords are given to the novel by the novel author. We applied our metrics to keywords of novels, and evaluate our proposed metrics.

The rest of this paper is organized as follows. Section 2 shows basic statistics of novels and words on syosetu.com briefly. In section 3, we introduce an information entropy based quantitative metrics to measure contents diversity. Section 4 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 5.

2 STATISTICS OF SYOSETU.COM

This section shows some statistics of novels and words on syosetu.com.

In this section, we explain the organization of the site syosetu.com, the number of novels and the number writers. We also describe a basic statistics concerning to the distribution of frequencies of words which were assigned to novels by users.

2.1 Novel metadata

Figure 1 illustrates a model of data structure and data flow of a novel in syosetu.com. When an author posts a novel to syosetu.com, the author can set the title, the author's name, genre, keywords and outline text as metadata. Author has to select a genre word from 15 genre words provided by the site. The author can give arbitrary words for keywords. Each novel is automatically assigned a novel ID, and novel ID is called as "Ncode" in syosetu.com.

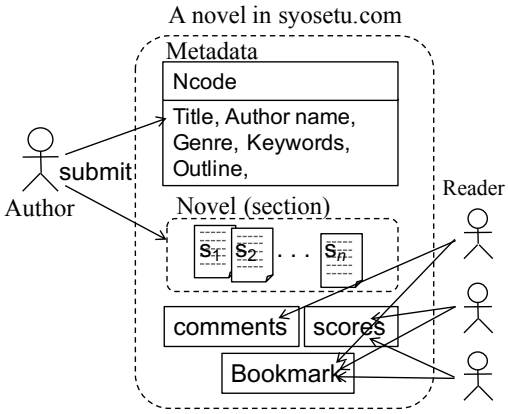

Figure 1. Metadata of syosetu.com

A novel in syosetu.com is classified into short novel or series novel. A serial novel consists of several sections, a short novel has one section. A novel of syosetu.com is also classified into completed novel or not. Short novel is certainly completed novel. Figure 2 shows the metadata page and the TOC (table of contents) page of a novel "Knight's & Magic" (Ncode: `n3556o`). This is a serial novel, and the TOC page shows section titles.

2.2 Novel metadata crawling

Web API named "Narou API" [4] is provided by syosetu.com. Using Web API, anyone can get novel metadata and author's bibliography (list of novels). Returned data is written in YAML format, and YAML is a format of structured data.

If you specify an Ncode in API, then you will get YAML format metadata of the novel. But we didn't use Ncode for metadata crawling, because Ncode assignment rule is not clear. On the other hand, author ID is a numerical number and assignment rule is very clear, that is serial number. Author ID (User ID) is assigned to each user in syosetu.com.

We created author's bibliography data crawler by Ruby language. The crawler increases the author ID number from 1 to 450,000, and gathers bibliography of the author ID. Since we checked in advance that the highest author ID may be about 400,000, we set 450,000 as upper limit of author ID to the crawler.

In order to efficiently crawling of 450,000 bibliography, we set 8 virtual machines for distributed parallel metadata crawling. We ran 8 crawlers since Nov. 2015 until Dec.2015. After crawling, we get 81,449 valid data. In other words, there are 81,449 authors in syosetu.com at Dec.2015.

Next, we created an extractor program. It extract each novel metadata (identified by Ncode) from 81,449 crawled author's bibliography data. Finally, we obtained 232,096 novel metadata. Table 1 shows the summary of data obtained by crawling and extraction. Table 2 shows a part of attributes of novel metadata.

Table 1. Crawled Metadata

Item	Description
Priod	Apr.2004–Oct.2015.
Novel	232,096
Writer	81,449

2.3 Novel posting trend

Figure 3 shows the number of monthly new novel posts. Until March 2014, the number of

Table 2. Attributes of metadata

Attribute	Description
ncode	Identifier of a novel
title	Novel title
story	Story outline of the novel
writer	Author name
keyword	Keyword(s) given by the author
genre	Genre word
writer	Author ID (User ID)
general_fistup	The first upload date

new novel posting is increase. Peak is March 2014. 5,306 new novel posted at March 2014. After that, new novel posting decreased, but more than 2,000 novels are newly posted to syosetu.com in a month.

2.4 Keyword trend

We counted the number of words in keyword field. Figure 4 shows monthly trend of the number of keywords. Blue line is total word count, and red line shows the number of unique words. As same as the number of new novel posting, the number of words is increase until March 2014. Peak is also March 2014. There are 7,149 unique words at March 2014. After that, the number of words decreased.

2.5 Keyword rank-frequency

We counted term frequency of each keyword. Figure 7 shows rank frequency plot of keywords. 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.

3 ENTROPY BASED DIVERSITY

As we mentioned in section 1, some articles [4] mentioned diversity decrease of novels in syosetu.com. To investigate whether the diversity decreases or not, quantitative metric is necessary.
We use the following symbols for expressions.

D : a document (content) set,
n : the number of documents ($|D| = n$),
W : word set,
$df(w)$: document frequency of word w ($w \in W$).

3.1 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 $df(w)$ be the documentary frequency of word w. If documents are perfectly uniform, all document are same. In this perfect uniform case, the same words will be given to all documents, and then, $df(w)$ will be n for all 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 word will be given to only one document, and then, $df(w)$ will be 1 for all w.
Actuary, $df(w)$ of a word w is between two extreme cases. Keywords, which are used as genre or category, are frequently appeared, and df of those words are high. Words, which represent creator nickname or content name, are appeared a few times, then df of those words are low.

3.2 Definition of tag entropy

Shannon estimated the entropy of real English documents[6]. He applied the information entropy to the words of document. The information entropy is calculated by the following expression (1). The unit of $H(W)$ is bit/word, if the bottom of a logarithm is 2.

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

In expression (1), $p(w)$ is the appearance probability of word w. In syotetu.com, one word can be given at most one time for one novel. Therefore, $p(w)$ of a keyword w is $df(w)/n$.

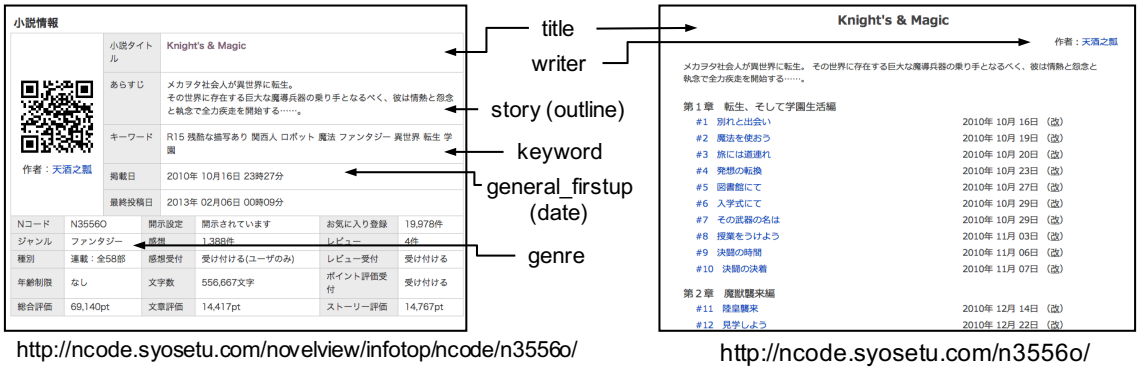

Figure 2. Novel TOC and metadata page (ID:n3556o)

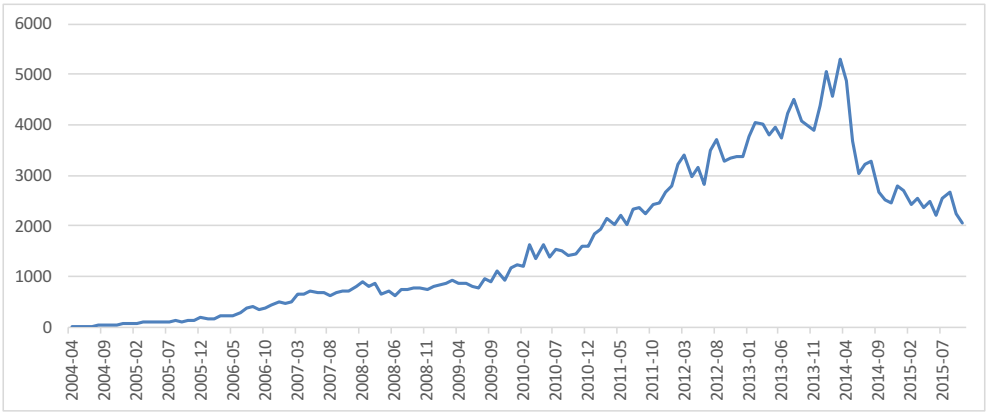

Figure 3. Number of posted novel (monthly)

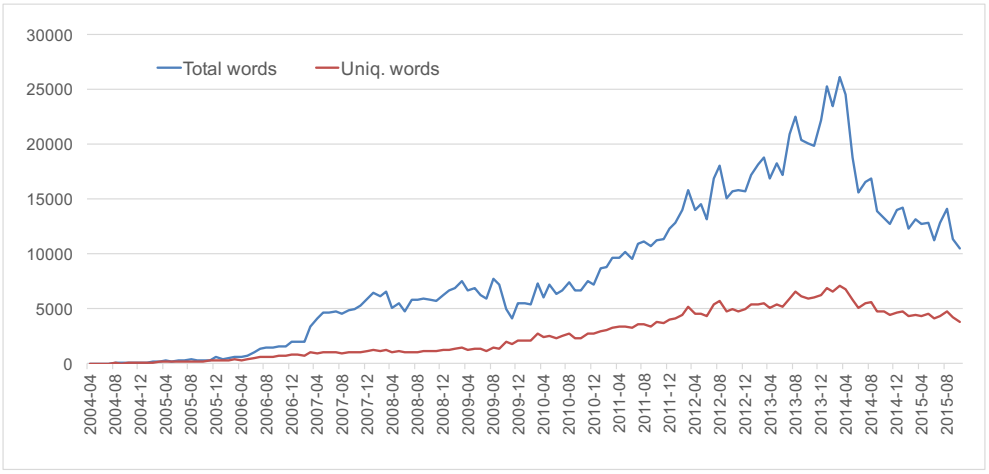

Figure 4. Number of (unique) words in keyword field (monthly)

3.3 Entropy trend of the keywords

Let D_m be the novel metadata document set, and all novel in D_m is posted at a specific month m of a year. For example, $D_{2010-01}$ is the novel metadata set, and all novel $d \in D_{2010-01}$ is posted at January 2010. We made the keyword set W_m extracted from D_m, and calculated $H(W_m)$ of keywords according to expression (1). Figure 5 shows the monthly trend of keyword entropy.

In Figure5, except during April 2017 to September 2009, the entropy of keywords gradually increases, and saturates about 10 bit/word. As shown in Figure 3 and Figure 4, the number of new novel posting and the number of (unique) keywords is increase until March 2014. After that, new novel posting and the number of (unique) keywords gradually decreased. Figure5 shows that the number of documents are not related entropy of word.

This result is similar with Shannon's result[6]. In [6], entropy of word is saturated about 11.8 bit/word for English documents.

During April 2017 to September 2009, entropy of keywords dented. This phenomenon will be mentioned in 4.4.

4 SIMILARITY BASED DIVERSITY

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

4.1 Basic idea of entropy based contents diversity

Figure 6 illustrates a model of contents diversity. One dot in Figure 6 corresponds to a document. If documents are diverse, then distance between two documents will be long and similarity of the two documents 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.

There are some definitions for distances and similarity, such as Euclidean distance, Manhattan distance, and so on for distance, cosine similarity, Pearson's correlation, Jaccard/Dice/Simpson coefficient for similarity. In this study, we decided to use cosine similarity because it is most used as index of similarity. So, sum of all similarity of all document pairs may be a diversity metrics for document set.

4.2 Definition of $SumCos$

We use term-document matrix to vectorize a document. In the term-document matrix, a document is expressed as a word vector. The model is also known as "Bag of Words" model. Figure 7 illustrates an example of term-document matrix M. Usually, $M_{i,k}$ element is term frequency of word w_k in document d_i.

Cosine similarity of document i and j is calculated by expression (2).

$$cos(i,j) = \frac{\sum_k M_{i,k} * M_{j,k}}{\sqrt{\sum_k M_{i,k}^2} * \sqrt{\sum_k M_{j,k}^2}} \qquad (2)$$

The range of cosine is from -1 to 1, normally. In case of term-document 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.

In case of syosetu.com, one keyword can be given at most one time for one novel. Then, an element $M_{i,k}$ must be 0 or 1 for keywords. Consequently, it is easy to calculate cosine similarity.

We calculated sum of cosine similarity ($SumCos$, for short) of all document pairs, according to the expression (3).

$$SumCos(D) = \sum_{i=1}^{n-1} \sum_{j=i}^{n} cos(i,j) \qquad (3)$$

For document set D, the number of pairs is ${}_nC_2 = n(n-1)/2$, where n is the number of documents ($n = |D|$). As shown in Figure 3, the number of posted documents (novels) in a month is different. If n is large, then the number of document pairs is more large, and $SumCos$ become larger value.

To modify the effect of the number of document, we propose $SumCos_{ave}$ in expression (4).

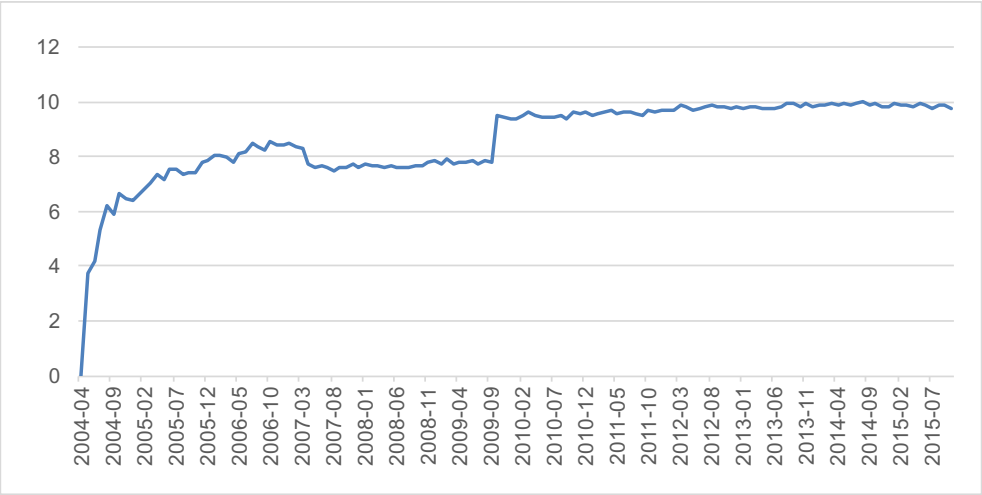

Figure 5. Entropy of keywords (monthly)

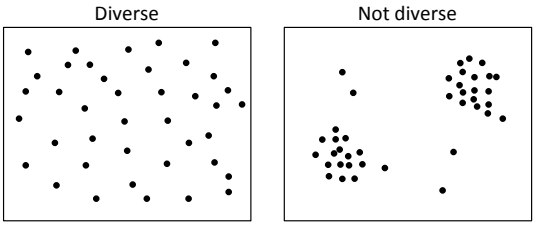

Figure 6. A model of diversity

word (term)

document	w_1	w_2	...	w_k	...	w_m
d_1	2	5		0		1
d_1	0	1		2		5
:				:		
d_i	1	1	...	M_{ik}	...	2
:				:		
d_n	1	0		0		10

Figure 7. Term-Document Matrix

$$SumCos_{ave}(D) = \frac{SumCos(D)}{{}_nC_2} \quad (4)$$

4.3 Monthly trend of $SumCos$

Let D_m be the novel metadata document set of month m of a year. According to the expression (4), we calculate $SumCos_{ave}(D_m)$ for each month. Figure 8 shows monthly trend of $SumCos_ave$ of keywords.

Except during April 2017 to September 2009, Figure8 shows that $SumCos_{ave}$ gradually increases. It indicates that the number of words commonly appearing in the keyword field of the novel has increased. This also indicates that the diversity of the novel keywords is decreasing.

4.4 Exception

During April 2017 to September 2009, entropy of keywords dented in Figure 5, and $SumCos$ of keywords is bumped in Figure 8. This period is exception of keyword data. We calculated ave_m with $ave_m = |W_m|/|D_m|$, where ave_m is average of the number of keywords per novel in month m. Figure 9 shows the monthly trend of ave_m.

Figure 9 obviously shows that ave_m is high during April 2017 to September 2009. Actually, keywords are increased in this exception period.

5 RELATED WORK

There are very few contents diversity trend analysis. Igami and Saka proposed a scientific paper mapping method called ”science map” [5] and investigate effectiveness of scientific policy of Japan. Science map is based on citation relations. A few top cited papers are core papers. The paper citing a core paper is a descendant paper. The descendant papers surround the core paper, and forms an island centered on the core. They reported evidence

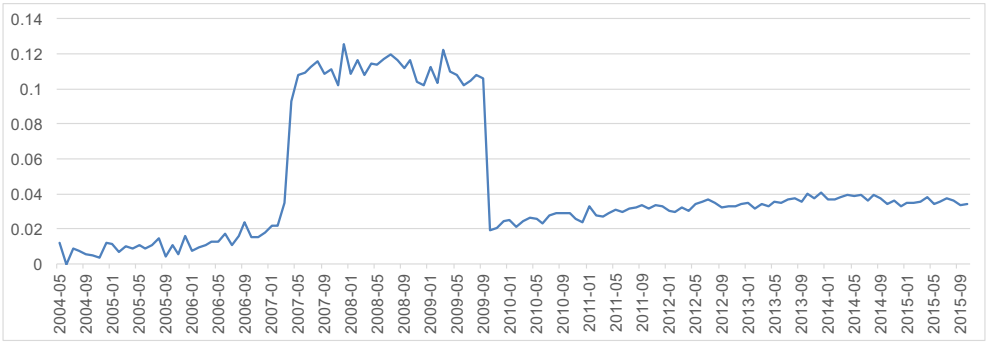

Figure 8. Average $SumCos_{ave}$ of keyword (monthly)

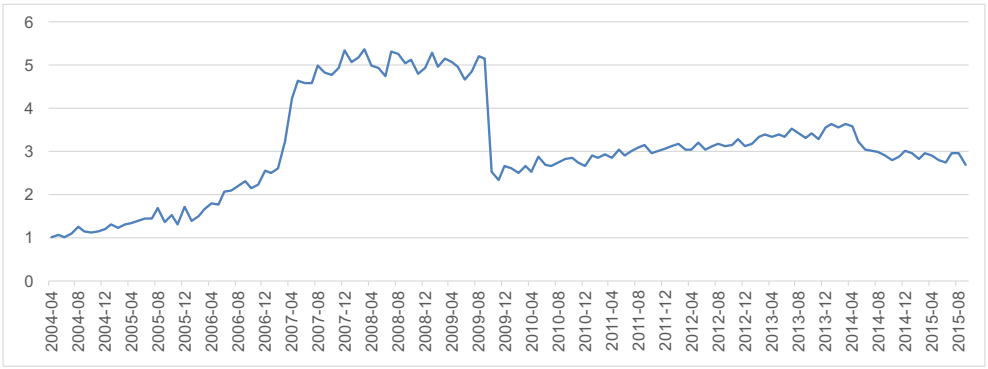

Figure 9. Average of the number of keywords (monthly)

of decreasing diversity of scientific article created by Japanese researchers[1]. Science map method is difficult to apply CGM contents. Because there are only few obvious direct links between novels.

6 CONCLUSION

Online novel become very popular 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 one is the sum of cosine similarity ($SumCos$). We applied proposed metrics to the set of syosetu.com novel metadata. The entropy of keywords is saturated about 10 bit/word. Word entropy is not suitable contents diversity metric.

The average of sum of the cosine similarity ($SumCos_{ave}$) is increased. It indicates increase the number of words commonly appearing in the keyword field of the novel. These may be quantitative evidences of decreasing contents diversity in syosetu.com.

In the future, we want to investigate trend of $SumCos$ not only for the keywords of the novel but also for the outline and the title. We want to investigate $SumCos$ trend of subsets for each genre, whether $SumCos$ trends are different or same. We also want to check diversity trend of other CGM contents such as movies or comic (cartoon). Finally, we want to establish user's contents selection model.

ACKNOWLEDGMENT

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

REFERENCES

[1] Igami, M., Saka, A.: Decreasing diversity in japanese science, evidence from in-depth analyses of science maps. Scientometrics 106(1), 383–403 (January 2016)

[2] Ito, E., Shimizu, K.: Frequency and link analysis of online novels toward social contents ranking. In: Proc. of SCA2012 (The 2nd International Conference on Social Computing and its Applications). pp. 531–536. IEEE (November 2012)

[3] Kawakami, N.: Nico Nico Philosophy. Nikkei BP (November 2014)

[4] Narou-Developer: Narou api. http://dev.syosetu.com/man/api/

[5] Saka, A., Igami, M.: Science map 2010&2012 (study on hot research area (2005 – 2010 and 2007-2012) by bibliometric method). Tech. Rep. 159, NISTAP (July 2014)

[6] Shannon, C.E.: Prediction and entropy of printed english. Bell System Technical Journal, 30(1), 50–64 (1951)

[7] Shimizu, K., Ito, E., Hirokawa, S.: Predicting future ranking of online novels based on collective intelligence. In: Proc. of ICDIPC2013 (The Third Int'l Conf on Digital Info. Processing and Communications). pp. 263–274. IEEE (January 2013)