Frequency and link analysis of online novels toward social contents ranking

Ito, Eisuke
Research Institute for Information Technology | Graduate school of ISEE

Shimizu, Kazunori
Graduate school of ISEE | Research Institute for Information Technology

出版情報：Proceedings of The 2nd International Conference on Social Computing and its Applications (SCA2012), pp.531-536, 2012-11-01. IEEE
バージョン：
権利関係：

Frequency and link analysis of online novels toward social contents ranking

Eisuke Ito
Research Institute for Information Technology
Kyushu University
Fukuoka, Japan
ito.eisuke.523@m.kyushu-u.ac.jp

Kazunori Shimizu
Graduate School of ISEE
Kyushu University
Fukuoka, Japan
2IE11061G@s.kyushu-u.ac.jp

***Abstract*—User generated contents service is a social service because users not only enjoy contents but also provide feedback on content by commenting, tagging and bookmarking. The authors are interested in social service ranking and categorization methods for online novels. Many readers comment and bookmark there favorite novels in online novel services. Comments and bookmarking are facilitated by readers, and it is possible to use the data as a resource for social ranking and recommendation. In this paper, we focused on an online novel service, and analyzed the frequency of keywords, number of authors, and links from readers to novels. Although the bipartite graph between readers and novels fulfills growth and preferential attachment conditions of scale-free network, distribution of links does not follow power law. The basic idea of social ranking using extracted bookmarked novels and favorite authors is also explored.**

Keywords-online novel; user comments; link structure analysis; social content recommendation

I. INTRODUCTION

User generated contents service is a social service. At present, youtube.com, China's youku.com, and Japan's nicovideo.jp are very popular, and many people create accounts on these services to post content(s), and massive amounts of viewers watch the posted contents. User generated online novel services, which is our research target, have become popular in recent years. For example, qidian.com of China and syosetu.com of Japan have become popular. The word in the address *syosetu* is a Japanese word that means a *novel*. Viewers not only watch (read) contents, but also engage in various social activities such as commenting, bookmarking, and tagging there favorite contents.

A search and recommendation engine plays an important role for finding good contents, as there is too much contents in which to find good contents. There are two important functions for contents search and recommendation engines. One is to measure the quality, and the other one is to categorize contents. In order to realize these two functions, we are developing search and recommendation methods using collective intelligence such as social comments and links given by many users.

We have been studying ranking (quality measurement) methods of movies in nicovideo.jp based on comments given by viewers [1], and also automatic categorization methods based on tags given by viewers [2]. The comments and tags are good resources for social ranking and categorization because they include the viewer's impression and knowledge. We also studied recommendation methods from scholarly papers using co-occurrence of contents access [3].

In this paper, our target is the online novel site syosetu.com of Japan. On the website syosetu.com, there is feedback from readers to the author of the novels. Readers can bookmark his/her favorite novels and authors on the site, and there is a notification service to alert readers if a bookmarked novel is updated or a brand-new novel is written by a favorite author. Moreover, readers can post comments on a novel, if the author permits commenting. Comments are not only encouragement from reader to the author, but also provide suggestions for improvement (pointing out typos, etc.).

A simple popularity ranking is already provided by syosetu.com, but it is not so useful. There are two problems for simple popularity ranking. The first one is accumulated value. Popularity is measured by the number of viewers, page views, bookmarked users, and scores given by reader. These are accumulated values, and old good novels will always stay in the top rank. To address this problem, syosetu.com has proposed periods of limited accumulated value such as, today, this week, this month, and this quarter. The second problem is that the readers' preference and capability are not reflected in the score. Reader's evaluation score may differ depending on the readers preferred genre, or past reading experience.

We have studied novel quality measurement methods based on the link structure of bookmarked favorite novels. A bookmarked novel expresses the reader's preference and support of the novel, and therefore, bookmarking data is a good mining resource to mine the hidden social knowledge of novel evaluation. Bookmarking can represent a bipartite graph between readers and novels. In this paper, we analyze the frequency and link structures of online novels in syosetu.com in the context of social ranking and recommendation.

The composition of this paper is as follows. In section 2, we describe related work. Section 3 describes the result of basic analysis of syosetu.com, and shows some basic frequency analysis. In Section 4, we describe the link structure of the bookmarking of novels. Additionally, we proposed a novel evaluation method based on the link structure briefly. Finally, we conclude our paper in section 6.

II. Related Work

Ido Guy and others [5] proposed a social recommendation system based on social media, such as SNS. They used the relation between items, persons, and tags. In syosetu.com, authors and readers are given ID numbers, and all novels, comments and bookmarking data can identify who post it because the user ID is attached to the data. Therefore, it is possible to apply Guy's technique for novel recommendation.

A lot of users may assign tags to many items, and these tags could be a good mining resource, but most of them are noise. To filter out the noise tags, H. Liang and others [6] proposed a weighting technique for determining noise tags based on the relation between the tag and the item. Their techniques are also applicable for online novel search and recommendation.

Using conventional collaborative filtering has its own problems in that too many already known items are recommended. Hijikata and others [7] proposed the concept of novelty as a measurement, which recommends new items. They also proposed and evaluated three novelty based recommendation algorithms. Their novelty concept will be required for online novel search and recommendation systems.

III. Basic Data of Target Site

In this section, we describe the data structure of the online novel site "syosetu.com", show the number of novels and authors, and the frequency of keywords.

A. Structure of syosetu.com

"syosetu.com" is an online novel service provided by the *Hina-project* Company. Almost all the metadata (HTML pages) of novels from syosetu.com were crawled in April 2012. The scores of novels given by readers, and the readers bookmarks of favorite novels lists were also collected. Table I shows the number of published novels, authors who have written at least one novel, genre words, and unique keywords given for all novels.

Figure 1 shows an outline of the structure of data in syosetu.com. The author writes a novel, and then uploads it to the site. One novel can consists of a single or multiple sections. When there is only one section, it becomes a short novel. The author supplies the metadata for his/her novel, such as title, author name, genre, keywords, and a short synopsis. The author must select a genre from 18 genre words, which are specified by the service manager. The author can create the synopsis and keywords freely, only limited by the number of bytes allocated for the synopsis and keywords.

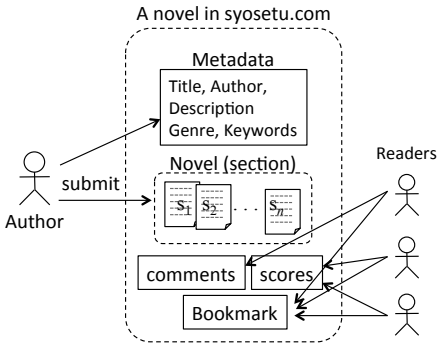

Figure 1. Data structure in syosetu.com

Anyone can read the novels on syosetu.com. If you have a syosetu.com account, it is possible to use convenient functions, such as bookmarking of favorite novels, notification of updates to favorite novels, and feedback to author. Registered user can also score novels, and send comments about a novel to the author.

Registered users are able to use mypage. Figure. 5 shows the structure of mypage in syosetu.com. Any reader is able to bookmark favorite novels, and register favorite authors in his/her mypage. The syosetu.com site notifies readers if a bookmarked novel is updated or the publication of a new novel that is written by a favorite author. If the reader opens his/her mypage, anybody can check the reader's preference.

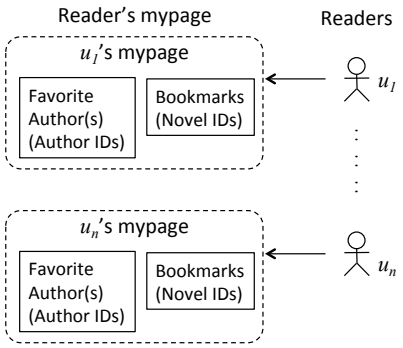

Figure 2. Mypage of syosetu.com

Table I shows the number of novels, readers, and authors. Authors are subset of readers, and an author writes at least one novel. Mypage readers are also subset of readers, and they open their own mypage.

TABLE I. Number of Novels, Readers and Authors

Item	Number
Novel	134,763
Reader	24,0730
Author	56,236
Mypage	92,488

B. *Keyword frequency*

We collected the metadata files of each novel and counted the frequency of words in the novel keyword field. There are 128,115 unique words. Table II shows the top 20 words and the occurrence frequency. Figure 2 shows a plot of the ranking and frequency of keywords. Both axes are on a logarithmic scale. The distribution of the frequency follows the power law or zipf's law.

TABLE II. TOP 20 KEYWORDS AND FREQUENCY

Rank	Keyword	Freq.
1	Cruel	27,696
2	romance	26,669
3	R15	21,718
4	modern	21,547
5	fantasy	20,247
6	high school	15,633
7	serious	12,900
8	tender	11,651
9	another world	11,433
10	youth	9,303
11	magic	8,673
12	girl	8,169
13	comedy	7,893
14	school	7,277
15	friendship	6,960
16	boy	6,696
17	campus	6,689
18	happy ending	6,685
19	literary	4,859
20	dark	4,742

Some high frequency words in Table II are caution words, which are specified by management side. In Table II, 1st "cruel" and 3rd "R15" are caution words. Reader can filter using caution words.

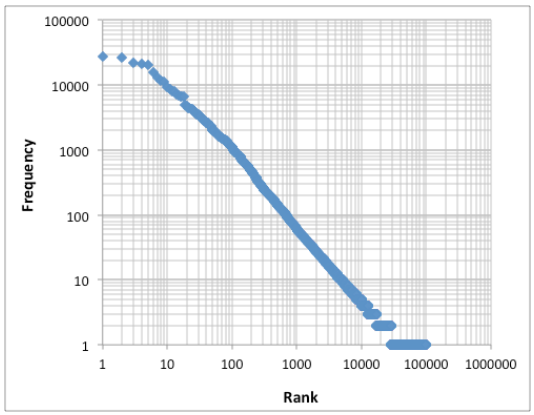

Figure 3. Rank-Frequency of keywords

Table III shows the number of low frequency words. Authors define a lot of low frequency words, with 77.7% words only appearing once. Low frequency words may be difficult to use as categorization words, or indexed words for search.

TABLE III. RATIO OF LOW FREQUENCY WORDS

Freq.	# of words	Ratio
1	99,483	77.7%
2	11,440	8.9%
3	4,760	3.7%
4	2,490	1.9%
5	1,667	1.3%

C. *Author's frequency*

To investigate the tendency for every author, we counted the number of posted novels per author from the metadata of all novels. Moreover, in order to investigate the deviation of the author's taste, we counted the number of keywords for every author.

Table IV shows the top 10 authors who have written the most novels on syosetu.com. We plotted the ranking and the number of novels written by the author in Figure 3. Both axes are on the log scale. The distribution of the number of novels follows the power law.

TABLE IV. TOP 10 WRITER

Rank	Author ID	# of novels
1	743	1,218
2	26055	491
3	107085	425
4	26407	300
5	153402	290
6	34969	265
7	9272	254
8	47590	233
9	126858	230
10	200	200

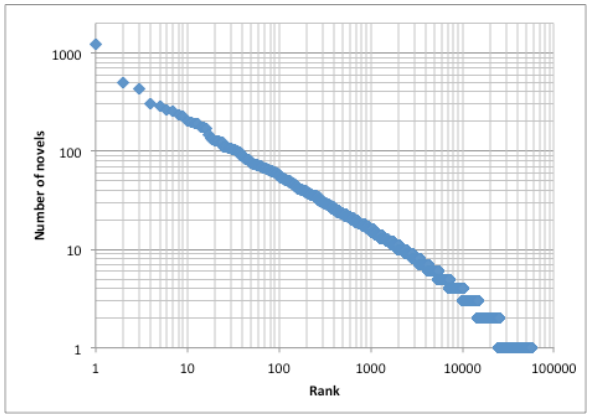

Figure 4. Rank-Number of novels per author

We investigated the correlation between the number of unique keywords and the number of novels for every author.

Figure 4 shows the scatter plot of the number of keywords and the number of novels for every author.

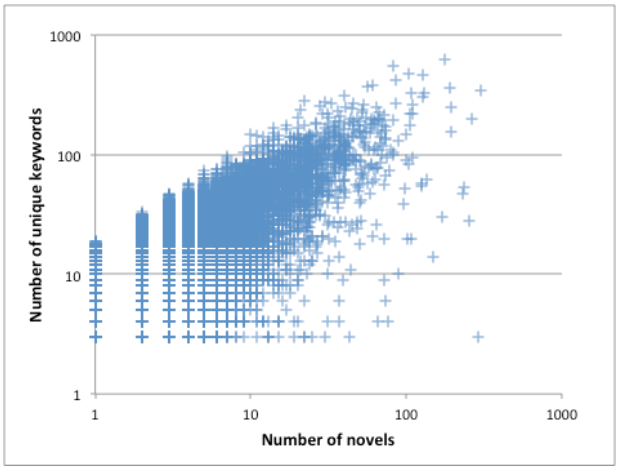

Figure 5. Novels-Keywords scatter plot per author

It seems that there is a straight line around the upper limit. In syosetu.com, the author can give 15 keywords at most to a novel. This is the reason the straight line is visible. If an author gives different keywords for every novel, this will become the number of unique keywords may be 15 times the number of novels, then it will appear as a straight line on the slope of 15. As for the results in the rightmost point of Figure 4, the author wrote 175 novels and gave 618 unique keywords.

IV. LINK ANALYSIS

We are interested in social ranking and recommendation methods based on the readers support, such as bookmarks, comments, and evaluation.

The reader's favorite can be represented as link from reader to novels, or authors. In order to understand the basic situation of link relation, we analyzed bookmark data, and favorite authors registered by readers.

A. Formalization

The entities of the link relations are authors, readers, and novels. Let A be the set of authors, U is the set of readers, and C is the set of novels. Let m be the number of authors, n be the number of readers, and s be the number of novels, then we can represent A, U and C as follows:

$$A = \{a_1, a_2, \ldots, a_m\},\ a_i\text{: author},$$
$$U=\{u_1, u_2, \ldots, u_n\},\ u_j\text{: reader (user)},$$
$$C=\{c_1, c_2, \ldots, c_s\},\ c_k\text{: novel (content)}.$$

Let the reader u_i bookmarks novel c_k, and the author a_i as the favorite author. We represent the links as $<u_i, c_k>$ and $<u_i, a_i>$. Each metadata page of the novel includes the author name and author ID, so it is possible to make link between the author and novel. If an author a_i posts two novels c_k and c_l, then link $<u_i, c_k>$ and $<u_i, c_k>$ are established. The total links between authors, readers, and novels will construct a tripartite graph as shown in Figure 5.

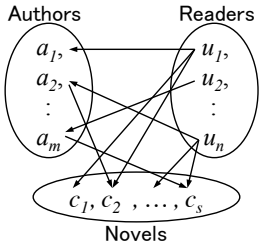

Figure 6. Tripartite graph of authors, readers and novels

B. Distribution of Links

Firstly, we analyzed the links from readers to novels. As shown in Table 1, there are 92,488 readers who open their mypage. We counted the links to favorite novels (bookmarked novels) for each of the readers mypage, and sorted them by the number of links. This number is the degree of links which goes to novels from a reader node. Figure 7 shows the scatter plot of the rank and the number of links for each reader. Both axes are on the log scale.

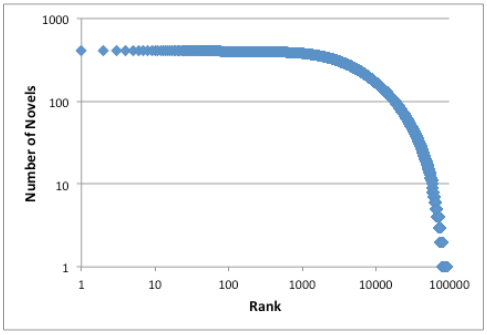

Figure 7. Rank-Number of favorite novels

The distribution of the number of links to bookmarked novels as shown in Figure 7 does not follow the power law. From first until 70th rank, the number of bookmarked novels is roughly the same, around the 410 mark. In other words, the top 70 mypage readers have bookmarked 410 favorite novels. 410 is also the maximum number of bookmarks in syosetu.com. We also pay attention to low rank nodes in Figure 7. There are not many readers who bookmarked less than ten novels. If the distribution follows the power law, the users who have zero bookmarks would be the majority. However, there is only 68. It is very few considering there is a total of 92,488 readers with opened mypages. This would be because opened mypages were crawled. A reader can select to open his/her mypage or not. Most mypage readers may not like to share only a few favorite in a lists on an open mypage.

Secondly, we analyzed the incoming links from readers to novel nodes. We sorted novels by the number of incoming links (by degree). Figure 8 is the scatter plot of the rank and

the number of incoming links for each novel. Both axes are on the log scale.

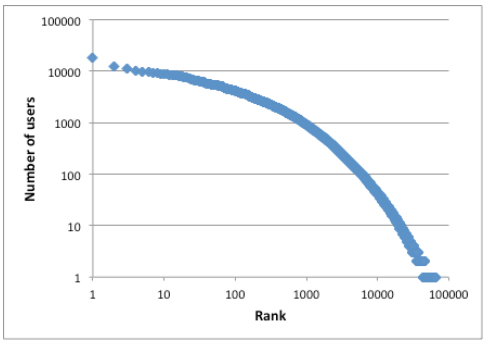

Figure 8. Rank-Number of bookmarked readers

The distribution of the plot in Figure 8 does not follow power law. Although it is close to the power distribution, it isn't a straight line, but instead a curve. The reason for this is not clear. We know that citation links between academic papers and the number of accesses in a Web site does not follow power law [9,10].

C. *Structure of Links*

Barabási and others [11,12] studied the topology structure of various graphs. They found that the distribution of web links to a page follows power law. They called the network like web a scale-free network. They also mentioned two conditions of scale-free networks:

(a) Growth: networks continuously grow by the addition of new vertices,

(b) Preferential attachment: new vertices connect preferentially to highly connected vertices.

We tried to investigate the structure of the bipartite graph. The graph continuously grows by the addition of new novels and readers. Readers tend to bookmark popular novels, because poplar novels are high ranked by popularity ranking, and it is easy for readers to check those high ranked novels. Then the bipartite graph fulfills growth and preferential attachment conditions of a scale-free network. However, as shown in Figure 7 and 8, the distribution of links does not follow the power law. The reason may be the upper limit on the number of bookmarks. We will investigate the structure of the graph in the future.

V. BASIC IDEA TOWARD SOCIAL RANKING OF ONLINE NOVELS

We will briefly describe the basic idea of the ranking method we are developing, which is based on readers' support.

There are two types ranking methods commonly provided in many contents sites. One is the ranking by popularity, and the other one is ranking by newness. Popularity ranking is implemented as the number of visitors, evaluation scores by viewers, and the number of bookmarked viewers.

The ranking by popularity is satisfactory for the beginners or light users of the syosetu.com site. Because beginners or light users have read no or few contents, they will be satisfied with the recommendation of higher popularity ranked contents. The ranking by newness is satisfactory for the heavy users. Heavy users spend a lot of time reading/viewing contents, and they know almost all of the popular contents. They like to find out new good contents from brand-new posted contents.

On the syosetu.com site it provides a simple popularity ranking. The novel i's weight w_i of popularity ranking is calculated using the following expression:

$$w_i = (L_i * 2) + \sum_{u \in U} (t_{u,i} + s_{u,i}) \quad ,$$

where, L_i is the degree of novel i, $t_{u,i}$ and $s_{u,i}$ are the story evaluation value for the story and scripting given by the reader u, respectively. Since the weight w_i is an accumulated value, the old good novels which were posted several years ago get high score, and they keep their higher rank. On the other hand, it is difficult for new novels to get high scores, even if it is a good novel.

The ranking for moderate users is problem. Moderate users like to read novels, but don't have enough time to read many novels. We are developing a new ranking method for moderate users based on social collective intelligence.

We hypothesize that there are only a few readers who can judge good contents, and we call them the selectors. Selectors initiatively select high quality novels from recently posted novels. They are heavy users because they read many novels.

Firstly, the ranking method we are developing finds selectors from the set of readers using the link structure between readers and novels. Secondly, an initial weight is assigned to the selectors, and the initial weight may be determined by the degree of links and evaluated score of novels. Thirdly, iteratively propagate the initial weight to readers and novels. After iteration, our method may amplifier the selectors weight to selected novels. We are currently developing the details of ranking method.

The novels highly ranked by our method will be high ranked in popularity ranking in the future. Early selectors find good novels, moderate readers read the novel and the bookmark it, and massive light-users will then read and evaluate it. Therefore, we will evaluate the ranking method we are developing by comparison between the ranking of our method using present data, and future popularity ranking. We will check the Spearman's rank correlation coefficient between the ranking of our method using present data, and future popularity ranking.

VI. CONCLUSION

User contents generation services have become popular recent years. Since the number of contents is so huge, contents ranking, classification, and recommendation methods play an important role in contents services. We

have been studying contents ranking and categorization methods, and are interested in the online novel service of Japan in this paper. There is no editor who assures the quality of novels, and no librarian who gives appropriate categories in online novel services. On the other hand, there are many readers who write comments and bookmark favorite novels. Readers' bookmark and comments express support of the novel. They are collective intelligence, and it is possible to use them as a resource for social ranking and recommendation.

In this paper, we focused on the Japanese online novel site "syosetu.com", and we analyzed the frequency of keywords, number of authors, and the links from readers to novels. Although the bipartite graph between readers and novels fulfills the growth and preferential attachment conditions of a scale-free network, the distribution of links does not follow the power law.

In the future, we will develop a social ranking method using readers' my page data. We extract bookmarked novels and favorite authors from my page files, and construct the tripartite graph of authors, readers and novels. After that, we will apply our developed algorithm, which applies a weight to each novel, and make new ranking list.

ACKNOWLEDGMENT

This work was supported by KAKENHI 2350099.

REFERENCES

[1] N. Murakami, E. Ito, "Video weighting method based on viewer's comments and its evaluation," Proc. of DEIM2012, March, 2012, p.F8-3. (in Japanese)

[2] N. Murakami, E. Ito, "Emotional video ranking based on user comments," Proc. of iiWAS2011, ACM, December 2011, pp.499-502.

[3] K. Baba, E. Ito, S. Hirokawa, "Co-occurrence Analysis of Access Log of Institutional Repository," Proc. of JCAICT2011, January 2011, pp.25-29.

[4] E. Ito, S. Hirokawa, K. Shimizu, "Introducing faceted views in diversity of online novels", Proc. ICDIM2012, IEEE, 2012, pp.??-??.

[5] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel, "Social Media Recommendation based on People and Tags," Proc. of SIGIR'10, ACM, 2010, pp. 194-201.

[6] H. Liang, Y. Xu, Y. Li, R. Nayak, X. Tao, "Connecting Users and Items with Weighted Tags for Personalized Item Recommendations," Proc. of ACM HT'10, ACM, 2010.

[7] Y. Hijikata, T. Shimizu, Shogo Nishida, "Discovery-oriented collaborative filtering for improving user satisfaction," Proc. of IUI 2009, ACM, 2009, pp.67-76.

[8] P. Kazienko, K. Musial, T. Kajdanowicz, Multidimensional Social Network in the Social Recommender System, IEEE Trans. on Sys., Man and Cybernetics, Part A, Vol.41, No.4, 2011, pp.746-759.

[9] I. Popescu, On a Zipf's Law Extension to Impact Factors, Glottometrics 6, 2003, pp.83–93.

[10] R. Mansillaa, E. Köppena, G. Cochob, P. Miramontes, On the behavior of journal impact factor rank-order distribution, Journal of Informetrics, vol. 1, issue 2, 2007, pp.155-160.

[11] A.-L. Barabási, R. Albert, H. Jeong, Mean-field theory for scale-free random networks, Physica A, vol. 272, 1999, pp.173-187.

[12] A.-L. Barabási, Linked: The New Science Of Networks Science Of Networks, 2002. (ISBN: 0738206679)