

## Introducing faceted views in diversity of online novels

Ito, Eisuke

Research Institute for Information Technology | Graduate school of ISEE

Hirokawa, Sachio

Graduate school of ISEE | Research Institute for Information Technology

Shimizu, Kazunori

Graduate school of ISEE | Research Institute for Information Technology

<https://hdl.handle.net/2324/24696>

---

出版情報 : Proceedings of ICDIM 2012 (Seventh International Conference on Digital Information Management), pp.145-148, 2012-08-30

バージョン :

権利関係 :

# Introducing faceted views in diversity of online novels

Eisuke Ito, Sachio Hirokawa

Research Institute for Information Technology

Kyushu University

Fukuoka, Japan.

{ito.eisuke.523@m, hirokawa@cc}.kyushu-u.ac.jp

Kazunori Shimizu

Graduate School of ISEE

Kyushu University

Fukuoka, Japan.

2IE11061G@s.kyushu-u.ac.jp

**Abstract**—In recent years, user generated content services have become popular. The authors are interested in online novel services. Classification of online novels is difficult because keywords and genre are assigned by the author of the novel without control. In order to overcome the problem faced when category classifying and searching online novels, faceted views were introduced and a cross tabulation search and analysis system was developed. This system can discover relations between novel genres and keywords, and can find the author's trend.

**Keywords**- user generated content; online novel; keyword frequency; search engine; recommendation; cross tabulation;

## I. INTRODUCTION

In recent years, user generated content services have become popular, like youtube.com, youku.com, and nicovideo.jp. Many movies are uploaded every day, and a huge number of movies have accumulated at these sites. If you are interested in photographs, flicker.com is popular. Online novel sharing services have also become popular, like qidian.cn in China, and syosetu.com in Japan. Most of the contents might be of a low quality, however a few have quite a high level of quality.

The amount of contents is increasing rapidly in user generated content services. As of May 1st of 2012, the most popular Japanese movie sharing service nicovideo.jp has over 15 million movies, and the online novel site syosetu.com, which is focused on in this paper, has over 130,000 novels. Although the number of novels is not extensive, it is increasing rapidly.

Search and recommendation engines play an important role in finding good contents, as there is too much contents in which to find good contents. There are important two functions for the search and recommendation of contents. The first function is the measurement of quality, and the other is the categorization of contents.

Traditional search and ranking systems are currently being used at youtube.com, nicovideo.jp, and syosetu.com. The user enters a query and the system returns a list of contents that match the query. The user can choose the ranking style, such as the number of replays, the length (bytes) of the content, recent commented on content, or the score given by viewers. Search and ranking by category restriction or tag specification are

possible. However, it is not enough to use the number of replays or viewers for content quality evaluation. Since these measures are accumulated overtime, older contents will tend to receive a higher quality evaluation. Collaborative filtering may be a good method for contents recommendation, but it has to collect preference data for each item from many users in order to be effective. A “cold start” is difficult to apply for new users and new items, because CF is based on the users' past preferences for items.

Tag clouds are a similar technique for automatic category classification. A tag cloud is useful for finding a major tag from not so many tags. However, when the number of contents is huge, then the number of tags also becomes very huge. It is impossible to take a glance at minor tags. Actually, more than 100 thousand tags (keywords) exist for only 100 thousand contents in syosetu.com. Clustering or hierarchical structure are necessary for massive contents categorization.

We previously considered ranking and categorizing contents using tags and comments by content viewers [1, 2]. The tags and comments are resource for data mining because they include user knowledge. We also studied scholarly papers that recommended using co-occurrence access [3].

In this paper, we focus on the Japanese online novel service “syosetu.com”. Compared with traditional printed books, it is difficult to classify the online novels in syosetu.com. In the case of traditional printed books, professional editors assure the quality of books, and librarians gave appropriate category words to each book. Category words are from a controlled vocabulary set, and there is little fluctuation for categorization. On the other hand, most online novel authors are amateurs. They are not trained in scripting, and do not know the controlled vocabulary used for categorization. The author may freely give keywords to their novel, some of which are not suitable as classification words. For example, a fantasy novel (like Harry Potter by J. K. Rowling), that has been specified as being in the “history” genre.

In order to overcome this problem faced when category classifying and searching online novels at syosetu.com, we propose the introduction of faceted views for a range of online novels, and the development of a cross tabulation search system [4]. This system has two search or classification axes, and a query phrase can be specified. Results are clustered into a matrix. The search results are displayed in table form.

The composition of this paper is as follows. In section 2, we describe the data structure of novels in syosetu.com, and provide some basic frequency analysis. Section 3 describes the cross tabulation search system which we developed, and simple evaluation of the system. Related work is shown in section 4, and we conclude our paper in section 5.

## II. BASIC FREQUENCY ANALYSIS

In this section, we describe the data structure of the online novel site “syosetu.com”, show the number of novels, the number of the authors, and the frequency of keywords (tags).

### A. The 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. Scores of novels given by readers, and the bookmarks of favorite novels lists of readers 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.

TABLE I. NUMBER OF NOVELS AND AUTHORS

Number of novels	134,763
Number of authors	56,236
Genres	18
Unique keywords	128,115

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 (tags), and 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.

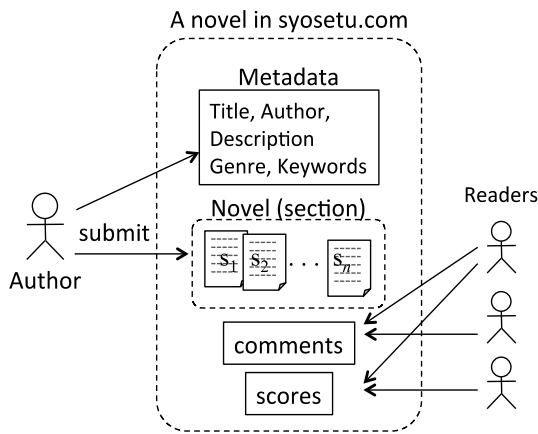


Figure 1. Data structure in syosetu.com

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.

### B. Keyword frequency

We collected the metadata files of each novel and count the frequency of words in the novel keyword field. There are 128,115 unique words. Table II shows the top 20 words and its 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 the service manager. In Table II, 1<sup>st</sup> “cruel”, 3<sup>rd</sup> “R15” are caution words. Readers can filter out novels by caution words.

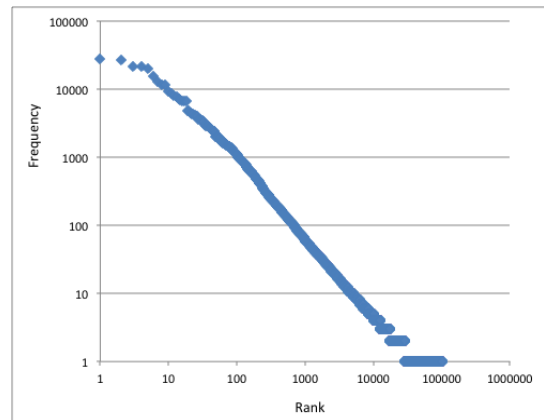


Figure 2. Rank-Frequency of keywords

Table III shows the number of low frequency words. Authors define a lot of the low frequency words, with 77.7% of words appearing only once.

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%

### III. CROSS TABULATION SYSTEM FOR ONLINE NOVELS

As previously mentioned, category classification of online novels on syosetu.com is difficult as the author assigns keywords freely and uncontrolled with a large variation in the use of words. In order to overcome this problem we developed a cross tabulation search and analysis system.

#### A. Outline

Figure 3 shows the outline data flow of our cross tabulation search system. This system is based on the inverted index file of the novel metadata. The keywords are categorized by their attributes. Even the same word may be indexed in several attributes.

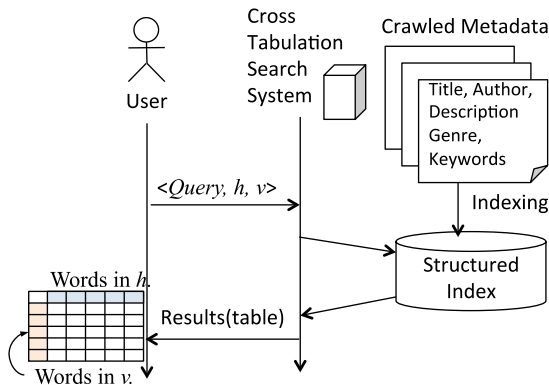


Figure 3. Data flow of cross tabulation search system

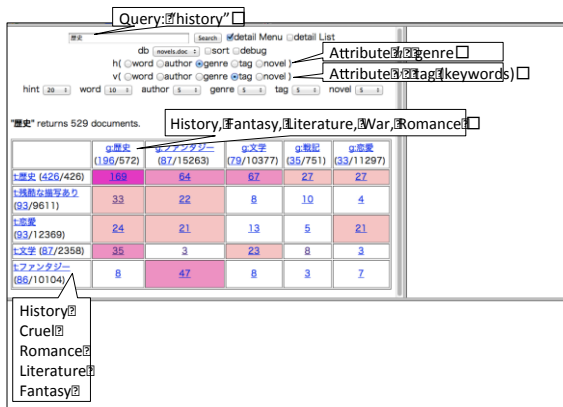


Figure 4. Web Interface (Query:history, Attributes  $h$ :genre,  $v$ :tag)

The search system has two search / classification axes. The user can enter not only a query, but also can specify two attributes for analysis. The first attribute  $h$  is used for horizontal cells, and second attribute  $v$  is used for vertical cells. Our search system returns a 6x6 size table, where 6x6 is default table size. The upper right horizontal cells are the top related words for the query in attribute  $h$  records. The left down vertical cells are the top related words for the query in attribute  $v$  records.

As an example in Figure 4, the user enters “history” as the query, and specifies the attributes “genre” and “keywords”, and the prototype search system would return the table as shown. In this case, the upper right cells are “History, Fantasy, Literature, War, and Romance” from genre, and the left down cells are “History, Cruel, Romance, and Fantasy” from keywords.

As shown in Figure 4, the background color of each cell is changed according to the number of documents. The cell with many documents is a dark color, and light color for middle number, and white for a few documents. Thereby, it is easy to understand which words are popular. It is also easy to grasp the trends of keywords or genre liked by authors.

Table IV shows the cross tabulation part of Figure 4 returned in the results from our system. We added the labels A, B,..., E to the horizontal cells, and 1,2,...,5 to the vertical cells for convenience. In the table, the number in the cell  $(i, j)$  ( $i=A..E, j=1..5$ ) means the number of the documents which contain the word of  $i$  in the attribute  $h$ , and word of  $j$  in the attribute  $v$ . For example, in Table IV, the value 27 in cell E1 shows the number of the novels that have “Romance” in the genre field, and “History” in the keywords field.

TABLE IV. CROSS TABULATION PART (QUERY: HISTORY, ATTRIBUTES H:GENRE, V:KEYWORDS)

	A	B	C	D	E	
	g:History	g:Fantasy	g:Literature	g:War	g:Romance	
1	t:History	169	64	67	27	27
2	t:Cruel	33	22	8	10	4
3	t:Romance	24	21	13	5	21
4	t:Literature	35	3	23	8	3
5	t:Fantasy	8	47	8	3	7

#### B. Evaluation

In our system, it is easy to understand the distribution of novels, and popular genre. Moreover, it provides the ability to understand the relations between genre and keyword. In this subsection, we evaluate our system qualitatively.

At first, consider the cells A3 and E1 in Table IV. A3 is the number of novels containing “history” for the genre, and “romance” in the keywords. E1 is the number of novels containing “romance” for the genre, and “history” in the keywords. The intersection set of E1 and A3 turns into an empty set. E1 contains 27 and A3 contains 24 novels. The total number of history genre novels is 572, but the total number of

romance genre is 11,297, and the ratio is about 20 times. From these data, history and romance can be applied together.

Next, consider cells A5 and B1. A5 is the number of novels containing “history” for the genre, and “fantasy” in the keywords. B1 is the number of novels containing “fantasy” for the genre, and “history” in the keywords. The intersection set of A5 and B5 also turns into an empty set. B1 contains 64, and A5 contains 8 novels. The total number of history genre novels is 572, but the total number of the fantasy genre is 15,263, and the ratio is about 20 times. From these data, authors who write history genre novels might not like fantasy.

It is possible to analyze the author's trend using our system, because it returns tabulation data based on specified attributes. For example, Tables V and VI show the trends of an author (author ID 50552) who posted 49 novels.

Table V shows the search results for the query: “a:50552”, both *h* and *v* are “genre”. It turns out that this author's favorite genres are literature, fantasy, romance, history and comedy.

TABLE V. GENRE TREND OF AUTHOR 50552

	g:Literature	g:Fantasy	g:Romance	g:History	g:Comedy
g:Literature	17				
g:Fantasy		15			
g:Romance			13		
g:History				3	
g:Comedy					1

Table VI shows the search results for the query: “a:50552” (author ID), *h* is “genre” and *v* is “keyword”. This table shows that this author is using various keywords for novels in the literary genre. On the other hand, he may not like to write love or romance in fantasy genre novels.

TABLE VI. GENRE AND KEYWORD TREND OF AUTHOR 50552

	g:Literature	g:Fantasy	g:Romance	g:History	g:Comedy
t:Fantasy	3	15	5		
t:Romance	6	4	12	1	
t:Serious	4	10	5		
t:Aother world	2	12	2		
t:Modern	5	1	2		

#### IV. RELATED WORK

Ido Guy and others [5] proposed a social recommendation system based on social media, such as SNS. They used relation between items, persons, and tags. In syosetu.com, authors and readers are identified by ID numbers, and from reader's

comments on a novel it is clear who posted the comment. Although the cross tabulation system was built for the analysis of a novel group this time, it may possible to apply Guy's technique for novel recommendation.

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

There is a problem with the conventional collaborative filtering in that too many already known items are recommended. Hijikata and others [7] proposed the concept of novelty as a measure, which recommends a new thing. They also proposed and evaluated three novelty based recommendation algorithms. Their novelty concept will be required for online novel search and recommendation systems.

#### V. CONCLUSION

Unlike metadata management of books in real libraries, in online novel services the author decides the genre and keywords of his/her novel. This makes it difficult to classify novels into an appropriate category because the genre and keywords are not controlled. We developed a cross tabulation system to solve this problem. Our system can discover relations between genre and keywords, and can find the author's trend.

Although trained librarians are not a part of online novel services, there are a lot of readers who assign comments and tags to novels. In the future, we plan to develop collective intelligence based methods of ranking, recommendation, and classification.

#### ACKNOWLEDGMENT

This work was supported by KAKENHI 2350099.

#### REFERENCES

- [1] Naomichi Murakami, Eisuke Ito, “Video weighting method based on viewer's comments and its evaluation,” Proc. of DEIM2012, March, 2012, pp.F8-3. (in Japanese)
- [2] Naomichi Murakami, Eisuke 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] C. Yin, S. Hirokawa, J. Yin-Kim Yau, T. Nakatoh, K. Hashimoto, Y. Tabata, “Analyzing Research Trends with Cross Tabulation Search Engine,” Int'l Journal of Distance Education Technologies, 2012. (to appear)
- [5] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, Erel Uziel, “Social Media Recommendation based on People and Tags,” Proc. of SIGIR'10, ACM, 2010, pp. 194-201.
- [6] Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak, Xiaohui 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.