Development of Facet Analysis System for Diverse Online Novels

Eisuke Ito, Kazunori Shimizu, Sachio
Hirokawa Kyushu University
Japan
ito.eisuke.523@m.kyushu-u.ac.jp, hirokawa@cc.kyushu-u.ac.jp, 2IE11061G@s.kyushu-u.ac.jp

ABSTRACT: *In recent years, user generated content services have become popular. The authors are interested in user generated 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 in classifying seaching 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 preference.*

Keywords: Cross Tabulation, Keyword Frequency, Online Novel, Recommendation, Search Engine, User Generated Content

Received: 4 June 2012, Revised 22 July 2012, Accepted 29 July 2012

1. 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 are too much 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 (CF) 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. The lack of enough evaluation date at the early stage of

recommendation system is known as the problem of "*cold start* ". It is difficult to find good recommendation 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 [5, 4, 8, 7]. The tags and comments are resource for data mining because they include user knowledge. We also studied recommendation of scholarly papers using co-occurrence access [1].

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, particularly in syosetu.com. In the case of traditional printed books, professional editors assure the quality of books, and librarians give appropriate category words to each book. Category words are chosen 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 authors 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 in classifying and searching online novels at syosetu.com, we propose the faceted views for a range of online novels, and developed a cross tabulation search system [9]. 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.

2. Basic Analysis of Data in syosetu.com

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.

2.1 The structure of syosetu.com

Syosetu.com is an online novel service provided by the *Hina-project* company. We crawled almost all the metadata (HTML pages) from syosetu.com in April 2012. Scores of novels given by readers, and the bookmarks of favorite novels lists of readers were also collected. Table 1 shows the number of published novels, authors who have written at least one novel, genre words, and unique keywords given for all novels.

Item	Number
Novel	134,763
Author	56,236
Genre	18
Unique keyword	128,115

Table 1. Number of 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 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, which is limited only by the number of bytes allocated for the synopsis and keywords.

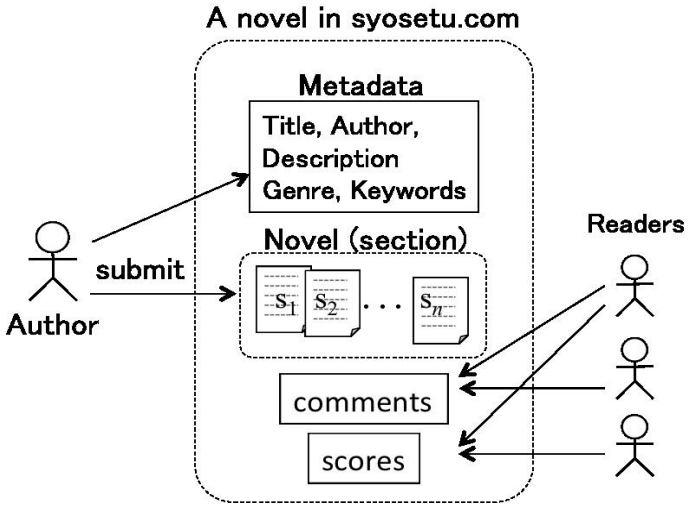

Figure 1. Data structure in syosetu.com

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.

2.2 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 2 shows the top 20 words with their frequencies. Some high frequency words in Table 2 are caution words which are specified by the service manager. In Table 2, 1st "*cruel* " and 3rd "*R*15" are caution words. Readers can filter out novels by caution words.

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

Table 2. Top 20 keywords and 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.

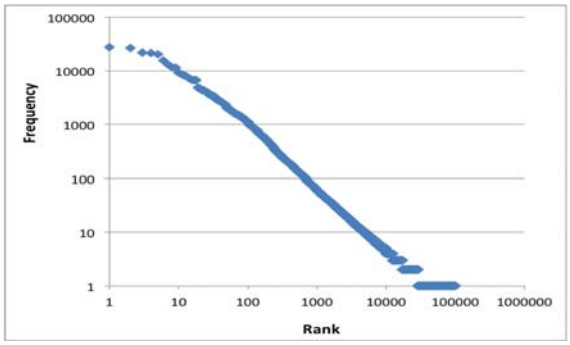

Figure 2. Rank-Frequency of keywords

Table 3 shows the number of low frequency words. Authors define a lot of the low frequency words. The 77.7% of all words appear only once.

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

Table 3. Ratio of low frequency words

3. 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.

3.1 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.

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 first row displays the ranked list of related words in attribute *h*. The first column displays the ranked list of related words in attribute *v*.

As an example in Figure 4, the user enters "*history*" as the query, and specifies the attributes "*genre*" and "*keywords*". The prototype search system returns the table as shown. In this case, we can see the selected genre "*History, Fantasy, Literature, War, and Romance*" in the first row. We cans see the selected keywords "*History, Cruel, Romance, and Fantasy*" in the first column.

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 preference of keywords and genres of authors.

Figure 5 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 Figure 5, the value 27 in cell *E*1 shows the number of the novels that have "*romance*" in the genre field, and "*history*" in the keywords field.

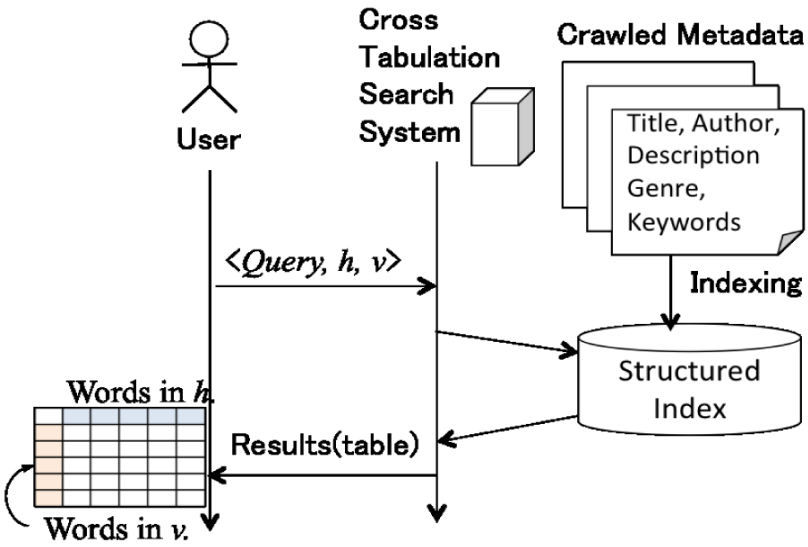

Figure 3. Data flow of cross tabulation search system

3.2 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 *A*3 and *E*1 in Figure 5. *A*3 is the number of novels containing "*history*" for the genre, and "*romance*" in the keywords. *E*1 is the number of novels containing "*romance*" for the genre, and "*history*" in the keywords. The intersection set of *E*1 and *A*3 turns into an empty set. *E*1 contains 27 and *A*3 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.

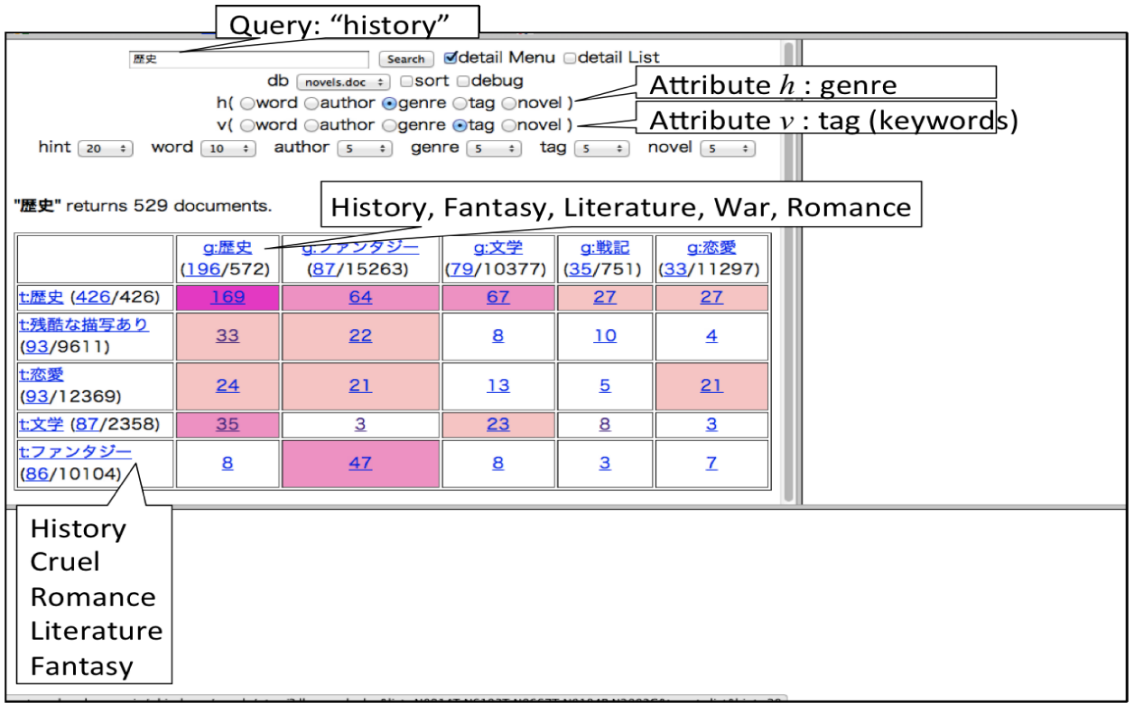

Figure 4. Web interface (Query:*history*, Attributes *h*:genre, *v*:tag)

	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

Figure 5. Cross tabulation part (Query:History, Attributes *h*:genre, *v*:keywords)

Next, consider cells *A*5 and *B*1. *A*5 is the number of novels containing "*history*" for the genre, and "*fantasy*" in the keywords. *B*1 is the number of novels containing "*fantasy*" for the genre, and "*history*" in the keywords. The intersection set of *A*5 and *B*5 also turns into an empty set. *B*1 contains 64, and *A*5 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 preference using our system, because it returns tabulation data based on specified attributes. For example, Figure 6 and 7 show the preference of an author (author ID 50552) who posted 49 novels.

Figure 6 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*".

Figure 7 shows the search results for the query: "*a*:50552" (this is an 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.

	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

Figure 6. Genre preference of author 50552 (Query:*a* : 50552, Attributes *h*:genre, *v*:genre)

4. Related work

Ido Guy and others [2] 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 fil-ter out the noise tags, 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 recom-mended. Hijikata and others [3] 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.

	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		

Figure 7. Genre and keyword preference of author 50552 (Query:*a* : 50552, Attributes *h*:genre, *v*:keyword)

5. 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 preference.

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.

References

[1] Kensuke Baba, Eisuke Ito, Sachio Hirokawa. (2011). Co-occurrence analysis of access log of institutional repository. *In*: Proc. of JCAICT2011, p. 25–29, January.

[2] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, Erel Uziel. (2010). Social media recommendation based on people and tags. *In*: Proc. of SIGIR 2010, p. 194–201. ACM.

[3] Yoshinori Hijikata, Takuya Shimizu, and Shogo Nishida. Discovery-oriented collaborative filtering for improving user satisfaction. *In*: Proc. of IUI2009, p. 67–76. ACM.

[4] Eisuke Ito, Sachio Hirokawa, and Kazunori Shimizu. Introducing faceted views in diversity of online novels. *In*: Proc. of ICDIM2012 (*Seventh International Conference on Digital Information Management*), p. 145–148. IEEE, August.

[5] Eisuke Ito and Kazunori Shimizu. 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*), p. 531–536. IEEE, November.

[6] Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak, Xiaohui Tao. (2010). Connecting users and items with weighted tags for personalized item recommendations. *In*: Proc. of HT2010, p. 51–60. ACM.

[7] Naomichi Murakami, Eisuke Ito. (2011). Emotional video ranking based on user comments. *In*: Proc. of ACM ii-WAS2011, p. 499–502. ACM, December.

[8] Naomichi Murakami and Eisuke Ito. Video weighting method based on viewer's comments and its evaluation. *In*: Proc. of DEIM2012, p. F8–3. IEICE, May 2012.

[9] Chengjiu Yin, Sachio Hirokawa, Jane Yin-Kim Yau, Tetsuya Nakatoh, Kiyota Hashimoto, Yoshiyuki Tabata. (2013). Analyzing research trends with cross tabulation search engine. *International Journal of Distance Education Technologies*, 11 (1) December.