Keywords frequency trend analysis of online novels

Ito, Eisuke
Research Institute for Information Technology

Urakawa, Takahiko
Research Institute for Information Technology

Flanagan, Brendan
Research Institute for Information Technology

Hirokawa, Sachio
Research Institute for Information Technology

出版情報：Proceedings of 2013 Second IIAI International Conference on Advanced Informatics, pp.68-73, 2013-09-10. IIAI
バージョン：
権利関係：

Keywords frequency trend analysis of online novels

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

BRENDAN FLANAGAN
Graduate School of Information Science and Electrical Engineering, Kyushu University.
Fukuoka, Japan.
b.flanagan.885@s.kyushu-u.ac.jp

TAKAHIRO URAKAWA
Graduate School of Systems of Life Science,
Kyushu University.
Fukuoka, Japan.
1te09084k@gmail.com

SACHIO HIROKAWA
Research Institute for Information Technology,
Kyushu University.
Fukuoka, Japan.
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—The authors are interested in online novel services as a user-generated media on the Web. A large number of novels are being uploaded, and a few novels become major. Novel writers like to create novel of current popular genre, then current popular genre words may frequently appear. In this paper, the authors apply the time series analysis to the keywords words given to an online novel by the creator. The authors construct a trend analysis tool. The tool not only shows the trend of posted query word(s), but also shows the trends of similar terms. This paper describes the trend analysis system, the used data, and some interesting analyses.**

Keywords: Online novel, user-generated media, keyword trend, frequency

I. INTRODUCTION

Online novel sharing service, which is a user-generated media on Web, becomes popular and a large number of novels are being uploaded. Most of online novels are written by amateur writers and might not be good quality. However, there are some high quality novels. Actually, the site "syosetu.com" in Japan [1], which is the research target of this paper, contains more than 14 million online novels as of September 2012 and is still drastically increasing the number of contents and viewers. The number of contents is also increasing in the site of "qidian.com" of China [2], as same as syosetsu.com.

There are no professional editor or no trained librarian concerning to UGM. Neither quality evaluation nor a categorization technique is clear. On the other hand, there are many viewers (readers) who perform comment and favorite registration. If we can use the collective intelligence gained by a large number of viewers, search, the ranking, and a classification of good quality may be possible.

We have been interesting in recommendation and categorization of contents, such as movies [2], and scientific papers [3]. We applied a faceted analysis of documents [4] as a part of categorization research. We also proposed a ranking method of the online novels based on viewer's bookmarks [5].

In this paper, we apply the time series analysis to the keywords of online novels, and analyze the changes trend. Novel writers like to write current popular genre. Most readers like to check current popular genre, and current popular genre novels are easy to be high-ranked on popularity ranking. Popularity ranking is based on the number of page view, unique viewing users, or bookmarks. So, current popular genre words may frequently appear in brand-new posted novels.

The composition of this paper is as follows. In section 2, we briefly describe service of syosetu.com, and show statistics of novels and keywords of the site. Section 3 describes the keyword trend analysis methods. In section 4, we explain the keyword trend analysis tool, which we developed. The tool not only shows the trend of posted query word(s), but also shows the trends of similar terms. In section 5, we discuss the developed tool with showing some interesting analyses. We conclude this paper in section 5.

II. STATISTICS OF SYOSETU.COM

This section shows some statistics of novels on syosetu.com, and illustrates an example of novel.

A. Number of novels

Table 1 shows the number of novels, registered readers, and writers on syosetu.com. Figure 1 shows the number of monthly novel posts. It is clear that the number of contents is drastically increasing.

Table 1 Number of novels, readers and writers

Item	Mar. 2012	May 2012	Jul. 2012	Sep. 2012
Novel	134,763	159,090	168,396	148,278
Reader	195,716	240,730	258,478	272,512
Writer	46,938	53,396	56,214	44,585

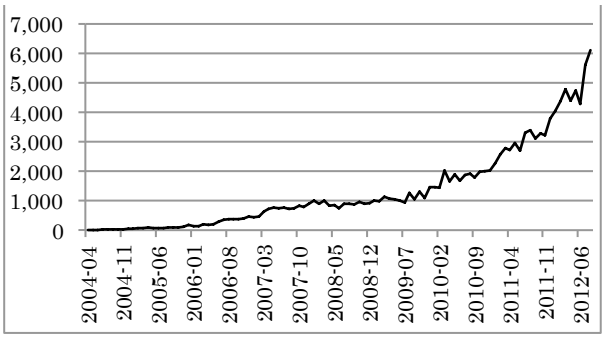

Figure 1 Monthly novel posts to syosetu.com

B. *Novel metadata*

A novel of 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 an example of novel metadata. It is the metadata of the novel "Knight's & Magic" (novel ID: n3556o). Figure 3 shows its TOC (table of contents) page. This is a serial novel, and the TOC page shows section titles.

小説情報

作者：天酒之瓢

小説タイトル	Knight's & Magic
あらすじ	メカヲタ社会人が異世界に転生。 その世界に存在する巨大な魔導兵器の乗り手となるべく、彼は情熱と怨念と執念で全力疾走を開始する……。
キーワード	R15 残酷な描写あり 関西人 ロボット 魔法 ファンタジー 異世界 転生 学園
掲載日	2010年 10月16日 23時27分
最終投稿日	2013年 02月06日 00時09分

Nコード	N3556O	開示設定	開示されています	お気に入り登録	19,978件
ジャンル	ファンタジー	感想	1,388件	レビュー	4件
種別	連載：全58部	感想受付	受け付ける(ユーザのみ)	レビュー受付	受け付ける
年齢制限	なし	文字数	556,667文字	ポイント評価受付	受け付ける
総合評価	69,140pt	文章評価	14,417pt	ストーリー評価	14,767pt

Figure 2 Metadata page (ID: n3556o)

Knight's & Magic

作者：天酒之瓢

メカヲタ社会人が異世界に転生。 その世界に存在する巨大な魔導兵器の乗り手となるべく、彼は情熱と怨念と執念で全力疾走を開始する……。

第１章　転生、そして学園生活編

#1 別れと出会い	2010年 10月 16日 （改）
#2 魔法を使おう	2010年 10月 19日 （改）
#3 旅には道連れ	2010年 10月 20日 （改）
#4 発想の転換	2010年 10月 23日 （改）
#5 図書館にて	2010年 10月 27日 （改）
#6 入学式にて	2010年 10月 29日 （改）
#7 その武器の名は	2010年 10月 29日 （改）
#8 授業をうけよう	2010年 11月 03日 （改）
#9 決闘の時間	2010年 11月 06日 （改）
#10 決闘の決着	2010年 11月 07日 （改）

第２章　魔獣襲来編

#11 陸皇襲来	2010年 12月 14日 （改）
#12 見学しよう	2010年 12月 22日 （改）

Figure 3 TOC page (ID: n3556o)

The author can set the title, the author's name, the genre, keywords and the outline as metadata, when he/she posts a novel. Genre has to be chosen from 15 keywords provided by the site manager. The author can choose arbitrary keywords freely with a limited length. The outline can be described as he/her wishes with a limited length.

C. *The number of keywords*

Table 2 shows the number of metadata files, unique words, and co-occurred keyword pair as of September 2012. We use the symbols D, W and P to represent the following data sets for convenient.

Table 2 Numbers at September, 2012.

Symbol	# of elements	Description
D	148,278	The set of novel metadata.
W	90,052	The set of keywords.
P	1,022,788	The set of pair of co-occurred keywords.

The number of elements of D is equal to the all posted and novels in syosetu.com. W is the set of the unique keyword, which appeared in the keyword column of the metadata. P is the set of keyword pair, which co-occurred in the keyword column.

III. KEYWORD TREND ANALYSIS WITH FREQUENCY

We use two methods for keyword trend analysis. One is the time plot of the keyword frequency for every period. This plot investigates the increase or the decrease of a word frequency. The other one is similar word analysis.

A. *Frequency to time plot*

The metadata of a novel has the post date (year, month, and day). Let m is a month, $D(m)$ is a subset of D, where d in $D(m)$ is a novel which is posted at m. Let $df(w, m)$ is the document frequency of the word w of $D(m)$. We plot $f(w, m)$ to the month m, $f(w, m) = df(w, m) / |D(m)|$, f is normalized value of df.

Trend is seen from time series change of $f(w, m)$. If $f(w, m)$ increases to a certain period, the word w is in fashion in the period. If it decreases, it can be judged that fashion has gone out of use.

1) *Target field*

It is possible to count word from title, outline, genre, and keyword. We only count words from keyword field in this time. The novel author gives the keywords by oneself. The author may believe that keywords express the feature of his own work. Therefore, it is suitable for the trend analysis. Since the number of words of the keyword field is small, it is suitable as the first trend analysis. In the future, we will check the words in other fields.

2) *Date*

There are two date fields in the metadata. One is submission date, and the other is the date of last update. In the plot of $df(w, m)$, we use the submission date as m. The

submission date may be reflecting the fashion at the starting time of a novel.

B. Similar words

Keywords are not controlled. Amateur writer gives keywords his/her own novel. There was some fluctuation in keyword. Then, the trend analysis tool not only show the trend of query words but also show the trend of related terms automatically calculated using similarity.

1) Why similar words?

In case of general things, it is easy to select query word. For example, the "weather" will be chosen as a query when you want to know weather forecast. You will add the place name to the query, if you want to know specific place weather. The "timetable" will be chosen as a query when you investigate the timetable of a train or a bus.

In the case of a special field of study, it is difficult to select a query word. If you are an expert of the field, you can select suitable word. For example, let us consider a case to investigate the performance of a smart phone. If you are an ICT engineer, it is easy for you to select the model name, OS name, the application name, the telecommunications standard, and so on. It is difficult to select suitable words, if you are not an expert.

In case of online novels, it is easy to select general genre words, such as history, sports, war, and so on. However, it is difficult to select suitable query word for a niche field. As shown in Table 2, the number of unique words is huge. Since a new word is continuously registered, diversity of keyword space becomes increasingly large. It may be a help to search novels by showing the similar word of a query word.

2) Two Similarities

Various similarities are defined. In this paper, we used cosine similarity and Jaccard coefficient. Cosine similarity can be calculable by two word vectors of document vector model. Jaccard coefficient is calculated as an operation during two sets. Let x and y are words, $D(x)$ is a subset of D, and document d in $D(x)$ has the word x in the document. We use Jaccard coefficient of $D(x)$ and $D(y)$ as the similarity of word x and y.

When calculating cosine similarity and a Jaccard coefficient, we have to care high frequency words and low frequency words. Let us consider low frequency words. If frequency of the words x and y are 1, x and y only co-occurs in one document, the $\cos(x, y)$ and $\text{Jaccard}(x, y)$ are 1, highest value. But x and y are not similar. To avoid this problem, we set a threshold and cut off low frequency words for similarity calculation.

Next, we consider high frequency words. A high frequency word appears in many documents, and it co-occurs with many words. Most of high frequent word is a general word. It may be an obstacle for investigation novels. We consider using IDF value (inversed document frequency) to reduce influence of a high frequency word. The effect of IDF is mentioned later.

IV. TREND ANALYSIS TOOL

In this section, we show our developed trend analysis tool. The tool consists of two parts. One is preprocessing part, and the other one is search part.

A. Preprocessor

Figure 4 shows the outline of the preprocessor.

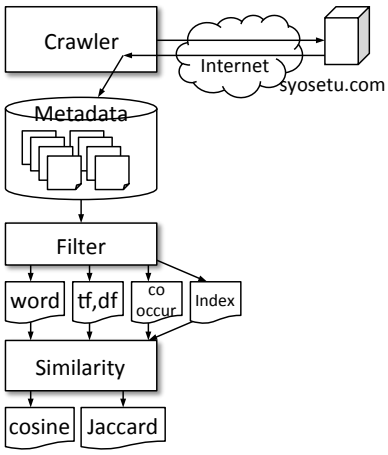

Figure 4 Preprocessing

1) Metadata crawler

We implemented the metadata crawler as a Ruby script using Web APIs (Narou-APIs). Hina-project Company provides APIs. Crawled metadata are written in YAML format.

2) Filter

We implement a word count program using Ruby language. We used Hash structure for word counting. And the filter program also calculates TF, DF, and extracts co-occurred words pair.

3) Similarity

As mentioned in section 3, we calculate similarity of each co-occurred words. The similarity program calculates cosine similarity, and Jaccard coefficient.

B. Search part

We implemented the search part (user Interface part) as a Web CGI program using the Apache web server. CGI programs are also implemented in Ruby language. Figure 6 shows the outline of the search part.

A user inputs query words from a web browser. The CGI program receives the query words, and it extracts top n similar words based on similarity of each co-occurred words. Normalized document frequency $f(w, m)$ for every month is extracted, and gunplot software makes a PNG format graph image using the extracted monthly frequency data. Figure 6 illustrates a web image of trend analysis.

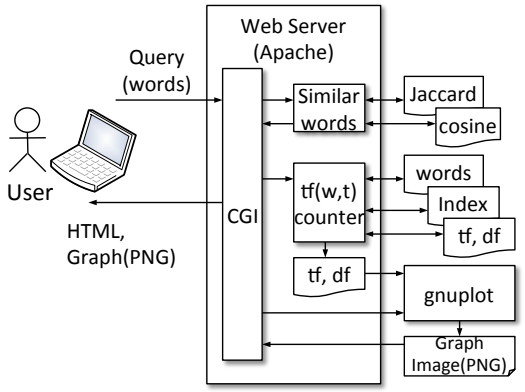

Figure 5 Search part (User interface part)

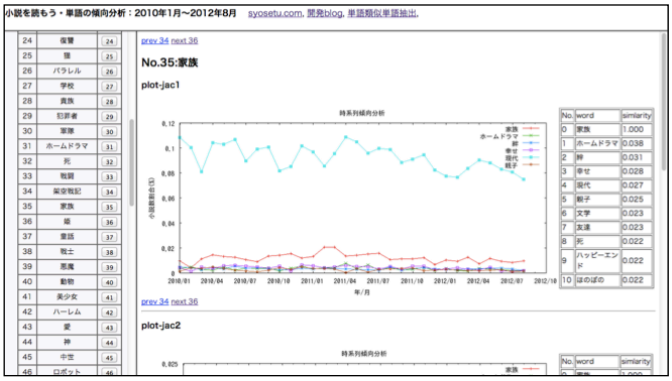

Figure 6 An image of the trend analysis

V. EXPERIMENT AND EVALUATION

This section describes the experiment and the evaluation of the proposed keyword trend analysis system.

A. *The data set*

Table 3 shows the data set, which we used as evaluation of our trend analysis. They are subset of data shown in table 2. We cut off low frequency words.

Table 3 Analysis data set

Symbol	# of elements	Description
D'	135,164	The set of novel metadata, which has at least one keyword.
W'	6,484	The set of keywords, which appear more than 5.
P'	36,804	The set of pair of co-occurred keywords, which co-occur more than 5.

B. *Time period for trend analysis*

The trend analysis period is from the date of syosetu.com service start until September 2012. We found that trend graphs radically changes at late of 2009. Figure 7 shows an example of radical curved graph.

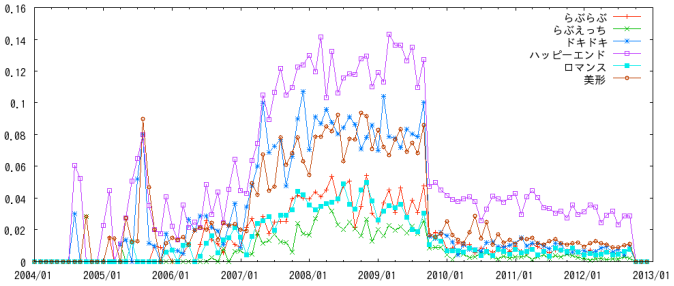

Figure 7 Radical curve at Nov. 2009.

We researched the reason of radical curve. A system trouble happed in 2009. As a result, the number of novel post was becoming fewer. Moreover, the regulation of novel post was changed in November 2009. After that, the maximum size of all keywords is less than 200 bytes. Then, the number of words in keyword field decreases, and normalized frequency $f(w, m)$ is also decreasing sharply. After January 2010, keyword trend is stable.

C. *Comparison of slimiarity*

We compared two similarities. Table 4 shows top 10 similar words for "family". (1) and (2) are Jaccard, and (3) and (4) are cosine. (2) and (4) are the value with IDF value of similar word.

Table 4 Comparison of similarity (for "family")

(1) Jaccard

No.	word	simlarity
0	family	1
1	home drama	0.038
2	ties	0.031
3	happiness	0.028
4	modern	0.027
5	parent and child	0.025
6	literature	0.023
7	friends	0.023
8	death	0.022
9	happy ending	0.022
10	hono-bono	0.022

(2) Jaccard * idf

No.	word	simlarity
0	family	1
1	home drama	0.294
2	ties	0.246
3	happiness	0.221
4	parent and child	0.205
5	friends	0.185
6	modern	0.184
7	brother	0.178
8	death	0.173
9	mind	0.171
10	impression	0.163

(3) Cosine

No.	word	simlarity
0	family	1
1	home drama	0.085
2	ties	0.07
3	parent and child	0.064
4	modern	0.062
5	happiness	0.061
6	hono-bono	0.061
7	literature	0.053
8	brother	0.052
9	happy ending	0.051
10	father	0.051

(4) Cosine*idf

No.	word	simlarity
0	family	1
1	home drama	0.656
2	ties	0.555
3	parent and child	0.524
4	happiness	0.49
5	wife	0.463
6	father	0.455
7	brother	0.435
8	mordern	0.426
9	hono-bono	0.401
10	mother	0.394

We inspect influence of IDF. In the right tables, (2) and (4), similarity are multiplied by IDF value of the similar word. Table (1) and (3) have the word “literature”, but (2) and (4) don’t have it. “Literature” is one of designated genre words. Its frequency is high, and it is a general word. IDF multiplied similarity has good effect to avoid general words. Next we compared Jaccard and cosine similarity, but couldn’t find clear difference between two. We thought that Jaccard may be a little bit better than cosine.

D. Case study

This subsection shows four interesting trend analyses. The period of trend analysis is from January 2010 until September 2012.

1) Case “Game”

Figure 8 shows the trend of “game”. Similar words of game are “RPG”, “Online”, “VRMMO”, “MMO” and “VR”. These are computer game related words. “Game” and similar words are rising after October 2011. We know that computer game like novels are become popular in syosetu.com.

2) Case “Sea”

Figure 9 shows the trends of “sea” and similar words of sea. They are ship, summer, mermaid, nature and pirate. In this graph, the trend of “summer” is significant. The blue line of figure 9 is the trend of “summer”, and it rise at summer season in every year.

3) Case “Tsun-dere”

“Tun-dere” [9] is a Japanese character development process that describes a person who is initially cold and even hostile towards another person before gradually showing his or her warm side over time. The red line is the trend of “Tsun-dere”, and the green line is the trend of “Yan-dere”, which is similar to “Tsun-dere”. Red line (tsun-dere) shows a declining trend, but grean line (yan-dere) shows an uptrend.

4) Case “Harem”

Figure 11 shows the trends of “harem” and similar words of it, cheat, strongest, beauty, love comedy, and slave. These topics become popular.

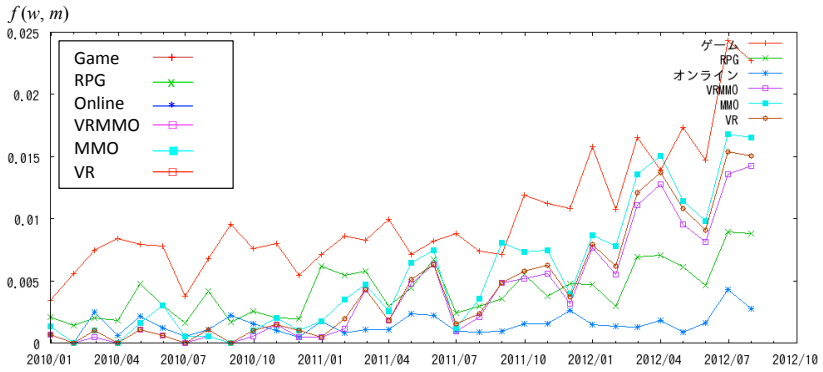

Figure 8 Trend of “Game”

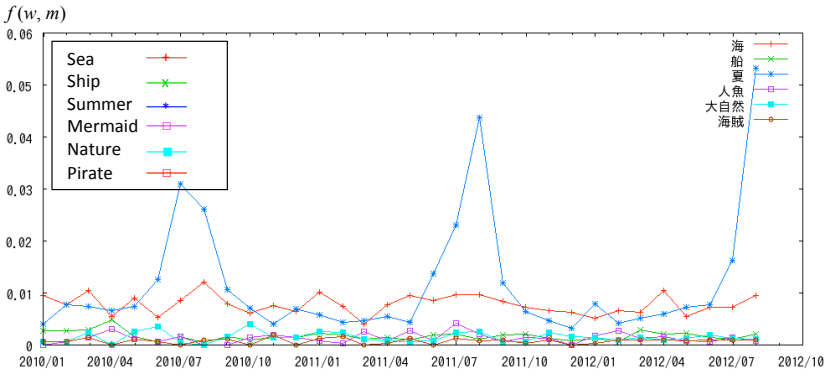

Figure 9 Trend of “Sea”

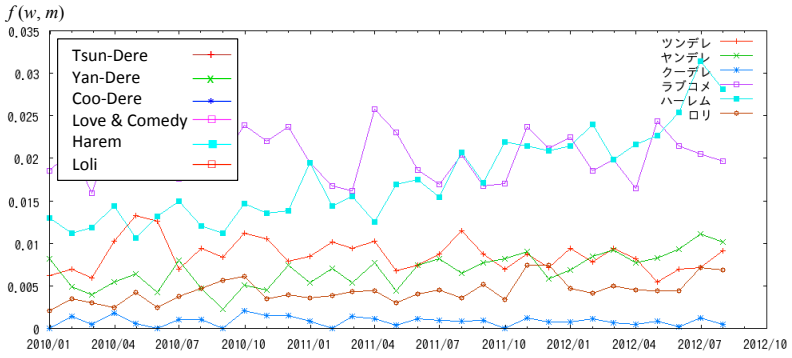

Figure 10 Trend of "Tsun-dere"

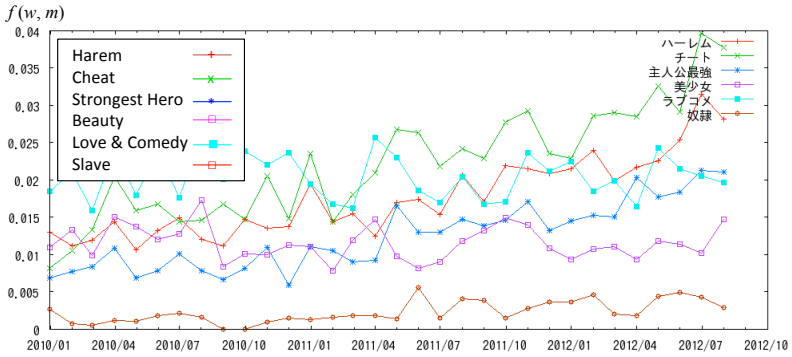

Figure 11 Trend of "Harem"

VI. Related Work

Yin et.al proposed a keywords trend analysis system [10]. They apply their system for research proposal documents. Figure 12 shows an analysis of their system. Although there are many researchers, discipline are diverse, researchers who works a subject may not be many. Therefore, keyword will appear in sparse.

Google trends [11] provides time trends analysis based on query words. Figure 13 showed trends of "game, RPG, online", VRMMO, MMO and VR". They are the same words in Figure 8. Present Google trends does not have similar words support.

Moving average is suitable technique to evaluate wide range trend. Trend of stock price are illustrated in moving average, and range is 13-week, 26-week. We should implement moving average into our trend analysis tool.

VII. Conclusion

Online novel become popular in recent years. Because a lot of contents are archived in UGM service, search engine plays an important role to find good contents and automatic categorization. In this paper, we apply the time series analysis to the keywords of online novels. We developed a trend analysis tool as a Web CGI system. The tool shows not only trend of query word, but also shows trend of similar words. We show four interesting trend analyses. The tool illustrates that relation of trends of similar words. In the future, we want to implement moving average for trend graph plot.

Acknowledgement

This work was supported by KAKENHI 2350099.

References

[1] Hina-project, Syosetuka-ni-narou, http://www.syosetu.com/ .

[2] QiDian, http://www.qidian.com/.

[3] N. Murakami, and E. Ito: Emotional video ranking based on user comments. Proc. of ACM iiWAS2011, pp. 499–502, ACM (2011).

[4] K. Baba, E. Ito and S. Hirokawa: Co-occurrence analysis of access log of institutional repository. Proc. of JCAICT2011, pp. 25-29 (2011).

[5] E. Ito, S. Hirokawa, and K. Shimizu: Introducing faceted views in diversity of online novels, Proc. of ICDIM2012 (Seventh International Conference on Digital Information Management), pp.145-148, IEEE, (2012).

[6] K. Shimizu, E. Ito and S. Hirokawa: Predicting Future Ranking of Online Novels based on Collective Intelligence, Proc. of ICDIPC2013 (The Third Int'l Conf on Digital Info. Processing and Communications), pp.261-272, SDIWC (2013).

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

[8] Hina-project, Narou-APIs, http://dev.syosetu.com/man/api/ .

[9] Tundere, http://en.wikipedia.org/wiki/Tsundere , Wikipedia.

[10] Yin, C., Hirokawa, S., Yau, J. Y., Nakatoh, T, Hashimoto, K., and Tabata, Y.: Analyzing research trends with cross tabulation search engine, International Journal of Distance Education Technologies, Vol.11, No.1, pp.31-44, (2013).

[11] Google, Google Trend, http://www.google.com/trends/ .