Keyword Relation Analysis Using Concept Graph Toward Automatic Categorization of Online Novels

Eisuke, Ito
Research Institute for IT, Kyushu University

Sachio, Hirokawa

http://hdl.handle.net/2324/1398574

出版情報：Proceedings of IIAI AIT 2013 (IIAI International Conference on Advanced Information Technologies), pp.24-, 2013-11-28. IIAI (International Institute of Applied Informatics)
バージョン：
権利関係：

Keyword Relation Analysis Using Concept Graph Toward Automatic Categorization of Online Novels

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

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

Abstract—**Contents of user-generated media (UGM) such as YouTube are popular in recent years. The authors are interested in online novel services as a UGM on the Web. A large number of novels are being uploaded, and a few novels become major. Because there are no professional editor or no trained librarian concerning to UGM, neither quality evaluation nor categorization of UGM contents is clear. Moreover, the creator of a novel gives keywords without control, and they often generate new words, then it is difficult to use keyword for categorization. Toward automatic categorization, we have to understand relation between keywords. In this paper, we focused on analysis of keywords relation using concept graph. The concept graph is based on statistical analysis of word frequency and co-occurrence. We also show several interesting relations on concept graph.**

***Keywords-** User generated media; online novel; categorization; concept graph*

I. INTRODUCTION

Contents of user-generated media (UGM, for short) such as YouTube are popular in recent years. Online novel sharing service, which is a UGM on Web, becomes popular and a large number of novels are being uploaded. We are interested the site “sy-osetu.com” in Japan [1] as the research target of this paper, and this site contains more than 14 million online novels as of September 2012 and is still drastically increasing the number of contents and readers. The number of contents is also increasing in the site of “qidian.com” of China [2], as same as Japan. Most of online novels are written by amateur writers and might not be good quality. However, there are a few high quality novels, and high quality novel is published as a paper book or e-book from major publisher.

We have been interesting in recommendation and categorization of UGM contents, such as movies [3], and scientific papers [4]. We also proposed a ranking method of the online novels based on viewer’s bookmarks [5]. We applied a faceted analysis of documents [6,7] as a part of categorization research.

In this paper, we focus on categorization of online novels. 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 at the time of publishing. Trained librarians gave appropriate category words to each book. Category words are from a controlled vocabulary set. Trained librarians select words of controlled vocabulary, and each word is assigned at a tree structure. The tree structure describes semantic hierarchy. There is little fluctuation for categorization.

On the other hand, there is no professional editor or no trained librarian concerning to UGM, neither quality evaluation nor categorization of UGM contents is clear. 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, is given “history” as a keyword. Moreover, they often generate new keywords, which is not referred on dictionary.

A tag cloud is useful for finding a major tag (keyword) from not so many keywords. However, when the number of contents is huge, then the number of keywords also becomes very huge. It is impossible to take a glance at minor keywords on tag cloud. Actually, more than 90 thousand unique keywords exist for 140 thousand contents in syosetu.com. Clustering or hierarchical structure are necessary for massive contents automatic categorization.

Toward automatic categorization, we have to understand relation between keywords. In this paper, we focused on analysis of keywords relation using concept graph [8,9]. Concept graph is based on statistical analysis of word frequency and co-occurrence. We also show several interesting relations on concept graph.

The composition of this paper is as follows. In section 2, we briefly describe some statistical data of syosetu.com, such as number of novels, readers, and keywords. Section 3 describes classification methods. We compare taxonomy, folksonomy, and fluxonomy. Section 4, describes the definition of concept graph. In section 5, we show some interesting results of concept graph analysis. We conclude this paper in section 6.

II. STATISTICS OF SYOSETU.COM

This section shows some statistics of syosetu.com, and illustrates an example of metadata of a novel.

A. Number of novels

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

TABLE I. 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
Author	46,938	53,396	56,214	44,585

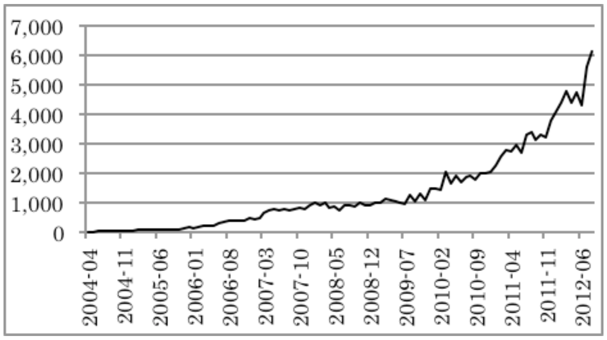
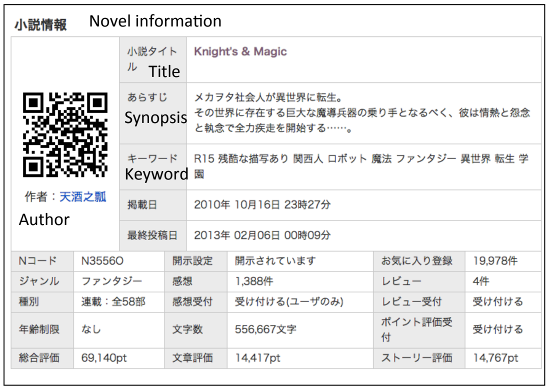


Figure 1. Monthly novel posts to syosetu.com

B. Novel metadata

A novel of syosetu.com is classified into short novel or series novel. A serial novel consists of several sections, and 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. Fig. 2 and 3 show an example of novel metadata and its TOC (table of contents) page. They are the metadata and TOC of the novel "Knight's & Magic" (novel ID: n3556o). This is a serial novel, and the TOC page shows section titles.



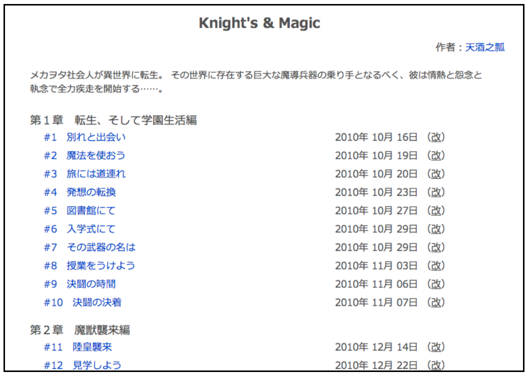
小説情報 Novel information

作者：天酒之瓢
Author

小説タイトル Title	Knight's & Magic
あらすじ Synopsis	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 Keyword	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. The metadata of a novel (Novel 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. The TOC page of a novel (Novel ID: n3556o)

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

C. The number of keywords

Table II shows the number of metadata files, unique keywords, 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. 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.

TABLE II. NUMBER OF METADATA, KEYWORDS AND CO-OCCURRED KEYWORDS (SEP. 2012)

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

III. CLASSIFICATION

This section describes several classification and categorization technics.

A. Taxonomy

Controlled vocabulary [10] is a carefully selected list of words and phrases, which are used to tag units of document or work. For instance, the society of ACM (association for computing and machinery) defined a controlled vocabulary named the ACM computing classification system [11]. Author must give a word and the number of a subject to the article. They may be more easily retrieved by a search, and classification of articles. Because scientific and technology articles are written by well-educated researcher, and research discipline are well defined, then controlled vocabulary is useful for classification of scientific and technology articles.

Academic society and community update name of research subject.

A library classification [12] is a system of coding, assorting and organizing documents, library materials or any information according to their subject and allocating a call number to that information resource. In Japan, there is NDC (Nippon Decimal Classification), and librarian in a library gives a code to a book.

However, controlled vocabulary and library classification may not be useful for classification of online novels. Most online novel authors are amateurs, and the author may freely give keywords to their novel. Moreover, they often generate new words and acronyms, which dot not recorded on dictionary, such as *chearem*[1], *TS*[2], and *VRMMO*[3].

WordNet [13] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is useful for natural language understanding. However, it may not be useful for classification of online novels.

B. *Folksonomy*

A folksonomy is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content [14]. Folksonomy, by Thomas Vander Wal, is a portmanteau of folk and taxonomy.

Li and others [15] used ISR (inter section ratio) to extract semantic hierarchy of tags (keywords) from a large number of social bookmarked web pages. Let *x* and *y* be a word, and *df*(*x*) be the document frequency of the word *x*. ISR is defined as follows:

$$df(x) > df(y), \qquad (1)$$

$$ISR(x,y) = df(x*y)/df(y). \qquad (2)$$

The *df*(*x* * *y*) is the document frequency of *x* and *y*, that is the number of documents which has *x* and *y* in keywords, and this is equal to |*D*(*x*, *y*)|.

For example, let us consider two words “Google” and “gmaps”. If most documents, which are given “gmaps”, are also given “Google” as keyword, but if a small number of documents, which are given “Google”, are given “gmaps” as keyword, then it is possible to infer that “gmaps” is a sub-word of “Google”.

If you calculate ISR of all co-occurred words, then you get partial ordered set of words. Let a word be a node and let a partial order be a directed edge, and then you get a directed acyclic graph. The graph may represent the semantic hierarchy of words. However, there are so many edges in the graph that it is difficult to understand relation of nodes (words). As shown in table 2, the number of keywords is about 0.9 million, and the number of co-occurred word pair is more than 1 million, then the graph size becomes so huge graph. If you cut off low frequency word, as shown in table 3, there are still more than 6 thousand words.

C. *Fluxonomy*

Fluxonomy is a portmanteau of fluxion and taxonomy. Hamano [16] introduced a tag selection system on nicovideo.jp. Nicovideo.jp is a YouTube like UGM, and it is popular in Japan. User can give at most ten tags to a movie. On nicoviddeo.jp, anybody can edit (delete and add) tags of a movie. Some tags are survive long time, but most tags are deleted soon. If audiences adopt a tag, then it is used long time. But if they don’t adopt, then the tag is deleted. So, tags of nicovideo.jp are not fixed, but stable eventually. He found that the tag selection mechanism is similar with natural selection of biological existence.

IV. CONCEPT GRAPH

To realize automatic categorization of huge amount of online novels, clustering and extraction of semantic hierarchy are necessary. Toward automatic categorization, we have to understand relation between keywords. To analyze keywords relation, we use concept graph [8, 9].

Concept graph is a directed acyclic graph of words introduced in [8]. It is a formulation of hypernym/hyponym relation of words according to frequencies. Let *D*(*q*) be the set of documents that satisfy the query *q*. A word *u* is a hypernym of *v* with respect to *D*(*q*) when they satisfy the following two conditions:

$$df(u) > df(v), \qquad (3)$$

$$df(u*v, D(q))/df(v, D(q)) > \alpha, \qquad (4)$$

where, the α in equation (4) is threshold.

Hypernym/hyponym relation determines an order structure among characteristic words and can be drawn as a directed acyclic graph. However, some words have too many hypernym and the graph may contain edges overlapped each other. To obtain more clear structure, we define "direct upper/lower" relation between words. Given a word *v*, the set *UP*(*v*) of upper words of *v*, and the set *DUP*(*v*) of direct upper words of *v* are defined as follows:

$$UP(v) = \{u \in D(q) \mid u\ is\ a\ hypernym\ of\ v\}, \qquad (5)$$

$$DUP(v) = \{u \in UP(q) \mid \forall w \in UP(v), u \in UP(w)\}. \qquad (6)$$

A hypernym *u* of *v* is a direct upper hypernym when there are no other hypernym of *v* between *u* and *v*. Visualization of a Concept Graph can be obtained by placing words of high frequencies on the left sides and ones with lower frequencies on the right sides. Thus a directed edge looks like an arrow from right to left.

[1] *Chearem* is a portmanteau of cheat and harem

[2] Trans Sexual

[3] Virtual Reality Massive Multi-player Online-game

V. Experiments

A. The data set

The row data shown in table II includes so many words, and then we cut off low frequency words. Table III shows the data set, which we used in this analysis. The cut off threshold of frequency is five.

TABLE III. Number of test data set

Data set	# 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. Keyword relation analysis

To evaluate effectiveness of concept graph for analysis of keyword relation, we checked some words. We show some cases of graph-based analysis.

1) Case 1: Efficient of threshold α

α in equation (4) is the threshold to cut off edges. If the probability of co-occurrence of two words *u* and *v* is less than α, *u* and *v* are not linked in the concept graph. If there is no link to *v*, then node of word *v* does not appear on the graph. The larger value you set to α, the less number of nodes is appearing.

To find suitable value for α, we set some values to α, and looked at the concept graph output as a result. Fig. 4 shows two concept graphs for query 'war'. The word 'war' is a high frequency word, and *df*('war')=2778. α is 0.2 in upper graph of Fig. 5, and α is 0.1 in lower one.

As shown in Fig. 4, lower one displays suitable relation of words. We checked other graphs for other queries, and then we judged that suitable value of α is 0.1 for keyword relation analysis.

2) Case 2: Concept graph of 'war' (α=0.1)

Fig. 5 shows the detail of the lower concept graph in Fig. 4. In the graph, we set the α as 0.10. The graph shows interesting four parts. The area, marked with (1) in Fig. 5, forms a straight line without any splitting branch. It includes "army, fictitious war, SF, nation, race, and record of war". These words are general words, and easy to understand semantic relation.

Left (lower) graph consists with three parts, and we labeled them as (2), (3) and (4). Part (2) may explain modern weapons and arms in 20 century. Part (3) shows 'middle age', 'history', 'world war II', and 'warring states period (in Japan)'. O*da-Nobunaga* (1534-1582) is a famous and popular war load in the period of Japan. Part (3) illustrates that there are some historical novels, and keywords are classified into the part. Part (4) shows future fictional words, such as "robot, space and hi-tech". They may show that some SFs (scientific fictions) are uploaded to syosetu.com, and keywords of SF are classified into part (4).

3) Case 3: Concept graph of 'game'

Frequency of 'game' is 455. So this word is a middle class frequency word. And novelization of computer game is a popular in syosetu.com. Computer game based novels have their origins in major game novels such as "Sword Art Online[4]" and "Log Horizon[5]".

Fig. 6 shows the concept graph of 'game'. We also labeled the interesting areas of the graph. First area is just below the root, which we labeled (1). It includes "cruel, fantasy, R15, high school student, and another world". These words are general words in syosetu.com. 'Cruel' and 'R15' are warning keywords, they are is used as filter.

Left (lower) part consists with three parts, and we labeled them with (2), (3) and (4). Part (2) may shows mystery and horror. Part (3) has "RPG (role playing game), sword, VRMMO (virtual reality massive multi-player online game)". Part (4) may explain novels, which is based on SF (science fiction) and war games.

C. Discussion

As shown in Fig. 5 and 6, concept graph extracts semantic relations of low frequency keywords. However, relations of high frequency keywords are not good. In Figure 5, the word 'war', which is near to the label (4), is placed to low node for 'game'. *df*('war')=2778, and *df*('game')=455, then 'war' should be upper word for 'game'.

The reason is clear that according to the equation (4), concept graph is generated for query *q*. In case 3, query is 'game', then the graph in Fig. 6 is generated for *D*('game'). For automatic categorization of all contents, we need the concept graph for all documents *D*. It includes the whole keywords, and the whole edges of hypernym/hyponym relation.

VI. 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 concept graph analysis to the keywords of online novels. Concept graph extracts semantic relations of keywords.

In the future, we will generate the concept graph for all documents. And we want to compare concept graph and other clustering.

Acknowledgment

This work was supported by JSPS Grant Number KAKENHI 2350099 and 24500176.

[4] Sword Art Online, ISBN-10: 4048677608

[5] Log Horizon, ISBN-10: 4047271454

REFERENCES

[1] Hina-project, Syosetuka-ni-narou, http://www.syosetu.com/, (accessed at June 24, 2013).

[2] QiDian, http://www.qidian.com/, (accessed at June 24, 2013).

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

[4] K. Baba, E. Ito and S. Hirokawa, "Co-occurrence analysis of access log of institutional re-pository," Proc. of JCAICT2011, 2011, pp.25-29.

[5] 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. Pro-cessing and Communications), SDIWC, 2013, pp.261-272.

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

[7] E. Ito, S. Hirokawa, and K. Shimizu, "Development of Facet Analysis System for Diverse Online Novels," Journal of Data Processing, volume 2, issue 3, DLINE, 2012, pp.113-119.

[8] Y. Shimoji, T. Wada and S. Hirokawa, "Dynamic Thesaurus Construction from English-Japanese Dictionary," Proc. of the Second International Conference on Complex, Intelligent and Software Intensive Systems, 2008, pp.918-923.

[9] K. Qian, S. Hirokawa, K. Ejima, X. Du, "A fast associative mining system based on search engine and concept graph for large-scale financial report texts," Proc. of 2nd IEEE ICIFE, IEEE, 2010, pp.675-679.

[10] Controlled vocabulary in Wikipedia, The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Controlled_vocabulary, (accessed at June. 24, 2013, UTC).

[11] ACM, "The ACM Computing Classification System (1998)," http://www.acm.org/about/class/ccs98-html, ACM (accessed at June 24, 2013).

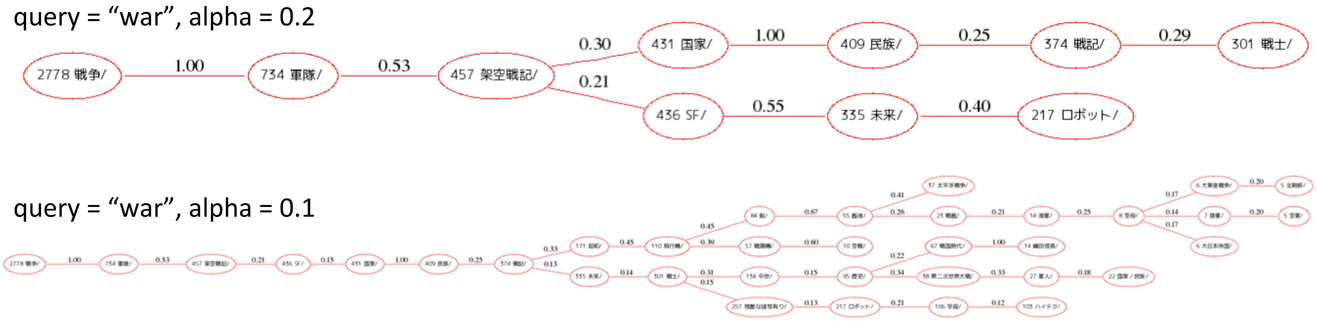
[12] Library Classification in Wikipedia: The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Library_classification, (accessed at June. 24, 2013, UTC).

[13] Wordnet, http://wordnet.princeton.edu/, Princeton University (accessed at June 24, 2013).

[14] Folksonomy in Wikipedia: The Free Encyclopedia. Retrieved from http://en.wikipedia.org/wiki/Folksonomy, (accessed at June. 24, 2013, UTC).

[15] R. Li, S. Bao, B. Fei, Z. Su, and Y. Yu, "Towards Effective Browsing of Large Scale Social Annotations," Proc. of WWW2007, ACM, 2007, pp.943-952.

[16] S. Hamano, et.al.: "CGM-no-genzai-to-mirai (Now and future of CGM)," IPSJ, ISBN:4915256839, 2012. http://www.ipsj.or.jp/magazine/5305.html


query = "war", alpha = 0.2

query = "war", alpha = 0.1

Figure 4. Concept graphs of 'war'

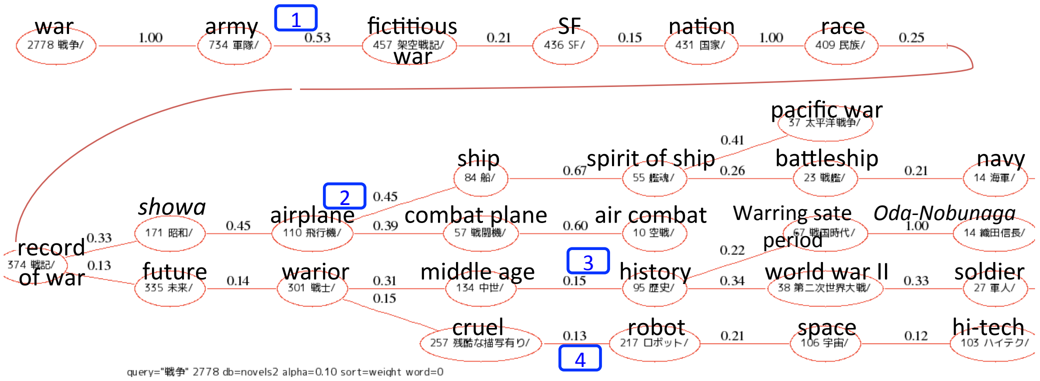


Figure 5. Concept graph of 'war' (α=0.1)

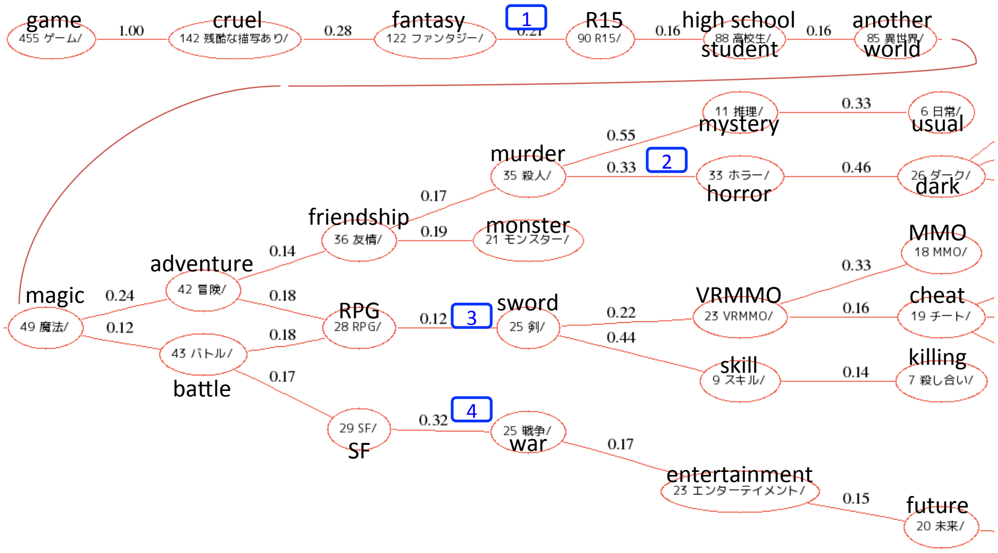


Figure 6. Concept graph of ‘game’