Information Engineering Express
International Institute of Applied Informatics
2017, Vol.3, No.2, P.45 - 57

Designing a Comic Exploration System Using a Hierarchical Topic Classification of Reviews

Byeongseon Park[*] , Kahori Okamoto[*],
Ryo Yamashita[†] , Mitsunori Matsushita[*]

Abstract

The purpose of our research is information access support based on the contents of a comic book. For this purpose, it is necessary to obtain information related to the story and the characters. In our previous research, we extracted the information using a review sentence and built a comic search system based on the extraction results. This system determined the relationship between comics using the TF-IDF method. However, the TF-IDF method cannot take into account the meaning of each word included in the review sentence. Therefore, this system may not be able to provide accurate results based on comic relationships. In this study, we analyze the review sentence using a hierarchical topic classification. On the other hand, the extracted topics often contain words that hinder the user from guessing the contents of the comic. Especially, the named entities defined in the comic tend to cause the problem. The amount of information obtained by the user varies greatly based on the user's knowledge about the named entities. Therefore, we investigated the influence of the named entity extracted from the comic reviews as the feature words. Our experiment revealed that the user's understandability was improved when the proportion of the named entities was decreased. Furthermore, we build an exploratory search system based on the topic.

1 Introduction

In Japan, the number of newly published comic titles is increasing every year with more than 12,000 titles published in 2015. A user who wants to select a new comic based on his/her taste can retrieve it through web services such as the search function of ComicCmoa and the search service provided by the Japan Society for Studies in Cartoons and Comics. Retrieval is based on bibliographic information (e.g., title and author), genre information (e.g., fantasy, humor, and sci-fi), and information obtained from other users (e.g., average preference, number of views). However, the web services are not suited for finding a comic that meets a user's taste because they do not take the contents of the comics into account.

To overcome this difficulty, we focus on the comics' reviews. Reviews of comics often contain both objective information (e.g., abstract of the story, episodes, and story settings)

[*] Kansai University, Osaka, Japan
[†] Nomura Research Institute, Ltd., Tokyo, Japan

and subjective information (e.g., likes/dislikes and construals derived from the story). In a previous study, we proposed a method to extract content information from a comic indirectly by using the comic's reviews[6, 9]. The proposed method evaluated comics more precisely than the conventional search services that only take into account genre information. In our previous study, we only utilized feature words for assessing the comics, which we obtained from the reviews using the term frequency-inverse document frequency (TF-IDF) method. We did not consider the meaning and variations of each word used in the review. This may lead to erroneous results when measuring the relationships between comics. Therefore, topics described in the reviews need to be determined more accurately, and the search system needs to provide results based on these topics. In this paper, we propose a method to determine the topics described in a review.

On the other hand, among the feature words extracted by the previous method[9] included many expressions based on unknown words defined in the comic, the characters that appeared in the comic, and the story of the comic are included. In this paper, these expressions are called named entities. The named entities tend to cause problems because the amount of information obtained by the user varies greatly depending on the user's knowledge about the named entities. For smooth search, it is desirable to present information that can capture the characteristics of the comic even if it is a comic that the user does not know. In this paper, we investigated the influence of the named entity being included in the feature word groups extracted from the review sentence on understanding the features of the comic. In addition, we propose a new search interface using hierarchical topics.

2 Related Work

Other research approaches also try to support the information access using content information of comics.

Iwama et al. proposed a comic access support system using comic information collected from DBpedia[1][8]. Comics that satisfy the user's needs, were explored and displayed in the form of a network, where a link connects two comics with the same information. A user can display a comic's information by selecting the link. By repeating the selection, the user could browse the related comics. This study solved the problem of existing comic search services, by retaining they had remained the keyword search by bibliographic information.

Okada et al. proposed a system for answering questions about the contents of novels[7]. The system guesses the question's intent from the words the user used to phrase the question, chooses a possible answer, and presents a part of the answer body. When a question arises while the user is reading a novel, he/she can use the system to answer the question.

In our study, we also aim to provide the information access based on the comics' content information as introduced by Iwama et al. For that, it is essential to understand the content information of each comic. When the content consists of a single modality like in the case of a text novel, it can be automatically processed using natural language processing technologies. Therefore, access support methods like the Okada's method are easy to achieve. However, processing the multi-modal content of comics is difficult. Processing comic contents requires a transcript of the content information or an indirect information acquisition from external information sources.

Iwama's approach leveraged the information available on Wikipedia. However, comic content included in this service consists only of bibliographic information, a short overview,

[1] http://www.dbpedia.org/

and information on the comic's characters. Therefore, extractable information is limited. In addition, Wikipedia articles about comics often do not contain much content. To acquire more information on the content of a comic, we need to extract it from other sources.

This study focuses on user reviews of comics, which we obtained from a review site that offers a vast collection of reviews and from blogs located on the website. We confirmed that the reviews contain the comic content information through an analysis of our previous work. Therefore, we can conclude that the reviews are a proper source for extracting the content information needed for this study[6].

3 Design Criteria

3.1 User Request for Comic Search

When a user accesses a comic that suits his/her taste using a search system, it is necessary to express his/her requirements and to convey them to the system. For requests concerning bibliographic information, the user can easily formulate a query, and the system will find an appropriate comic because the bibliographic information is available for every comic. However, content information such as the amount of information depicted in a comic (e.g., the number of background stories about a character) or the manner in which the story is told (e.g., narration, monolog) differs from comic to comic. In addition, the comics often utilize a technique that makes a reader guess the details of the contents by continuously arranging frames and story context. As a consequence, one reader's impression about the comic may differ from that of other readers. Therefore, there are no typical keyword patterns for a user's request for content information, and it is difficult to generate a query that represents his/her requirement appropriately.

There are many different forms of requirements for information about comics, from concrete requirements (e.g., a requirement to find a scene in a comic that she had read) to vague requirements (e.g., a requirement to find an unknown comic that suits her taste). Taylor[5] classified such requirements into four classes, namely "visceral need," "conscious need," "formalized need," and "compromised need." Based on the classification, concrete requirements about specific comics (e.g., "I want to read a new title written by Akira Toriyama") correspond to "formalized needs" or "compromised needs," and vague requirements about unspecified comics (e.g., "I want to read an interesting book") correspond to "visceral needs" or "conscious needs." If the comics' content information is structured, a user with a specific request can access the desired information easily. However, if the user only has a vague request, articulating a query is difficult, and the system cannot present the requested information immediately after receiving the query.

Based on the above considerations, this research intends to support the provision of appropriate information to users who present a vague request. With this support, a user's vague request will be clarified, and we expect his/her satisfaction from the search results to improve.

3.2 Supporting Information Access with Exploratory Search

As mentioned in Section 3.1, a user who only has a vague request faces difficulty expressing his/her preferences accurately. For such a user, an information access method that allows intuitive information access without the burden of generating a query will be suitable. To satisfy this requirement, we employ exploratory search, which is an information retrieval method to support users in articulating their thoughts and requests while accessing various types of information through a search [4]. The user clarifies his/her vague requirements incrementally and approaches the target information by conducting exploratory browsing and focused search repeatedly. Thus, this method supports users who only have vague information. In addition, exploratory search provides the following two advantages.

Satisfaction During an exploratory search, the user repeatedly evaluates the retrieved results and selects the one that best meets his/her requirement. To actively select the information, the user accesses the information along with his/her request and obtains satisfaction from the fact that it was his/her choice.

Serendipity By repeating the search, the user can access a number of comics. There is a high possibility that the user is presented with information about a genre that he/she has not read. Because the user might discover a new genre that he/she likes, we expect exploratory search to produce a high serendipity.

These two aspects are essential for a successful user support. Therefore, we adopt exploratory search for supporting information access, where the candidate information is presented to the user in response to his/her selection behavior and not displayed at the beginning of the search. This implies that the system cannot support the user, and, therefore, we need to support not only exploration but also query generation.

In our previous research, we employed "preferable title" as the input query to reduce the cost of query generation[9]. As mentioned in the previous section, query patterns that respond to content-related requests may vary. By limiting the patterns, we expect to reduce the cost of generating queries and, therefore, adopt a favorite comic title as the input query. Furthermore, comics are sometimes seriesized, and one work may be composed of multiple books. However, In the search in this system, the user aims to touch as many comics as possible. Therefore, in this paper, the search unit in the implementation system shall be the comic title.

4 Preparation

In this study, we use the reviews as the information source to characterize the contents of each comic. The reviews found on e-book and online shopping sites are not all positive but can also be negative and warn readers about purchasing the comic. We collected reviews from two Japanese sites, namely "MangaReview.com[2]" and "comic database[3]" for this research. We selected the top 1000 titles with more than 20 reviews and collected 70,639 reviews in total. As a result of collection, the maximum number of sentences of each review was 12,051("ONE PIECE", Eichiro Oda), the minimum was 63("EDEN〜It's an Endless World!〜", Hiroki Endo), and the average was 536.

4.1 Analysis Method

In our previous study, we adopted the TF-IDF value as a reference for assessing the similarity of content in comics. TF-IDF is an analytical method widely used as an index representing relative importance of words included in document data. TF-IDF is expressed by

[2] `http://www.manngareview.com/`

[3] `http://sakuhindb.com/`

the product of the frequency tf of words in one document and the inverse of df which is the frequency of words in all documents(see Equation 1).

$$\text{tf-idf} = \text{tf} \times \log \frac{N}{\text{df}} \tag{1}$$

In this case, $N/\text{df}(N$ is the total number of documents) is represented by a logarithm, which makes it possible to reduce the influence of words appearing in an extremely large number of documents when analyzing a huge number of documents. TF-IDF values represent the relative importance of words contained in a document. Hence, a word's frequency in a document will affect the TF-IDF value. Therefore, this approach might not provide an accurate content-based characterization of comics. To consider the meaning of words when analyzing a document, we need to use a topic model. In our previous research, we studied the Latent Dirichlet Allocation (LDA)[2] method for classifying comic reviews[6]. LDA is a popular statistical topic model that estimates a text's topics under the assumption that the text contains several topics. However, it was difficult to use the estimated topics for classifying the reviews. One reason why it was difficult was the vagueness of the main topic. For instance, the topics contained in newspaper articles and academic papers (e.g., international affairs, science) tend to be explicit. Therefore, the topics can easily be estimated and used for classification.

Reviews, on the other hand, do not necessarily represent contain a clear topic. In addition, in newspaper articles, the words used in the title are also likely to be used in the document. However, in the case of reviews, the topic is not usually described in the document. Taking this missing information into account, we need to analyze while estimating the subject information.

This paper utilizes a hierarchical LDA(hLDA) method[1], which is based on the LDA method. The hLDA is an unsupervised hierarchical topic model and assumes that the topics contained in the documents have a hierarchical structure. The hLDA method automatically joins the topics together in a hierarchy. Words contained in various topics tend to be classified as an upper layer, and the words that characterize each topic tend to be classified as a lower layer. By using this model, a more natural topic classification is possible.

In our previous studies [6], we confirmed the usefulness of the TF-IDF method for extracting a word that represents a feature of the comic title. Therefore, our research adopts this approach to extract the features of each comic title. When using the TF-IDF method, normalization is performed by dividing the TF value of each review by the number of reviews.

4.2 Analysis Result

As described above, reviews contain different perspectives such as impressions, opinions, and synopses. The system needs to extract the information comprehensively.

In this paper, we try to extract information by focusing on the nouns and adjectives that are part of the review statement. The nouns appear together with characteristic words such as "fantasy," "football," and "the Middle Age." From the words, we expect to determine the information about the world setting and the theme described in the comic. Moreover, adjectives in the review include words representing the state and comments such as "hot" and "scared." Thus, we can conclude that reviews are suitable resources for extracting the views of the reviewers, and we will use the nouns and adjectives in the comic reviews as

Table 1: Extended named entity class

Title	Words
"BOODY MONDAY"	hacker, Death Note, Third Eye, Brain game, betrayal, spy, Psychological wrafare, archery, terrorism
"Hyouge-mono"	tubulent age, unhappiness man, unhappiness, Wabi, chimney, history, dutifulness, china bowl, material desire, tea ceremony
"ONE PIECE"	Luffy, island, Chapter of Alabasta, Grand line, Usopp, Pirates, One piece, Fellows, Haki, Dragon ball

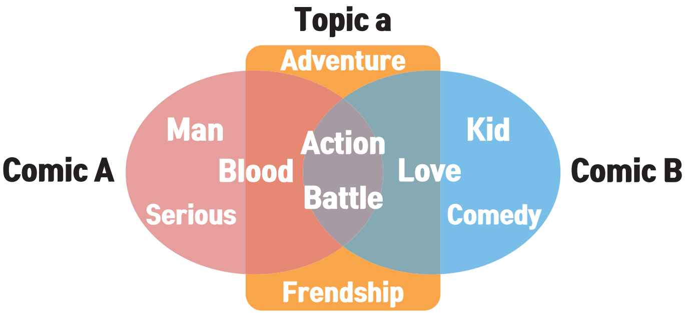


Figure 1: A related comic through topic

the source for our analysis. We do not consider the named entities such as the character's names because these words are not relevant for grasping the story of the comic.

TF-IDF method Figure 1 shows an example of words that have a high TF-IDF value. The TF-IDF method assesses words in the document quantitatively. The higher the value of a word, the better it characterizes the document. For instance, in "BLOODY MONDAY," words such as "hacker," "hacking," and "terrorism" were included. Based on the words, we can estimate that the theme of the comic title is "cyber crime." The estimation result allows the user to guess the comic's content better as compared to the genre assigned by the publisher. In this paper, we consider the 50 words of the comic title with highest TF-IDF as the "feature words."

hLDA method When conducting an analysis using the hLDA method, it is necessary to determine hyperparameters (α , γ , η) and the number of hierarchies before the analysis[1]. In this study, we examined several parameter settings and determined the most appropriate set of parameters, namely $\alpha = 1.0$, $\eta = 0.1$, $\gamma = 2.0$. The number of hierarchies was determined to be 3. We used the hLDA function developed by Mallet[4], a Java-based open source package for machine learning, to conduct a hierarchical topic analysis.

In Figure 1, we outline the usage policy of information obtained from the two types of analysis results mentioned above.

[4] http://mallet.cs.umass.edu/index.php

4.3 Named Entity in Comic Reviews

As mentioned above, in this research, we concluded that the named entity in the review sentence is information that can not be referred to for grasping the contents, and such part of speech is excluded from the analysis target in advance. However, it was confirmed that some named entities were classified as general nouns in the analysis results[9]. This is because the MeCab[5] morphological analysier was unable to analyze all the existing named entities during the comic review.

An column of " ONE PIECE " in Figure 1 is an example. Some of the feature words in " ONE PIECE " include general nouns such as "Pirates" and "Fellows." Even users who do not know the work can grasp the meaning of such words. Hence, at the time of retrieval, the user can infer the contents of the comic based on these words. Meanwhile, the feature words include named entities including character names such as "Luffy" and "Usopp" or such as terms relating to the story of comic such as "Alabasta" and "Haki". For users who have not read the comic, these words can hinder them from inferring the contents and grasping its meaning.

In order to solve these problems, it is desirable to register named entities in advance in a dictionary used for morphological analysis. Wikipedia[6], which is a multilingual encyclopedia on the worldwide web, can be cited as an information source that can collect named entities from comics. However, since it does not correspond to all comics subject to this system at present, it is difficult to solve using only Wikipedia. For the above reasons, in order to identify the named entities present in the review sentence, a method for automatically extracting them is necessary.

5 Extraction of named entities in comic review

the Conditional Random Fields (CRF)[10] algorithm was used for machine learning, in this study. CRF is an identification model for sequence labeling, and it is a method often used for morphological analysis and named entity extraction. In CRF,

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(|D|)}, y^{(|D|)})\} \tag{2}$$

When data is presented in this manner, the conditional probability P is represented by

$$\begin{aligned} P(y|x) &= \frac{1}{Z_{x,w}} exp(\omega \cdot \phi(x,y)) && (3) \\ Z_{x,w} &= \sum_{y} exp(\omega \cdot \phi(x,y)) && (4) \end{aligned}$$

(x : series, y: a label string given to the series, w: weight vector, $P(y|x)$: feature vector)

In addtion, we used a named entity class defined by Information Retrieval and Extraction Exercise (IREX) as a named entity class to be attached to the corpus. In the Japanese named entity extraction task of IREX, eight types of named entities are defined, and the named entities are not overlapped. However, Sekine et al. pointed out that with the widening spread of adaptation fields of information extraction and text-based answers to questions, the small number of existing classifications is insufficient and it is necessary to ac-

[5] http://taku910.github.io/mecab/

[6] https://www.wikipedia.org/

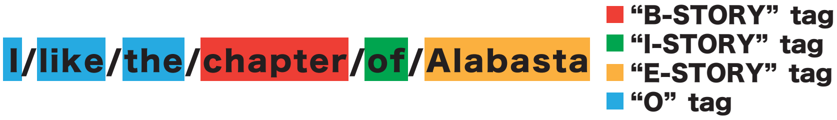


Figure 2: An example of tagging reviews of comics

count for a larger number of named entity types Then, we proposed an extended named entity which greatly increases the named entity type to 200 kinds.

In this paper, in order to smoothly extract unique proprietary expressions used in comic reviews, which have a description pattern different from existing named entity classes defined in IREX, three extension classes - STORY, ITEM, SKILL - were established. STORY is a word indicating a specific episode or story(eg.,"アラバスタ編 (Chapter of Alabsta)" や "開戦の章 (Chapter of war)"), ITEM is a word for a tool or an object represented in the comic(eg.,"どこでもドア (Anywhere Door)" や "ゴムゴムの実 (Rubber rubber fruit)"), and SKILL is a word depicting a technique possessed by a character, such as those seen in boys' mangas(eg.,"北斗百烈拳 (Hokuto hakuretsuken)" や "かめはめ波 (Kamehameha)").

In addition, we adopted "SE (Start / End) format"[12] as the chunk tag indicating the start and end positions of tags when tagging. In the SE format, B denotes the start position of the named entity tag, I denotes the inside of the named entity tag, and E denotes the ending position of the named entity tag. S is used when a single word itself is a named entity.

5.1 Model Implementation

From the following we will describe the procedure that we did to create a corpus for machine learning in this paper. First of all, we collected words existing on the Wikipedia page of 20 collected comic books, and created a named entity dictionary for each comic. At this time, a named entity class is assigned to each word based on the adopted named entity class. Next, review sentences were collected from "MangaReview.com" and "Comic database", and morphological analysis was applied to each sentence. Finally, we created a corpus tagging for each morpheme(y in Equation 2), based on the named entity dictionary of each comic2. There are five features of this learninig(x in Equation 2): (1) Surface, (2) Part-of-speech classification, (3) Character type, (4) The surface layer of the morpheme at the current position i to $i \pm 1 \sim 2$ position and the part of speech.

We used Python (version 3.5.1) for implementation, MeCab Python library for morphological analysis, and CRFsuite[7] Python library (Version 0.8.4) for machine learning. As a result of evaluating the named entity extractor adopted, the Precison rate was 0.788, the Recall rate was 0.458, and the F value was 0.526. The Precison rate is the proportion of those that are actually positive among the data predicted positive and the recall rate is the proportion of those that were predicted to be positive among those that are actually positive. In addition, the F value refers to the evaluation index of the prediction accuracy of machine learning as seen from Precision and Recall.

[7] http://www.chokkan.org/software/index.html.en/

6 Experiment on Influence of Named Entity in Feature Words and Topics

In the previous research[9], the words included in the topic were generated using the characteristic word of each comic extracted using the TF-IDF method and the topic generated using the hLDA method. When the user searches for a comic using the comic search system that contained words that were not recognized by the system, it was a burden on the user to guess the contents of the comic (see Figure 1). Especially, the named entity is not suitable as the information presented to the user because the amount of information obtained by the user varies extensively along with the user's knowledge about the named entities. On the other hand, according to the experiment conducted in the previous study, it was confirmed that the name of a work of a famous work and the name of a famous character exist as a word that was helpful when guessing the contents of a comic that the user does not know[9]. In other words, it is thought that the well-known named entity has the effect of making the image to the work more specific, such as "Comic similar to ○○" or "Character similar to □□." Therefore, in this paper, three conditions are set in order to grasp the influence of the named entity on the user.

Condition (1) Previous information

Condition (2) Information excluding all the named entities extracted by the named entity extractor

Condition (3) Information containing only *well-known word* among the named entities extracted by the named entity extractor

Well-known word in this experiment was determined to have an IDF value calculated from the review using the TF-IDF method of 6.5 or higher. We set a word as well-known word when the IDF value calculated using the TF-IDF method from the review was 6.0 or less in this experiment.

In addition, In the comic exploratory system, comics are related based on co-occurrence of each comic's feature words and words classified in each topic(see Figure 1). If it is not possible to accurately measure the relation of the comic via this topic, information different from the user's thought at comic search will be presented, making smooth searching difficult. Therefore, in this experiment, by comparing three conditions using named entities in different ways, when the user guesses the relation (topic) between comics defined in the system, the influence of the named entity. Therefore, in this experiment, by comparing the above three conditions, we confirm the influence of the named entity on the user when guessing the relation between comics.

6.1 Experimental Procedure

The experiment collaborators are college students belonging to the laboratory belonging to the author and 10 graduate students (4 males, 6 females). In this experiment, firstly, the purpose of the experiment and the contents of the task were explained to the experiment collaborators. The contents of the task of this experiment are (1) selection of comics judged to be related to the presented topic, and (2) selection of characteristic words related to topics included in the presented topic. At the same time, we told that there is no limit on the number of words that can be selected. In addition, the authors showed examples of

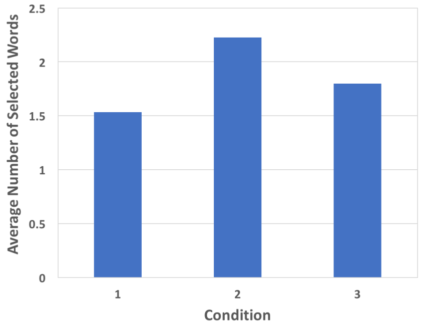


Figure 3: Average number of selected words per condition

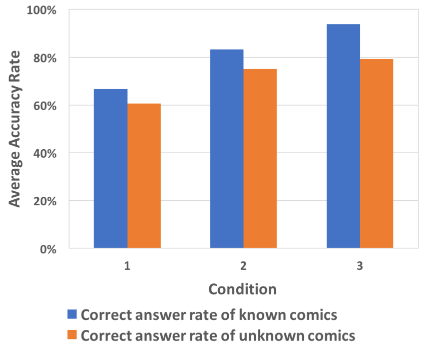


Figure 4: Correct answer rate based on awareness for each condition

responses using examples so that experiment collaborators can work on tasks smoothly. After informing the subjects that the experiment can be canceled by their own intention, we got consent from participating experiments after confirming the understanding of the explanation so far.

The subject to be tackled by experiment collaborators included four studies on one topic as depicted in the example and this time we asked for nine topics in all, including three topics for each condition. In addition, after having the problem addressed, if there were points that the subjects were aware of when making answers, they were asked to reply without any constraints in terms of the description. Finally, we had three comments(1. I have read, 2. I never read, 3. I have never read but I know) on the comic used in the experiment. In this experiment, no response time was set and the experiment was terminated at the time when all the subjects were addressed.

6.2 Experimental Result

Figures 3 and 4 show the average values of the response results of experimental collaborators for all items of each condition. Figure 3 shows the average number of words, which is the basis for estimating the relation between the topic and the comic, and Figure 4 shows the average correct answer rate for the relation between the topic and the comic in the recognition. From Figures 3 and 4, it can be confirmed that the number of words for which the conditions (2) and (3) are referenced from the condition (1) are many and the correct answer rate is also high. Particularly, the condition (3) shows the correct answer rate of 79% against "comics the user did not know." Furthermore, as a result of Welch's t-test verification, a significant difference was observed between Condition 1 and Condition 2($t(15) = 0.044$, $p < .05$), and Condition 1 and Condition 3($t(18) = 0.042$, $p < .05$) in the correct answer rate for comics unknown to experiment participants. However, there was no difference between Condition 2 and Condition 3($t(14) = 0.567$, $p < .05$).

7 System Implementation

Figures 5 - 7 shows our proposed system. With this system, we can start the exploration by input title of the preferable work. When the user enters his/her favorite work title into the system, the main search screen (see Figure 5) appears. The user can explore comics on the screen in an exploratory manner. A central comic is a "selected comic" (see A in Figure 5). The selected comic is an input title of the comic. It has a feature word of "selected comic"

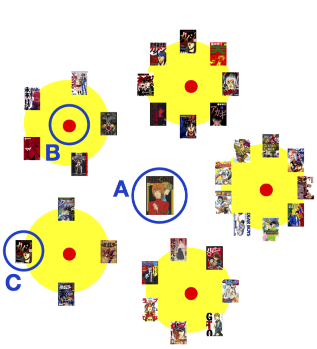
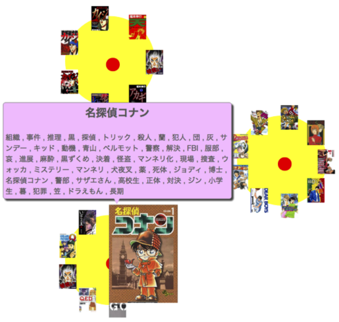


Figure 5: Search screen



名探偵コナン

組織 , 事件 , 推理 , 黒 , 探偵 , トリック , 殺人 , 蘭 , 犯人 , 団 , 灰 , サンデー , キッド , 動機 , 青山 , ベルモット , 警察 , 解決 , FBI , 服部 , 哀 , 進展 , 麻酔 , 黒ずくめ , 決着 , 怪盗 , マンネリ化 , 現場 , 捜査 , ウォッカ , ミステリー , マンネリ , 犬夜叉 , 薬 , 死体 , ジョディ , 博士 , 名探偵コナン , 警部 , サザエさん , 高校生 , 正体 , 対決 , ジン , 小学生 , 暮 , 犯罪 , 笠 , ドラえもん , 長期

Figure 6: Search screen at the time of mouseover

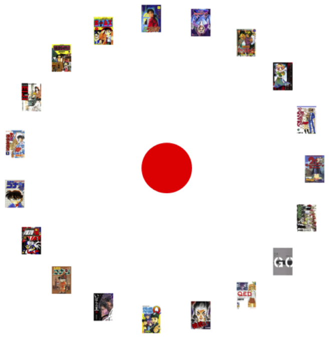


Figure 7: Search screen at the time of topic selection

and a “related topic” (see B in Figure 5). The related topic is a topic classified based on cooccurring words have classified into each topic. A method for estimating a topic was mentioned in Section 4.2. In addition, a “related comic” (see C in Figure 5) is presented to the surroundings. The related comic is a comic that has a feature word co-occurring in each related topic. Figure 5 is a representation of the result obtained from the following conditions: “related topics in which there is more than co-occurring word” and “comics in which there are more than three co-occurring words.” A presentation of the selected comic and the related comic is used as a cover for each comic. The picture can be selected when the comic is selected[9]. Presenting large amounts of information increases the burden on the user when selecting the information because for every repeated information seeking he/she needs to make a selection from the presented information[3]. To reduce the burden on the user, our proposed system limits the amount of presented information. We set the upper limit of the presented topics to six, the upper limit of the related comics to each topic to ten, and the implemented system features to three.

Explore function When the related comic is selected, it moves to the selected comic section. The related topic and the related comic are expanded. When we click the related topic, all the related comics that include the topic are presented (see Figure 7). This function allows exploring the starting point of the topic. When we select the related comic, the state returns to Figure 5.

Browse the detailed information function We expand the presented comic cover by using the mouse-over (see Figure 6). At the same time, we can view the comic’s feature words. The related topics can be viewed by using the mouse-over.

History function The selected comic is presented at the bottom of Figure 5 and can be a reference for comic selection. We can select the comic by clicking it and the comic search starts again.

By repeating these three acts, he/she can access to various comic information.

8 Discussion

Our proposed system presents information using two types of screens as shown in Figures 5, 6, and 7. In Figure 5, we present the related topic of different contents and the comic information to represent the selected comic. Presenting this information helps a user with a vague request to clarify the request, i.e., if it fits into the framework of Exploratory Search, he/she can perform a search conforming to Exploratory Browsing.

On the other hand, in Figure 7, we present the related comic as a starting point for the topic of interest. An action of selecting one of the topics is presumed to be a state in which the search purpose was settled because the user selects from a wide range of topics. Therefore, this search action corresponds to Focused Searching. The goal of this study is to select a comic for a user who has only a vague request using our proposed system. We achieve this by helping the user to clarify his/her requirements and preferences. Our proposed system conforms to the Exploratory Search framework and achieves the set goal. In the future, we will do user reviews and evaluate the system.

In addtion, Experiments in Section 6 confirmed that it is easy to predict the contents of comics with feature words whose proportion of named entities is smaller than feature words in which named entities are mixed. Furthermore, it was confirmed that the content inference with a reduced proportion of named entity is more significant than existing ones even when the subject does not know the comic.

On the other hand, however, there was no significant difference between Condition 2 and Condition 3 established in this experiment. The cause of this is the performance of the named entity extractor created in this paper. As a result of extracting using the named entity extractor proposed in this research. As a result of the extraction using the proposed named entity extractor, there is a difference between the number of substantive named entities when Condition 1 and Condition 2 are compared. There is a large difference from Condition 3 in which extraction is performed with IDF value as the reference. Further research is needed on this subject to improve the precision and versatility of the extractor, such as re-learning based on extraction results and adopting new corpus.

9 Conclusion

This paper proposed a search system that intends to support exploratory information access based on comic content information. The system requires information for characterizing each comic and provides the relations among the comics to the user. We extracted the information from the comics' reviews using the TF-IDF method. To assess the relationships between comics, we also conducted a hierarchical topic classification of the reviews using an hLDA topic model. In addtion, we investigated the influence of the named entity extracted from the comic reviews as the feature words. Our experiment revealed that the user's understandability was improved when the proportion of the named entities was decreased.

The proposed system made it possible to connect the comic with others through each topic. The system presents a variety of information about the comic and reveals the user's request. The user needs to enter his/her request to start the exploratory search by the system. For further research, we plan to study a new information presentation method based on the results of the experiment about the named entity. In addition, we will conduct a usability test of the system design to determine how suitable it is for the target user.

Acknowledgement

This research is supported by JSPS KAKENHI Grant Numbers JP15K12103.

References

[1] David. M. Blei, Thomas. L. Griffiths, and Michael. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, Vol. 57, No. 2, 2010, Article 7.

[2] David. M. Blei, Andrew. Y. Ng, and Michael. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993–1022.

[3] Sheena. S. Iyenger and Mark. R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of Presonality and Social Psychology*, Vol. 79, No. 6, 2000, pp. 995–1996.

[4] White. W. Ryen and Roth A. Resa. Exploratory search: Beyond the query-response paradigm. *Morgan & Claypool Publishers*, 2009.

[5] Robert. S. Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, Vol. 29, No. 3, 1968, pp. 178–194.

[6] Yamashita, R. and Matsushita, M. Content discrimination of comics based on users' reviews. *The Third Asian Conference ofn Information Systems*, 2014, pp.79–85.

[7] Okada, S. and Arakawa, T. A proposal on support system for reading of novels using question answering technology. *Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 27, No. 2, 2015, pp. 608–615, in Japanese.

[8] Iwama, Y. and Mihara, T. and Nagamori, M. and Sugimoto, S. Facet-based visualization of a Manga collection based on ontology and LOD resources. *IEICE Human Communication Group Symposium*, 2014, pp. 357–361, in Japanese.

[9] Yamashita, R. and Park, B. and Matsushita, M. Supporting exploratory information access based on comic content information. *Transaction of the Japanese Society for Artifical Intelligence*, Vol. 32, No. 1, 2017, pp. WII–D_1–11.

[10] Lafferty, J., McCallum, A. and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 282–289.

[11] Sekine, S., Sudo, K. and Nobata, C. Extended named entity hierarchy *Proceedings ofthe 3rd International Conference on Language Re-sources and Evaluation*, 2002, pp. 1818–1824.

[12] Kiyotaka, U., Qing, M., Masaki, M., Hiromi, O. and Hitoshi, I. Named Entity extraction based on a maximum entropy model and transformation rules. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 326–335, 2000.