Exploratory Search System Based on Comic Content Information Using a Hierarchical Topic Classification

Ryo Yamashita [†] **Kahori Okamoto**[†] **Mitsunori Matsushita**[†]
[†]Graduate School of Informatics, Kansai University
{k809231, k317680, m_mat}@kansai-u.ac.jp

Abstract

The purpose of our research is to provide 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 does not include the meaning of each word used in the review sentence. Therefore, this system may not be able to provide accurate comic relationships. In this study, we analyze the review sentence using a hierarchical topic classification. If we can extract the topic from the review sentence, it is possible to establish a relationship between comics based on the comic content information. Furthermore, we built an exploratory search system based on the topic.

Keywords: comic computing, latent Dirichlet allocation, review sentence, term frequency-inverse document frequency

1 Introduction

In Japan, the number of newly published comic titles is increasing every year with more than 12,000 titles published in 2014. 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[1] and the search service provided by the Japan Society for Studies in Cartoons and Comics[2]. 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 viewing). 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 let the search system to provide results that reflect the taste of the user better, we need to extract content-related information from the comics. It is, however, difficult to acquire such information from all the contents of a comic. A comic's content consists of several modalities such as images, texts, and symbols to express stories in a complementary style, and current image processing techniques are not sufficient for extracting content information precisely.

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) and subjective information (e.g., like/dislike and construals derived from the story). In a previous study, we proposed a method to extract content information of a comic indirectly by using the comic's reviews [8]. The proposed method discriminated comics more precisely than the conventional search services that only consider genre information.

In our previous study, we only utilized feature words for discriminating the comics, which we obtained from the reviews using the 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. The method applies a hierarchical topic classification method to the review and presents a new search interface that utilizes the determined topics.

2 Related Works

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

[1] http://www.cmoa.jp/

[2] http://ww1.yes.ne.jp/akitanet/shosekikensaku.htm

comics.

Iwama *et al.* proposed a comic access support system using comic information corrected from DBpedia[3] [3]. Comics, which 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, namely that they had remained the keyword search by bibliographic information.

Okada *et al.* proposed a system for answering questions about the contents of novels [5]. 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 support 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, 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 [8].

3 Design Criteria

This section presents the design criteria for developing the content-based search system for comics.

3.1 Target Users

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 adequate 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, monologue) 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, the 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 adequately.

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 to her taste). Taylor [7] 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."

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

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 consideration, this research intends to support information access of users who only have a vague request. Through the support, a user's vague request will be clear, and we expect his/her satisfaction 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.

Exploratory search is an information retrieval method to support users in understanding their thoughts and requests while accessing various types of information through a search [6]. 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 for 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.

4 Prepare for Building the System

4.1 Collecting Data

In this study, we use 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. However, such reviews express personal opinions and are therefore not usable for our purposes. For our system, we need to extract content information from the reviews and need to select sites that are not in the business of selling comics. Comic reviews are not only posted on review sites but also on blog sites and social networking services (SNS). In addition, one article may contain more than one comic review. We need to consider that the more articles and reviews we collect, the higher the cost for their classification.

Based on the above consideration, this paper collects reviews from two Japanese sites, namely "MangaReview.com[4]" and "comic database[5]." We selected the top 1000 titles with more than 20 reviews and collected 70,639 reviews in total.

[4] http://www.manngareview.com/

[5] http://sakuhindb.com/

4.2 Analysis Method

In our previous study, we adopted the TF-IDF value as a reference for measuring the content similarity of comics. 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] for classifying comic reviews [8]. 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 contain a clear topic. In addition, in newspaper articles, the words used in the title are likely to be also 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 Latent Dirichlet Allocation (hLDA) method [1] that 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 [8], 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.

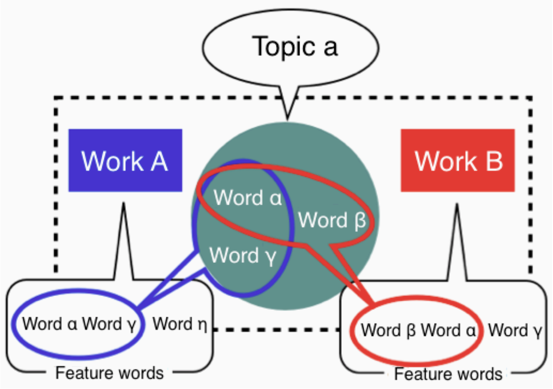


Figure 1. A related comic through topic a.

4.3 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 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. In the following, we present the analysis results of the TF-IDF method and hLDA method mentioned in Section 4.2.

- TF-IDF method

  Table 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

Table 1. TF-IDF value of the top ten words of each comic

BLOODY MONDAY	へうげもの **(Hyouge-mono)**	金色のガッシュ!! **(Zatch Bell!)**
ハッカー (hacker)	戦国 (turbulent age)	ガッシュ (Zatch)
デスノート (Death Note)	数奇者 (unhappiness man)	ファウード編 (Faudo edition)
ハッキング (hacking)	数奇 (unhappiness)	魔物 (goblin)
サードアイ (Third Eye)	わび (Wabi)	麿 (Maro)
頭脳戦 (brain game)	ほや (chimney)	ゼオン (Zeon)
裏切り (betrayal)	歴史 (history)	クリア (clear)
スパイ (spy)	忠実 (dutifulness)	魔界 (hell)
心理戦 (psychological warfare)	茶碗 (china bowl)	魔物編 (goblin edition)
アーチェリー (archery)	物欲 (material desire)	ブラゴ (Brago)
テロ (terrorism)	茶道 (tea ceremony)	雷句 (Raiku)

Table 2. example of topics classified in each layer with hLDA

Topic	**Words**
topic 1	の (of)，良い (good)，作品 (work)，漫画 (comic)，評価 (evaluation)，点 (point)，ない (no)，こと (thing)，よう (Yo)，面白い (interesting)，ん (n)，悪い (bad)，キャラ (character)，話 (story)，さ (sa)，いい (good)，方 (direction)，好き (like)，絵 (picture)，もの (thing)
topic 2	の (of)，良い (good)，絵 (picture)，バトル (battle)，点 (point)，キャラ (character)，悪い (bad)，評価 (evaluation)，設定 (setting), 総合 (total)，主人公 (hero)，漫画 (comic)，展開 (unfoldment)，ん (n)，戦闘 (fighting)，ない (no)，気 (ki)，敵 (enemy)，シーン (scene)，ストーリー (story)
topic 3-1	音楽 (music)，漫画 (comic)，熱い (hot)，主人公 (hero)，だめ (no)，試合 (match)，作品 (work)，サッカー (soccer)，チーム (team)，天才 (genious)，音 (sound)，人 (people)，才能 (talent)，の (no)，千秋 (Chiaki)，監督 (foreman)，マンガ (comic)，演奏 (play)，選手 (player)，さ (sa)
topic 3-2	三国志 (Sangokushi)，曹操 (Soso)，蜀 (Shoku)，演義 (Engi)，羽 (wing)，人物 (person)，呂 (Ro)，作品 (work)，さ (sa)，備 (Bi)，魏 (Gi)，武将 (military commander)，最後 (last)，葛 (Shoku)，歴史 (history)，本作 (this)，布 (fabric)，飛 (fly)，正史 (correct history)，魅力的 (seductive)

of the comic title is “cyber crime.” The estimation result allows the user to guess the comic’s content better 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 $\alpha = 10.0$, $\gamma = 100.0$, and $\eta = 0.01$. The number of hierarchies was determined to be 3. We used the hLDA function provided by Mallet[6], a Java-based open source package for machine learning, to conduct a hierarchical topic analysis. Table 2 shows an example of topics classified in each layer. As mentioned in Section 4.2, the topics in the upper layer contain the lower topics. For instance, the topic 2 in the second layer contains topic 3-1 and topic 3-2. Words that may be included in some topics (e.g., “character,” “interesting,” “story”) tend to be categorized into topics of the upper layer. On the other hand, words included in the classified topics of the lower layers tend to be characteristic words appearing in each topic. For example, hierarchy 3-2 contains words, such as “mystery,” “incident,” and “Thief.” We can estimate from these words that the topic will be the “detective story” topic. If a topic contains many abstract and ambiguous words, it is difficult to guess the reason why the topic is classified. Since the ambiguous words are not suitable to estimate the contents of the comic, the topics employed by the proposed system must be clear.

  Based on the above discussion, we employed topics that were classified in the third level (# of topic: 64). In Figure 1, we

[6] http://mallet.cs.umass.edu/

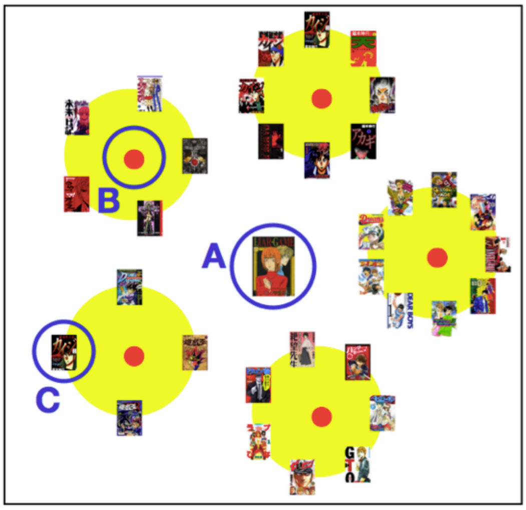
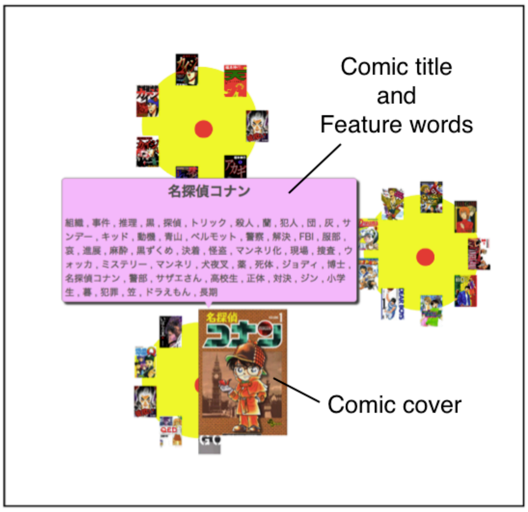


Figure 2. Search screen.



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 3. Information presentation screen at the time of mouseover.

outline the usage policy of information obtained from the two types of analysis results mentioned above.

5 Implementation

Figure 2-4 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 2) appears. The user can explore comics on the screen in an exploratory manner.

A central comic is a “selected comic” (see A in Figure 2). The selected comic is an input title of the comic. It has a feature word of “selected comic” and a “related topic” (see B in Figure 2). The related topic is a topic that co-occurring words have classified into each topic. A method for estimating a topic was mentioned in Section 4.3. In addition, a “related comic” (see C in Figure 2) is presented to the surroundings. The related comic is a comic that has a feature word co-occurring in each related topic. Figure 2 is the result of setting the two conditions. The conditions are “related topic that co-occurrence words are included in more than one word” and “comic that co-occurrence words are included in more than three 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].

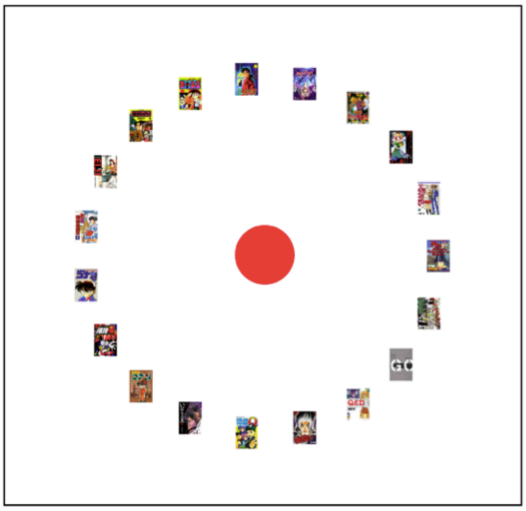


Figure 4. Search screen at the time of topic selection.

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 [4]. 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 4). This function allows exploring the starting point of the topic. When we select the related comic, the state returns to Figure 2.

- Browse the detailed information function

  We expand the presented comic cover by mouseover (see Figure 3). At the same time, we can view the comic's feature words. The related topics can be viewed by mouseover.

- History function

  The selected comic is presented at the bottom of Figure 2 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.

6 Discussion

Our proposed system presents information using two types of screens as shown in Figures 2 and 4. In Figure 2, 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 4, we present the related comic as a starting point for the interested topic. 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.

7 Conclusion

In this paper, we proposed a search system that intends to support an exploratory information access based on the 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.

The proposed system made it possible to connect the comic with others through each topic. The system presents various comic information and reveals the user's request. The user needs to enter his/her request to start the exploratory search by the system. As future work, we plan to identify more appropriate analysis parameters using the hLDA method by performing an experiment to compare the results of various parameters. 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 Grant-in-Aid for Challenging Exploratory Research (15K12103).

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, Issue. 2, Article No. 7, 2010.

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

[3] Yusuke Iwama and Tetsuya Mihara and Mitsuharu Nagamori and Shigeo Sugimoto. Facet-based visualization of a Manga collection based on ontology and LOD resources. *IEICE Human Communication Group Symposium 2014*, pp. 357–361, 2014. in Japanese.

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

[5] Satoru Okada and Tatsuya Arakawa. 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, pp. 608–615, 2015. in Japanese.

[6] White. W. Ryen and Roth A. Resa. *Exploratory Search—Beyond the Query-Response Paradigm*, Morgan & Claypool Publishers, 2009.

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

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

[9] Ryo Yamashita and Mitsunori Matsushita. Supporting exploratory search in Comic books using genre information. *The 29th Annual Conference of the Japanese Society for Artificial Intelligence*, 1H2-3in, 2015. in Japanese.