Content Discrimination of Comics Based on Users' Reviews

Ryo Yamashita[†] **Mitsunori Matsushita**[‡]
[†]Graduate School of Informatics, Kansai University
[‡]Faculty of Informatics, Kansai University
k809231@kansai-u.ac.jp, mat@res.kutc.kansai-u.ac.jp

Abstract

The objective of our research is to develop a search system to support the exploration of comics. To achieve this objective, it is necessary to obtain content information from comics, such as the world setting and characters. Consequently, we propose a method that obtains content information about comics from their reviews on the Web. Specifically, the proposed method determines a set of keywords that reflect the content of target comics in an appropriate manner and visualizes the relationships between comics by connecting them using their common keywords. Term frequency-inverted document frequency (TF-IDF) and latent Dirichlet allocation (LDA) algorithms are used to determine these keywords. TF-IDF determines keywords that reflect *explicit* information in the reviews, while LDA determines keywords that reflect *implicit* underlying topics in the comics. The results of two evaluations conducted to confirm the performance of these algorithms demonstrate that TF-IDF helps to obtain meaningful keyword sets, while LDA currently produces relatively unclear keyword sets.

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

1 Introduction

In Japan, over 12,000 new comic book titles appear every year. People who read these comics often communicate with other readers online by providing outlines or comments. Social networking services (SNSs) such as Twitter and Facebook are examples of the forums where this communication occurs. We can acquire reputation and outline information (e.g., the story and subjective recommendations) about a comic from such forums. In the case of unread comics, for instance, the reputation information obtained from such media can be used as a barometer when the user judges whether the comic is worth reading or not. However, much of the information disseminated by these media tends to be related to the newest comics and comics that have attracted attention. Therefore, these methods might not provide comic readers with information that is suited to their tastes. Search services are available for comics, but executing a search that focuses on the contents of comics is difficult.

To address this problem, we developed a search system based on comic contents that helps users to obtain information about unread comics intuitively and efficiently. In the present study, comic content information (such as story and character information) obtained from the Web was analyzed and the results visualized. Consequently, we propose a technique that allows the contents of comics to be understood by visualization.

2 Research focus and difficulty searching for comics

2.1 Difficulty searching for comics

In order to develop a system to search for comic contents, the extraction of comic content information is desirable. However, it is difficult to acquire such information from the entire contents of a comic because the contents of comics are multi-modal (e.g., illustration information and text information). Of course, it is possible to extract sentences about dialogue and to discern the appearance of a specific character in a comic, but it is difficult to obtain story information from this information. Therefore, it is necessary to examine the techniques used to obtain comic content information.

In existing comic search services, content information (e.g., a battle or love story) is entered manually to facilitate searches based on comic content information. As stated above, a large number of comics exist and extended periods of time are required to assess the information from each comic. In addition, detailed information is

not available for each comic because it is entered manually. Thus, it is difficult to differentiate each comic. If comics cannot be differentiated, the search results obtained will be vast. As a result, it is difficult to search for comics that correspond to one's tastes. For example, Cmoa, a comic search site (http://www.cmoa.jp/ (last accessed September 2014)), presented 4,638 comic titles for the search term "Love Story" and the user has to search for comics that suit their tastes among these results. To solve this problem, it is necessary to provide information that can differentiate each comic. Thus, obtaining comic content information is desirable to facilitate comic differentiation. To address this issue, the following section considers the techniques used to extract comic content information.

2.2 Selection of information to extract

As mentioned in Section 1, the market for comics is expanding in Japan and the number of comics is also increasing. Thus, it is preferable to obtain the content information automatically from the vast resources available. Our research is focused on the extraction of relevant information from the Web.

In the present study, the aim is to find comics that are suitable for readers. The system needs to assess the contents of the comic as a whole, which we achieve by extracting a desirable comic title unit. In addition, the stories in comics may contain two or more elements (e.g., world setting, characters, and items) and the preferences of reader may differ. Therefore, to obtain matches with the broad tastes of readers, it is desirable to obtain information from two or more viewpoints for each comic.

To meet this requirement, we obtained information from online review sites where users describe their impressions of each comic title. The information for each comic title can be extracted as a comic title unit. Since a review found on a review site is articulated for every target comic, we can obtain the information for each comic. In general, two or more people are likely to review each comic; hence, the comic information characteristics obtained are from two or more viewpoints.

A user who describes a review sentence (i.e., writer) is not the same as a user who interprets the review sentence (i.e., reader), which means that the writer's review sentences might not be interpreted correctly by the reader. To avoid this incompatibility, it is necessary to examine how a review sentence should be analyzed. A review sentence is described by the user from various viewpoints. If the viewpoints differ, the same text may also have a different meaning. The meaning of a text cannot be grasped based only on the explicit information in the text. Thus, we also consider the implicit information in a text in order to understand its meaning. Section 3 describes the analysis method, which uses both explicit and implicit information.

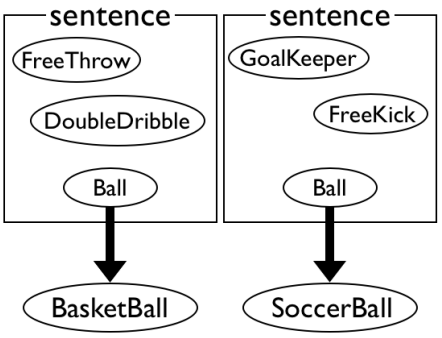

Figure 1. The diagram of implicit information

3 Analytical method

3.1 Focus of the analysis

As described in the previous section, comic content information is obtained based on the explicit and implicit information extracted from review sentences related to comics, resulting in a wider understanding of the information content of comics.

Explicit information is information obtained from each characteristic word in a text. Using explicit information, we can estimate how readers feel about specific comics. For example, if a review sentence contains the word "basketball," it is assumed that "basketball" is a theme in the comic. In addition, the contents that are currently considered by each user also differ. Therefore, two or more types of explicit information can be obtained from each comic review. After combining the explicit information extracted from a review sentence, it is possible to estimate the contents of the comic. For example, if words such as "disabled person" and "wheelchair" are present in addition to "basketball," it can be assumed that "disabled person and basketball" is a major theme, (i.e., "wheelchair basketball").

Implicit information might not appear in a text but the actual information in the text may refer to the same concept. An example is shown in Figure 1. We assume that the word "ball" appears in a review sentence. Based only on the word "ball," it is not possible to determine whether the focus is a "soccer ball" or a "basketball." However, if words such as "free kick" and "goalkeeper" also appear in the same sentence, it can be assumed that the ball is a "soccer ball." In addition, if a word that relates to "soccer" is present, it can be expected that another soccer-related term will also be present in the text. In our example, if the words "free kick" and "goalkeeper" appear in a comic review, it can be assumed that the comic discussed in the review is related to soccer. The advantage of using this information to generate topics related to a word is that the subject can be employed in addition to the actual meaning of the word.

In this paper, we focus on review sentences collected from manngareview.com (http://www.manngareview.com/ (last accessed September 2014)). At manngareview.com, a review is described based on the comic title unit. Thus, information can be collected for comic title units. In addition, mangareview.com requires the assessment of a target comic in 10 steps when completing a review. In this paper, the focus of the analysis comprises the comics with the top 150 average scores based on the evaluations, with reviews collected on February 25, 2014. Further, the number of reviews for each comic differed. Most of the reviews were for "SLAMDUNK," with 259. Six comic titles shared the lowest number of reviews with 10, including "がんばれ元気 (GANBARE GENKI)."

3.2 Analysis of explicit information

The term frequency-inverse document frequency (TF-IDF) method was used to analyze the explicit information included in the review sentences. TF-IDF provides an indicator of the relative importance of the words contained in a document. This indicator is the product of the reciprocal of the document frequency (DF) value, which represents the document frequency, and the term frequency (TF) value, which represents the frequency of words. In this study, the comics used in the analysis differed in terms of the number of reviews for each comic. The number of reviews was biased in terms of the amount of information they contained, which could have affected the results of the analysis. Therefore, we normalized the results by dividing the number of reviews for each comic by the TF value of each comic.

3.3 Analysis of implicit information

The analysis of the implicit information included in the review sentences used the latent Dirichlet allocation (LDA) method [1]. LDA classifies words in a document into topics by taking the meaning of the word into account. This method is a potential semantic analysis technique for expressing two or more topics contained in one document. Further, it can express the multinomial distribution of the topic that constitutes a document and the multinomial distribution of the word constituting each topic. The subject of a document can then be presumed from the constituted topic. The method has been applied to several tasks, such as multi-document summarization [2; 3] and recommendation of the information suitable for a user's taste [4]. The results of a previous work demonstrated that topics could be presumed from review sentences using this method. In all the experiments conducted using the LDA algorithm, we set beta = 0.01 and alpha = 50/K (where K is the number of topics that need to be estimated), with the sum of the Dirichlet hyperparameters remaining constant [5]. In addition, we used the Gibbs sampling method to estimate each topic, with the number of attempts set at 100.

4 Results and analysis

4.1 Explicit information

Table 1 shows the analysis of some of the results obtained using TF-IDF. The information characteristics of each comic are also described in Table 1. For example, it can be seen that the information characteristics of the comic "MOONLIGHT MILE" are "宇宙 (universe)," "開発 (development)," "月 (moon)," "ロボット (robot)," and "未来 (future)." In this analysis, the top 50 words of a TF-IDF value were defined as explicit information. However, words that were not related to the comic contents were also included in the information, which were defined as explicit information. For this reason, we excluded these words manually. The words that were not related

Table 1. Ten words and feature quantity with high feature value of each comic

7SEEDS		MOONLIGHT MILE		SF 全短篇 (SF all the short stories)	
Noun	Feature value	Noun	Feature value	Noun	Feature value
サバイバル (survival)	2.916	宇宙 (universe)	3.639	島 (island)	1.944
キリギリス (Katydid)	2.119	開発 (development)	1.354	皮肉 (irony)	1.639
夏 (summer)	1.889	中国 (China)	0.903	PERFECT 版 (PERFECT version)	1.541
チーム (team)	1.594	月 (moon)	0.804	絶滅 (extinction)	1.504
少女 (girl)	1.056	天空 (the sky)	0.733	侵略者 (aggression)	1.328
アリ (Ant)	0.779	征服 (conquest)	0.733	オジサン (a strange man)	1.328
冬 (winter)	0.664	～ (a wavy line)	0.733	モンスター (monster)	1.238
春 (spring)	0.664	SF	0.671	怪物語 (Monster words)	1.156
未来 (future)	0.639	ロボット (robot)	0.541	ガク (Gaku)	1.156
花 (flower)	0.624	過渡期 (a transition period)	0.539	日本語訳 (Japanese translation)	1.156

プラネテス (PLANETES)		銃夢-GUNNM-(GUNNM)	
Noun	Feature value	Noun	Feature value
宇宙 (universe)	1.624	ガリィ (Gally)	1.594
ハチ (Hachi)	1.203	サイボーグ (cyborg)	1.066
マキ (Maki)	0.978	クズ (iron filings)	0.711
愛 (love)	0.389	LO	0.683
フィー (Fee)	0.326	鉄町 (town of iron)	0.683
デビュー作 (Debut work)	0.300	SF	0.570
木星 (Jupiter)	0.260	一昔 (decode)	0.533
哲学的 (philosophical)	0.228	サイバーパンク (cyberpunk)	0.455
ロック (rock)	0.221	ザレム (Zarem)	0.455
未来 (future)	0.216	賞金 (prize)	0.455

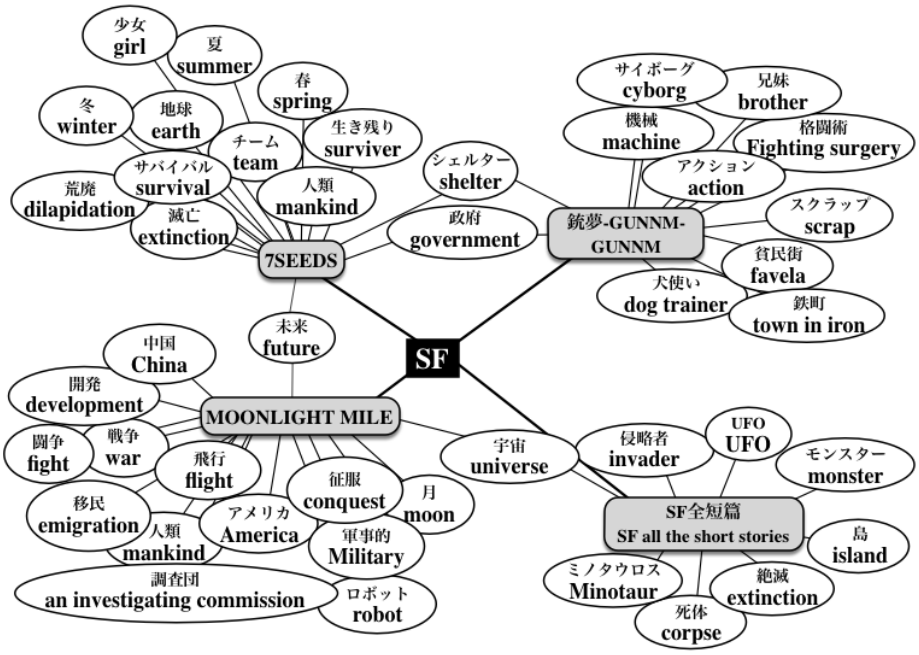

Figure 2. The graph linked to each comic explicit information

to the comic contents were classified into the following three types:

- A word related to a character

  The names of the characters that appeared in the comics, such as "ハチ (Hachi)," are shown in Table 1. The comic contents cannot be determined even if the names of the comic characters are shown.

- A word that cannot be used to determine the contents of the comic

  Words that did not provide good estimates of the comic contents such as " (a wavy line)," "過渡期 (a transitional period)," and "哲学的 (philosophical)." If this type of information was included in the system, it would prevent users from determining the contents of the comic.

- A word used in a bibliography

Words used in a bibliography, such as “デビュー作 (debut work)” and “PERFECT 版 (PERFECT version).” These words do not convey explicit information related to the comic contents.

Figure 2 provides a visualization of the explicit information found in four comics (“7 SEEDS,” “MOONLIGHT MILE,” “SF 全短篇 (SF all the short stories),” and “銃夢-GUNNM (GUNNM)”) among the comics that contained explicit information related to “SF.” The comic contents can be determined using the explicit information connected to each comic, as shown in Figure 2. For example, it can be determined that “7 SEEDS” is “the tale of a girl who risked survival on a ruined Earth” based on the explicit information. In addition, the contents of each comic can be determined and compared with the contents of other comics. For example, “MOONLIGHT MILE” may be estimated as a “story where the United States and China fight for possession of the moon” based on the explicit information. If this estimate is compared with that for “7SEEDS,” it is possible that the contents of “MOONLIGHT MILE” might be more suitable. There are a few tags in the existing comic search service. Therefore, performing a genre search using the tag information “SF” would place these two comics in the same list of search results. However, we found that the contents of each comic differed significantly based on our analysis using the explicit information from a review sentence. Therefore, the use of explicit information rather than tag information is more efficient and intuitive. Our method also provides greater detail than the tag information used in existing comic search services, thereby making it easier to understand the contents of comics.

4.2 Implicit information

Table 2 shows examples of the results obtained using LDA, which is an unsupervised learning method. In unsupervised learning, it is necessary to determine subjective classifications based on the results. However, it was difficult to specify each topic based on the results obtained in the present study. The words identified were ranked probabilistically for each topic, which made it difficult to assess each topic. Next, we analyzed the types of topics found in a comic review for “SLAMDUNK,” a basketball comic. The results of this analysis are shown in Table 3. For example, topic 15 suggests that this comic describes a “story where friendship is the focus of the final scene” based on the words “sensation,” “youth,” “adolescence,” “last,” and “friendship.” However, the same word appeared in other similar topics in this analysis, which made it difficult to assess the topics subjectively. In this study, we used the term-score to address this [6]. The term-score is a dignity index that is attached to a word in TF-IDF that computes the feature quantity for each topic. For the TF value, the term-score is generated for a word in a topic, while the DF value is the document (topic) frequency. The dignity of a word that appears in two or more topics decreases using this index, which is expected to make it easier to determine the topic. The results obtained after analyzing the comic contents using the term-score are shown in Table 4. For example, it can be seen that the dignity of the words “作品 (work)” in topic 16 and “漫画 (comic)” (10 places below) in topic 21 decreased using the term-score. However, it is still difficult to estimate the contents of each topic by assessing their information characteristics. Using the term-score for some words (e.g., “雰囲気 (atmosphere)” and “言葉 (language)”) made it difficult to estimate the comic contents, but the dignity of the word assigned to each topic decreased for other words (e.g., “個人的 (individual)” and “私 (I)”) that were used often. Thus, it will be necessary to reconsider the type of information related to the comic contents in future research and also to treat certain information as stop-words.

5 Related work

In our study, our objective is to extract comic content information. Conversely, other researchers have tried to structure comic information. Morozumi et al. [7] collected comic information from Wikipedia and structured it based on the functional requirements for bibliographic records [8]. Various types of information related to a comic can be obtained if this structure is constructed successfully. For example, it is possible to access the bibliographic information from a magazine that carries an article, the author, etc., as well as access content information, such as a scene and a character. It is easy to obtain bibliographic information from Wikipedia. In contrast, comic content informa-

Table 2. The estimated result of an object comic (10 words generated higher probability)

topic	generated words
topic 0	絵 (graphic)，漫画 (comic)，人 (people)，確か (assurance)，私 (I)，為 (for) レベル (level)，一言 (a word)，涙 (tear)，手 (hand)
topic 1	最後 (last)，漫画 (comic)，好き (like)，作品 (work)，アニメ (anime)，話 (story) 世界観 (interpretation of the world)、展開 (expansion)，巻 (volume)，評価 (evaluation)
topic 2	他 (other)，マンガ (comic)，個人的 (individual)，意味 (mean)，部 (a part)，人物 (character) 話 (story)，オススメ (recommend)，素直 (obedient)，男 (man)
topic 3	方 (side)，先生 (teacher)，感じ (filling)，描写 (description)，題材 (subject)，以外 (except) 世界 (world)，何 (what)，迫力 (force)，移入 (bring in)
topic 5	漫画 (comic)，事 (thing)，点 (point)，頃 (about)，今 (now)，現実 (reality) 作品 (work)，バトル (battle)，好み (taste)，絵柄 (picture)

Table 3. The estimated result of “SLAMDUNK” (10 words generated higher probability)

topic	generated words
topic 6	名作 (masterpiece)，他 (other)，キャラクター (character)，全て (all)，存在 (existence)，言葉 (language) 中学 (junior high school)，小学生 (elementary school kid)，評価 (evaluation)，満点 (full points)
topic 15	キャラ (character)，時 (occasion)，最後 (last)，青春 (adolescence)，上 (upper)，秒 (second) 野球 (baseball)，ラスト (last)，感動 (sensation)，友情 (friendship)
topic 19	最高 (cap)，事 (thing)，涙 (tear)，最終 (final)，人物 (person)，力 (power)，無理 (impossible) 過去 (past)，理解 (understanding)，勝手 (liberty)
topic 20	漫画 (comic)，描写 (description)，人物 (person)，本 (book)，最近 (recently)，影響 (infection) 天才 (brilliant mind)，演出 (interpretation)，賛否 (yeas and nays)，初心者 (abecedarian)
topic 22	漫画 (comic)，ギャグ (gag)，レベル (level)，読者 (reader)，文句 (complain)，全国 (all parts of the country) アニメ (anime)，ボール (ball)，人間 (human)，能力 (ability)

Table 4. The estimated result using term-score (10 words generated higher probability)

topic	generated words
topic 12	ストーリー (story)，個人的 (individual)，一番 (first)，人生 (life)，内容 (content)，魅力 (charm) 主人公 (central character)，秀逸 (excellence)，最高 (marvelous)，作品 (work)
topic 16	巻 (volume)，絵 (graphic)，セリフ (a comic caption)，宇宙 (universe)，後半 (the latter half) 作品 (work)，雰囲気 (atmosphere)，テーマ (theme)，ネタ (news)，完全 (perfect)
topic 21	人物 (person)，キャラ (character)，展開 (expansion)，当時 (in those days)，視点 (a point of view) 結局 (after all)，感想 (one’s impression)，感動 (sensation)，能力 (ability)，冊 (volume)
topic 24	表現 (impression)，私 (I)，作者 (author)，子供 (child)，画力 (drawing ability)，言葉 (language) 主人公 (central character)，天才 (brilliant mind)，とら (Tora)，ジャンル (genre)
topic 26	感じ (filling)，人間 (human)，敵 (enemy)，自分 (self)，上 (upper)，愛 (love)，最近 (recently) 世界観 (interpretation of the world)，好き (like)，絶対 (absolute)

tion is not detailed, which makes information retrieval difficult. Therefore, information needs to be obtained from other sources, such as the approach we proposed in this paper.

6 Conclusion

In this paper, we proposed a search system that explores comic content information using comic content information obtained from review sentences on the Web The proposed system utilizes TF-IDF and LDA to obtain information. It uses TF-IDF to analyze explicit information in order to assess the information characteristics of each comic. Thus, it was possible to determine the comic contents from the information characteristics. In the analysis of implicit information using LDA, the feature quantities of the words were generated for each topic and the term-score computed. As a result, it was possible to confirm the characteristic words for each topic. However, it was difficult to determine the topic from these words.

In this paper, we considered extraction of comic content information. It is also necessary to distinguish the information required for a system from the extracted information. However, distinction of the information required for comic search cannot yet be performed. Therefore, we

plan to analyze the processes required for comic search and collect the information required for each process in future work.

Implicit information was analyzed using the LDA method. However, it was difficult to determine the topic contained in a review sentence using this method. This problem may be because of the fact that information similar to many review sentences was present. When a large quantity of similar information is available, it is difficult to determine the topic automatically considering delicate nuances. Therefore, in order to improve topic determination, the information needs to be distinguished by people. We plan to propose a new method for adding constraints on the topic after topic determination or determination using learning data such as ITM [9] and sLDA [10].

References

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

[2] Rachit Arora and Balaraman Ravindran. Latent dirichlet allocation based multi-document summarization. *In Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*, pages 91–97, 2008.

[3] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Interactive, topic-based visual text summarization and analysis. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 543–552, 2009.

[4] Kenya Sudo, Shogo Nagasawa, Kuniaki Kobayashi adn Tadahiro Taniguchi, and Toshiaki Takano. Encouraging user interaction of social network through tweet recommendation using community structure. *Conference on Technologies and Applications of Artificial Intelligence*, pages 300–305, 2013.

[5] Thomas. L. Griffiths and Mark. Steyvers. Finding scienific topics. *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235, 2004.

[6] David. M. Blei and John. D. Lafferty. *TOPIC MODELS*. Text Mining: Theory and Application Taylor and Francis, 2009.

[7] Morozumi Ayako, Nomura Satomi, Nagamori Mitsuharu, and Sugimoto Shigeo. Metadata framework for manga: A multi-paradigm metadata description framework for digital comics. *Proceedings of the 2009 International Conference on Dublin Core and Metadata Applications*, pages 61–70, 2009.

[8] International Federation of Library Associations and Institutions. Functional requirements for bibliographic records final report. http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf (last accessed October 2014).

[9] Yuening. Hu, Jordan. Boyd-Graber, and Brianna. Satinoff. Interactive topic modeling. *Proceedings of the 49th Annual Meeting of the Assocition for Computational Linguistics: Human Langurage Technologies*, volume 1, pages 248–257, 2011.

[10] David. M. Blei and Jon. D. McAuliffe. Supervised topic models. *In Advances in Neural Information Processing Systems*, volume 21, pages 121–128, 2007.