# Analyzing Textual Sources Attributes of Comics Based on Word Frequency and Meaning

Ryota Higuchi(✉), Ryosuke Yamanishi, and Mitsunori Matsushita

Kansai University, 2-1-1 Ryozenijicho, Takatsukishi, Osaka 569-1095, Japan
{k896930,ryama,m_mat}@kansai-u.ac.jp
https://mtstlab.org/

**Abstract.** The purpose of this research is to analyze the textual source attributes of explanations and reviews about comics. Comics are difficult to process in terms of the intended story because they are primarily composed of pictures and text. One of the processing methods is to analyze comics text on the Web, particularly the description of characters and reviews including the reader's impression about the comic. Sources of textual information, such as explanations or reviews, are selected according to the application of the study. However, differences among textual sources regarding comics are not taken into consideration in the analysis. This paper classifies words appearing frequently in the text semantically, with results showing that explanations include words that express the story, for example, the family structure, physical information, and sex of the characters for describing the characters. Conversely, the review frequently uses words that provide meta-information about comics, such as illustrations and style. The proposed method revealed that explanations of comics are more useful as textual sources for analyzing story information than reviews.

**Keywords:** Differences in Data Sources · Review Sentences of Comics · Explanation Texts of Characters · Characteristics of Comic Story

## 1 Introduction

### 1.1 Current Situation of Comics

The total number of comic books in circulation today is enormous, with more than 10,000 new comic books being published each year in Japan. A user who wants to find a new comic from a large number of comics that appeals to their interests retrieves it using web services such as e-comics and book sales websites. Typical comic retrieve methods use meta-information such as bibliographic information (e.g., title, author, journal) and genre (e.g., romantic comedy, action/adventure, human suspense) as queries. In MechaComic[1], one of

[1] https://sp.comics.mecha.cc, (confirmed September 2nd, 2022).


J.-J. Rousseau and B. Kapralos (Eds.): ICPR 2022 Workshops, LNCS 13644, pp. 107–118, 2023.
https://doi.org/10.1007/978-3-031-37742-6_8

the most popular e-comic stores in Japan, users search by tag information such as new arrivals and popular keywords ranking, genres (e.g., boys' manga, girls' manga), categories (e.g., fantasy, mystery, spin-offs). However, meta-information has no deep connection to the story. The amount of information in the meta-information is not sufficient for retrieving comics based on user preferences. As comics are cross-modal contents that contain both image information (e.g., characters and cartoons) and text information (e.g., dialogues and onomatopoeia), to recommend the comic that matches the user's interests, it is necessary to understand the story of the comic using its image and text information.

There are two approaches to understanding the story of comics. The first is a direct analysis of the images of the comic, where the story information is extracted from images of comic books by combining several procedures such as estimating comic panels [10], extracting characters [18], and constructing a dataset by adding metadata [8]. Some studies shows high extraction and estimation accuracy for individual elements; however, it is challenging to obtain story information automatically by combining these elemental technologies. For example, Manga109 [1,7] is an annotated dataset of comics that is open to the public. However, the only way to obtain story information about new comics is to use the aforementioned direct approach. The second approach is to obtain story information indirectly. This approach is realized by extracting content information from other resources and interpreting them. Studies analyzing reviews [15,17] and texts describing details of works or characters from websites [12] such as Wikipedia and Pixiv encyclopedia (pixiv 百科事典) are related to this approach. Other texts of comics on the Web exist in Q&A sentences about the comic [9], as well as in the outline of the comic in MechaComic. The advantage of the indirect approach is that considerable data exist on the We and is easy to collect.

### 1.2 Problem Statement

When studying content, there are multiple sources of information on the same topic. For example, in cooking informatics [4], recipes and reviews are textual sources about cooking content, while in the tourism field [2], texts about the same accommodations, such as reviews and explanations, are sources of textual information. However, few studies have considered how should choose datasets by quantitative analysis of textual source attributes.

Some types of information sources can be used to analyze comic content, such as the comics' review texts, explanations, and outline sentences. While these texts from different information sources represent the same content, the textual details vary from each source. For example, the review texts consist of descriptions intended to provide feedback and evaluation of the work, as well as the explanations and outline sentences consist of descriptions intended to provide an overview of the work. Reviews for available products provide content details to enable other people to evaluate the content before their purchases. At the same time, reviews for comics must exclude spoilers [6] to keep the entertainment values of the comics; too much detailed information about the story may be harmful to reviews of comics. Thus, it is reasonable to say that the reviews

**Table 1.** Explanation Sources and the number of data

| Source of explanation | the number of data |
|---|---|
| Wikipedia | 1,950 |
| Niconico Pedia (ニコニコ大百科) | 1,281 |
| pixiv encyclopedia (pixiv 百科事典) | 1,879 |
| Aniotawiki (アニヲタwiki) | 1,140 |
| Sum | 6,250 |

for comics would have less information of content itself than reviews for other products. We consider that this peculiarity is caused by the fact that comics is story-oriented content. Not only the reviews about comics, but also the explanations differ in some respects from those of general products. In the explanations of a general product, the features of the product are described in detail. Since the product is already complete, these features do not change. The explanations of comics, on the other hand, describes variable contents corresponding to the progress of the story such as the features of new characters, significant episodes and relationships between characters. As the above suggests, significant difference exist between reviews and explanations for comics rather than other products. It is, therefore, necessary to conduct a study using suitable information sources depending on the specific application that is the aim of each study. However, differences in textual sources in comics have received little attention. There are few quantitative criteria for selecting the appropriate source according to the purpose of use. To determine the criteria, this study explores what vocabulary is common (or different) between the two sources. This attempt provides two advantages. The first is that one can select appropriate information resources according to the different attributes. The second is that one has access to a large amount of data by appropriately combining different types of sources.

In this study, the attributes of each textual source about comics are analyzed by clustering frequently appearing words in the text semantically, using two analysis object sources from comics. One is explanations that convey the detail of comic content and the other is reviews that express readers' impression about the comic.

## 2 Analysis Method

This study analyzed attributes of different information resources from the same content. Frequently appearing words in the texts on the Web were clustered semantically using the following procedure to get the attributes:

1. Texts about comics gathered from selected websites are preprocessed.
2. Using word embedding, the vector of appearing words is acquired and a dictionary to use for classification is constructed.

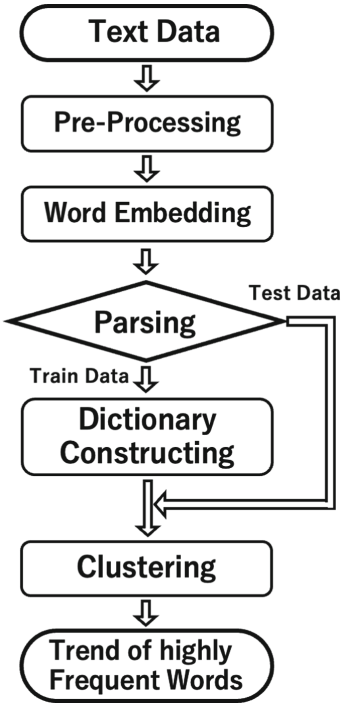


**Fig. 1.** Flowchart of Analysis Method

3. Words appearing frequently are clustered semantically with the dictionary obtained in Step 2.

Figure 1 shows a flowchart of the analysis method.

### 2.1 Dataset Construction

This study gathered explanations and reviews (6,250 each) from the Web as comic textual data. The differences in nature of description between the two types of sources are then analyzed.

First, character's explanations were collected from four types of websites, namely: Wikipedia[2], Niconico Pedia[3], pixiv encyclopedia[4], and Aniotawiki[5]. These websites consist of information written by several people, rather than a single author and the sentences on these websites includes a lot of story information about comics. Table 1 shows the number of sentences collected from each

[2] https://ja.wikipedia.org/wiki/ (confirmed September 2nd, 2022).
[3] https://dic.nicovideo.jp (confirmed September 2nd, 2022).
[4] https://dic.pixiv.net (confirmed September 2nd, 2022).
[5] https://w.atwiki.jp/aniwotawiki/ (confirmed September 2nd, 2022).

website. The collected explanations are about 2,067 comic characters. Most e-book and manga stores show the outline of the story for each work of comics, and it can be used as another textual resources of the story. The outline expresses the story briefly, using sentences, which is better than other sources in that they summarize all the symbolic information of the comic work but is inferior in that it is difficult to collect and may contain poor expressions. To express the comic story objectively, this study focused on explanations rather than synopsis as a more appropriate means of expression. The character explanation describes the character's actions and feelings, focusing on episodes in which the character stood out. Then, many sentences are used to explain the story in detail. This study also focused on character explanation as an appropriate means for acquiring comic story information.

Second, reviews were collected from the manga category of Sakuhin Database[6]. Among them, the top 200 most recently viewed works were included. The purpose of this site is different from shopping sites such as Amazon.com because, whilst the purpose of shopping sites is to make the customer purchase comics, the purpose of review sites like Sakuhin Database is to evaluate and collect information. Reviews from shopping sites [5] can also be used to obtain story information about comics; however, there is the risk of mixing information that has almost no connection with the story with relevant information. Therefore, this study gathered text information only from review sites. To avoid bias in the number of reviews for each work, the maximum number of reviews per work was defined as 45.

### 2.2 Word Segmentation and Data Cleaning

This study created a dictionary to semantically classify words in the texts. The explanations and reviews had a large variety of information such as contents of work and readers' feedback. It was thus necessary to extract this information comprehensively to research the attribute of each information source. This study focused on nouns, which can express a broad range of meanings, such as persons, things, and places. The characters' explanation contains words that describe the character's feature and content such as "Energy (元気)" and "Victory (勝利)." Reviews have words used to describe impression such "royal road (王道)" and "sentiment (感動)." Nouns in the texts were analyzed to study the attribute of each information source. Proper nouns, such as the name and technique of the appearing character, provide information that cannot be used to understand the story of comics. Therefore, these words were eliminated from the analysis texts in advance. This study used MeCab (Version0.996) as the morphological analyzer and Neologd (Version0.0.7) as the Japanese dictionary. There are also considerable proper nouns dealing with comic content. Collected texts include some words that do not describe comic content directly such as "that" and "when" in Japanese and this study defined stopwords to eliminate these words as noise. The stopword list was constructed with Slothlib [11] that holds 310 common Japanese

[6] https://sakuhindb.com/ (confirmed September 2nd, 2022).

**Table 2.** Example of characteristic words clustered together in the same class

| Class A | Class B | Class C |
|---|---|---|
| hard battle (激戦) | black (黒) | idol (アイドル) |
| comrades in arms (戦友) | white (白) | shortcoming (コンプレックス) |
| first game (初戦) | brown (褐色) | class (クラス) |
| hard fight (苦戦) | complexion (顔色) | gym (ジム) |
| mind-game (頭脳戦) | youth (青春) | position (ポジション) |
| strategy (作戦) | clear (明白) | earrings (ピアス) |
| warring States (戦国) | white coat (白衣) | badminton (バドミントン) |
| battle (対戦) | love (色恋) | jungle (ジャングル) |
| war (戦乱) | red blood cell (赤血球) | diving (ダイビング) |
| war situation (戦況) | dark (暗黒) | apple (リンゴ) |

words. A Japanese single word (hiragana, katakana), numbers, and symbols were also added to the list because there meanings are difficult to determine. The text data were then cleaned using the list constructed.

Low frequency words may become noise when performing semantic classification of word (described later in Sect. 2.3) used in each source. Therefore, this study removed low frequency words that appeared less than 10 times, a threshold limit that was determined empirically. The aforementioned process was then applied to both sources, explanations and reviews. The total number of unique words was 7,136 and 3,092 for explanations and reviews, respectively. Both the textual sources, each of which had 6,250 data points, were split in a ratio of 8:2 for the next classification. The number of training data and test data points was 10,000 and 2,500, respectively.

### 2.3 Classification of Frequently Appearing Words

This study classified the preprocessed words semantically and used Japanese Wikipedia entity vector (i.e., Wiki vector), which was proposed by Suzuki [14]. The Wiki vector was constructed using the full text of Japanese Wikipedia as

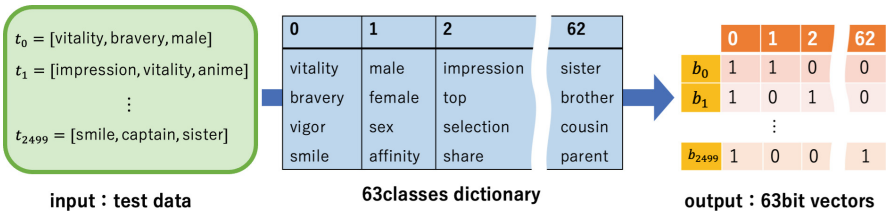


**Fig. 2.** Classification of Highly Frequent Words by Using the Class Dictionary

training data. The features of this vector are robust to common words and vocabulary ambiguity. The number of dimensions of the vector was set to 200, with a window size of five words. The prepared dataset was then vectorized with Wiki vector.

This study classified the word vector obtained from the aforementioned process using k-means clustering and conducted an exploratory search of the ideal number of classes from 2–100 using the elbow method. The elbow method is one of the most popular methods to determine the ideal number of classes. The results showed that 63 classes were appropriate. There were examples of characteristic word class in some of these classes (shown in Table 2). Some words that mean "battle (戦)" and "color (色)" were classified into class A and class B in the table, whilst some words in Class C were not similar in meaning or shared the common feature of being foreign words. The average number of words and the standard deviation per class was 118.8 points and 107.1 points respectively, while the maximum and minimum word counts per class was 495 and 8 points, respectively. This number shows that words belonging to each class was uneven. The resulting class sets of word was a class dictionary to use for content analysis of comics. Figure 2 shows the process of classifying frequent words using the class dictionary. The attribute of some textual sources was obtained by calculating the appearance frequency of a word in the 2,500 test data points and construing a word's meaning in the class dictionary. For example, there is "apple (リンゴ)" in class C in Table 2. When "apple" is in one test data point, it indicates that this data contains an element of class C. Then, this study can put "1" in class C in place of this data. The output data was a binary string that contains a 63 bit vectors for each of the 2,500 test data points.

## 3 Results and Discussion

There are two steps to consider the result. First, this study investigated the class that corresponded to the highly frequent words in each information source. Second, the appearance ratio of each class was calculated to compare some differences of attribute between the two types of sources. The relative difference for each class was defined as an absolute value of the difference ratio in each source.

**Table 3.** Relative difference between frequent words in explanation and compared to those in review

| Class words | Relative difference |
|---|---|
| body(身), body length(身長), familiar(身近), rank(身分) | 74.2 |
| condition(条件), no condition(無条件), belief(信条), vote(票) | 69.0 |
| parent(親), brother(兄), sister(姉), cousin(従兄弟) | 63.7 |

**Table 4.** Relative difference between frequent words in reviews and compared to those in explanation

| Class words | Relative difference |
|---|---|
| comic(漫画), movie(映画), illustration(イラスト), style(画風) | 35.5 |
| work(作品), cartoonist(作家), drawing(作画), masterpiece(傑作) | 19.8 |
| season(節), volume(巻), generation(世代), sequel(続編) | 3.1 |

### 3.1 Attribute of Explanation

There are many words that describe a character's feature and the content of works in the explanation source. Table 3 shows some classes of explanation that have a higher appearance ratio than that in review. The class of max relative difference (74%) includes many words that used "身", such as "body length (身長)" and "familiar (身近)." For example, the description in Wikipedia for "Tanjiro Kamado," who is a character from "Demon slayer (鬼滅の刃)" includes the information that "His body length is 165cm." In addition, "familiar (身近)" is often used to explain an episode that happened around the character. This word contained in the explanation source is another reason why this class showed a high percentage.

The class of the second largest relative difference includes words such as "condition (条件)" and "belief (信条)." This class has a relatively high ratio in explanations (69%) despite containing the lowest number of words (eight words) of all classes. An example of explanations containing words in this class is from the pixib encyclopedia for "Ken Kaneki" who is a character from "Tokyo Ghoul (東京喰種)." The sentence is as follows: "He proposed conditions wherein he

would sacrifice himself in exchange for the freedom of his friends." Data containing this class of words tended to describe information about the story. Sometimes, the review text did not refer to the story directly to avoid spoilers, whilst words that appeared in the part describing the episode tended to appear more often in explanation sources. "Condition (条件)" also appeared frequently when describing a character's abilities. There is a description that "Raining makes the trigger conditions of his ability more severe." in the Aniotawiki describing "Roy Mustang" who is a character from "Fullmetal Alchemist (鋼の錬金術師)."

The class that corresponded to the third largest number of test data contained some words referring to family structure (e.g., parent, brother, grandchild). These words were often used to explain the relation between characters [13]. The text in the pixib encyclopedia for "Sabo" who is a character from "One Piece (ワンピース)" explains "He is Luffy's other brother. He spent his childhood with Ace and Luffy."

### 3.2 Attribute of Review

Review sources contained sets of words that represent meta-information about comic works such as illustration, style, and cartoonist. Table 4 shows several classes of reviews with a higher appearance ratio than that in explanation.

Review sources contained more interpretive information about how the author of the review text felt after reading the work than information relating to the story. For example, "This cartoonist style will have a great influence on future generations." The result suggest that review source could be used for research on genre analysis [3] and topic classification [16]. Genre and topic information express the category of the work, which is metadata and not data on content. Interestingly, unlike description sources, review sources do not have classes with significantly relative differences. The largest relative difference was only 35.5%. This could be because the review data included very short sentences such as "It was interesting, and I want to read it again."

Review sources have no class with a large percentage of difference relative to explanation sources. Even classes with the largest differences have values of less than approximately 40%. The results of the analysis differed significantly between the two types of information sources. The sum of the applicable classes for each source was calculated to investigate the causes. A total of 35,694 points for explanation and 20,429 points for review were calculated, indicating that explanation data corresponds to approximately 1.75 times more classes than review data. This may be owing to the fact that the adjectives were eliminated from the subject. For example, there are some descriptions using adjectives such as "the thrilling and exciting story is hot." and "the expression was scary." In this research, only common nouns were considered in the analysis. Information such as impressions of the work using adjectives was eliminated. Future work will focus on the analysis of different parts of speech.

**Table 5.** Class words for which the most data corresponded

| |
|---|
| hairstyle(ヘアスタイル), check(チェック), plastic model(プラモ), tire(タイヤ), character(キャラ), hip(ヒップ), waist(ウエスト), animation(アニメ), diet(ダイエット), part - time job(バイト), cake(ケーキ), television(テレビ), motorbike(バイク), up(アップ), mascot(マスコット), rehabilitation(リハビリ), piano(ピアノ), pianist(ピアニスト), type(タイプ), captain(キャプテン), stupid(バカ), Popular(モテモテ), basket(バスケット) |

### 3.3 Overall Discussion and Future Work

The class with the most data in all classes included a significant amount of foreign words as showed in Table 5. Approximately 73% of all data belonged to this class. This result would due to representations peculiar to the Japanese language. The Japanese language is composed of three types of characters: hiragana, katakana, and kanji. Words in this class share the superficiality of being written in katakana. Kanji are ideographic characters imported from China in the earliest times, and hiragana are phonetic characters derived from them. For this reason, things that have existed in Japan since ancient times are often expressed using a combination of kanji and hiragana. In contrast, katakana is often used to express concepts newly introduced overseas, such as "animation," or words derived from foreign languages, such as "check." There are two other classes that included a large amount of foreign words; however, the words in these classes were not semantically clustered. From this result, it is not clear whether combining sources with similar vocabulary is possible. These classifications can be improved by using a different Japanese dictionary in the future.

Although this research embedded words in the text using Wiki vector, Hotto link Inc.[7] has built a Japanese corpus formed of SNS data such as blogs and Twitter, Japanese Wikipedia, and automatically collected web pages. Character's explanation and reviews of works include new and unknown words such as net slang and broken expressions. This corpus contains many new and unknown words and in the future, it is expected that this corpus will be applied to interpret the meaning of katakana words, such as those listed in Table 5.

## 4 Conclusion

This paper targeted explanation and review texts among textual information sources dealing with comic contents and analyzed sources attributes based on word frequencies. Results showed that explanation sources frequently contained words that described characters and content of works, suggesting that they are

[7] http://www.hottolink.co.jp/english/ (confirmed September 2nd, 2022).

a suitable source for analyzing the comic story. Conversely, review sources frequently contained words that provided meta-information about the works such as illustrations, styles, and cartoonists.

For future research, it is necessary to consider the part of speech to be focused on as the analysis target. In future research, it will be necessary to consider different parts of speech as the analysis target.

**Acknowledgement.** This work is supported by JSPS KAKENHI Grant Number #22K12338.

## References

1. Aizawa, K., et al.: Building a manga dataset "Manga109" with annotations for multimedia applications. IEEE Multimedia **27**(2), 8–18 (2020)
2. Coenders, G., Ferrer-Rosell, B.: Compositional data analysis in tourism: review and future directions. Tour. Anal. **25**(1), 153–168 (2020)
3. Daiku, Y., Iwata, M., Augereau, O., Kise, K.: Comics story representation system based on genre. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 257–262 (2018)
4. Kikuchi, Y., Kumano, M., Kimura, M.: Analyzing dynamical activities of co-occurrence patterns for cooking ingredients. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 17–24. IEEE (2017)
5. Liu, H., Wan, X.: Neural review summarization leveraging user and product information. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2389–2392. Association for Computing Machinery (2019)
6. Maki, Y., Shiratori, Y., Sato, K., Nakamura, S.: A method to construct coimc spoiler dataset and the analysis of comic spoilers. IEICE Technical report (2020)
7. Matsui, Y., et al.: Sketch-based manga retrieval using Manga109 dataset. Multimedia Tools Appl. **76**(20), 21811–21838 (2017)
8. Mihara, T., Hagiwara, A., Nagamori, M., Sugimoto, S.: A manga creator support tool based on a manga production process model-improving productivity by metadata. In: iConference 2014 Proceedings. iSchools (2014)
9. Moriyama, Y., Park, B., Iwaoki, S., Matsushita, M.: Designing a question-answering system for comic contents. In: Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding (2016)
10. Nguyen Nhu, V., Rigaud, C., Burie, J.C.: What do we expect from comic panel extraction? In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 1, pp. 44–49 (2019)
11. Ohshima, H., Nakamura, S., Tanaka, K.: SlothLib: a programming library for research on web search. Database Soc. Jpn. (DBSJ Lett.) **6**(1), 113–116 (2007)
12. Park, B., Ibayashi, K., Matsushita, M.: Classifying personalities of comic characters based on egograms. In: International Symposium on Affective Science and Engineering, ISASE2018, pp. 1–6. Japan Society of Kansei Engineering (2018)
13. Ruch, W., Gander, F., Wagner, L., Giuliani, F.: The structure of character: on the relationships between character strengths and virtues. J. Posit. Psychol. **16**(1), 116–128 (2021)

14. Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., Inui, K.: Fine-grained named entity classification with Wikipedia article vectors. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 483–486. IEEE (2016)
15. Ueno, A., Kamoda, Y., Takubo, T.: A spoiler detection method for Japanese-written reviews of stories. Int. J. Innov. Comput. Inf. Control **15**(1), 189–198 (2019)
16. Xu, A., Qi, T., Dong, X.: Analysis of the Douban online review of the MCU: based on LDA topic model. J. Phys. Conf. Ser. **1437**(1), 012102 (2020)
17. Yamashita, R., Okamoto, K., Matsushita, M.: Exploratory search system based on comic content information using a hierarchical topic classification. In: Proceedings of the Asian Conference on Information Systems, pp. 310–317 (2016)
18. Yanagisawa, H., Kyogoku, K., Ravi, J., Watanabe, H.: Automatic classification of manga characters using density-based clustering. In: 2020 International Workshop on Advanced Imaging Technology (IWAIT), vol. 11515, p. 115150F. International Society for Optics and Photonics (2020)