輔仁管理評論

中華民國 112 年 9 月，第三十卷第三期，1-20

Text Mining Techniques for Similarity of High Imbalanced Length Texts - A Case of The Adapted Anime and Their Original Light Novel

YI-NING TU, PENG-HSUAN LEE [*]

(Received Nov. 08 2022; Accepted Nov. 29 2022)

ABSTRACT

The similarity calculation will highly impact the text mining model training results especially the well-known cosine similarity. There is a serious problem especially the length of the compared texts are highly imbalanced. In this kind of situation, the bag of words will be quite large and not easily have the same terms in the word vector matrix. Even if there are hits the terms also will be diluted by the large number of different texts between the two texts. The study proposed an algorithm and tried to solve the problems and give a case study between adapted anime and their original light novel. The new proposed similarity will replace the traditional cosine similarity to handle the case study. This study uses 32 online original light novels that are published after 2000 and 32 corresponding online adapted anime episodes' summaries to find the key terms and calculate the similarity between. The result shows different genres of anime have different relationships of similarity and popularity. Besides, the proposed work also provides some strategies based on the analytic results.

Keywords: Text mining, Similarity of high imbalanced length texts, Adapted anime, Light novels

[*] Yi-Ning Tu, Associate Professor, Fu Jen Catholic University, Statistics and Information Science, E-mail: eniddu@gmail.com, Corresponding Author. Peng-Hsuan Lee, Waseda university, Fundamental Science and Engineering (Japan).

I. INTRODUCTION

Since many anime are adapted from their original light novel, the quality of the adaptation often has a direct effect on audiences' rating for the adapted anime. The differences between the anime and the light novels may vary according to the additional scenes or deleted details in the anime. Based on the past related studies, different forms of art or mediums may convey a new message compared to their original novels.

Comparing and analyzing the theme and scenes between Fei-Yu Bi's novels and the adapted movies of his, Wu(2018) has demonstrated literature and movies express different interpretations on the urban-rural relation, women status, and power through their ways of present and point of view(Wu, 2018). In Liao's (2017) intertextuality study of Liao Hui-Ying's novels and film adaptations, to find the similarities and differences between the novels and the adapted movies, he discussed and analyzed the characterization, added and deleted textual plot, and themes between two forms of art. Based on the result, it has appeared that adapted movies would ignore some important details and revise the character traits given in the original novels since they have to take into consideration the market (Liao, 2017). While movies have to consider the market, novels are the places for Liao Hui-Ying to empower women by creating awareness without limitations.

According to the previous research, most only discussed the difference and similarities between the original novels and the adapted films, which caused them to convey different messages to the audiences. However, they rarely connected the difference and similarities of the text structures to the popularity of the adapted film, especially in light novels and anime. If there is a relationship between the faithfulness of the anime adaptation and the series popularity, the result can help anime production companies with the consideration of the market.

Previous research focuses on the different ways of interpretation and the different conveyed messages from the novel and its adapted films. To our best knowledge, none have discussed the effect of similarities of textual structure between light novels and anime episode summaries on the popularity of the anime.

However, the similarities between the original light novels and the adapted anime might affect the ratings and reviews of the audiences. Audiences may have seen the light novel first, which is expecting the content of the adapted anime to be the same. Therefore, the more similar the light novel and the anime are, the more popular the anime is going to be. Yet, some viewers may watch the anime without reading the light novel first and give a bad rating for the anime. Since there are many scenarios for this situation, this research devotes itself to examine the relationship between the similarities of the light novels with their adapted anime and the popularity of the anime.

This research gathers 32 anime adapted from its original light novels. Each genre, which are action, romance, mystery, and comedy, contained 8 anime, while each adapted anime's release year, production company, and popularity (rating on Bahamut[1]) and its original light novel's release year, popularity (sales), and year of the end of calculating the sales are collected and recorded. To find the relationship between the faithfulness of the anime adaptation and the series' popularity, the study will investigate the repeated words between light novel chapters and the summary of corresponding anime episodes by utilizing CKIP Chinese word segmentation system[2] and Chinese stop words[3] . Then, compare the results with the popularity of the anime based on their rating on Bahamut.

The following are the purposes of the research:

(1) . Develop a new method to calculate the similarity between original work and adapted work.

(2) . Demonstrate that different genres (action, romance, mystery, and comedy) have different relationships between the faithfulness of the anime adaptation and the series' popularity.

(3) . Provide a reference for anime production companies to determine the degree of changes that should be made for the adapted anime from its original light novel considering the market.

[1] Bahamut https://ani.gamer.com.tw/

[2] CKIP Chinese word segmentation system http://ckipsvr.iis.sinica.edu.tw/

[3] Chinese stop words https://www.itread01.com/content/1494661154.html

II. Literature Review

1. Problem History

In the modern movie, anime, and film industry, the works have referenced the plots from novels to create adapted films or anime. When rewriting a story, the production companies consider the current market, trends, budgets, and audiences' tastes. These factors often affect the degree of changes between the original novels and the adapted films. Liao(2017), as previously mentioned, has compared Liao Hui-Ying's three novels with their adapted movies. When comparing, he focused on the three aspects: characters' characterizations, text plots, and main themes. The result showed the most significant difference between the original novels and the adapted movies is that movie production companies have to consider the market while the author of the novels doesn't. Therefore, some important details in the novel did not appear in the movie due to production budget consideration (Liao, 2017). Wu(2018) has discussed how literature and movies present the messages to the audience differently by comparing Fei-Yu Bi's three novels with their adapted movies. From a point of view, literature and movies are related and similar. They both convey their messages, artistic choices, and thoughts through storytelling. However, literature and movies are different. While literature utilizes words to leave spaces for imagination, movies provide concrete scenes through set-up images and videos, which leave limited spaces for imagination (Wu, 2018). Furthermore, with the change in the ways of presenting the story, the audience either feels more understandable because of the clear images the movie provides or dislikes since the movie isn't what they were expecting to be like the novel.

The relationship between novels and movies is similar to light novels and anime. Similarly, anime has a big market in modern society. Huang (2010) conducted lifestyle research on different types of animation consumer groups. The study has discovered anime fans consumed anime-related daily necessities, anime 3D models, cards and stickers, comics, and light novels(Huang, 2010). With a big amount of anime merchandise to sell, anime production companies can make a huge profit when the anime based on an original light novel has adapted well. Therefore, anime production companies, similar to the movie industry, often have to consider the market.

2. Similarity between Original Work v.s. Adapted Work

Tu and Wang(2020) has utilized text mining techniques on movie reviews, movie plots, and the introduction of movies in order to find the similarity between them and predict whether the movie review has spoiled any important plot from its content. Therefore, the author used the formula of cosine similarity((A·B)/(‖ A ‖ ‖ B ‖)) to calculate the similarity. The result revealed the similarity between movie reviews and movie plot does help determine whether the review spoils. However, cosine similarity usually compares two similar length works whereas the research will pick the key terms from an anime episode's summary, which is a short paragraph, as a benchmark to compare with the corresponding chapter of the light novel, which includes a few pages with a couple of long paragraphs. Therefore, if the research applies cosine similarity, the boolean values(1 or 0) that represent the relationship will show that one of the matrix values will always be 1, which is likely to be the anime's episode summary. Moreover, the number of terms in the anime episode summary is extremely few compared to the number of terms in a chapter of the light novel. As a result, the study can not use cosine similarity to calculate the similarity between the adapted anime and its original light novel.

Similarly, Tu and Lin(2020) has used novel text dialogues to identify the character relationship in a chinese romance novel-Ruyi's Royal Love in the Palace. Their research separates the text based on terms through the use of n-gram and Jieba, and calculates the word frequency that appears in the novel text. The result showed the method can indeed identify 25 relationships between the 18 main characters in the novel.

Overall, the past research has not discussed whether the similarity between adapted anime and original light novels affects the popularity of the adapted anime. They are mostly investigating movies with the original novels but rarely talk about anime and light novels.

III. Similarity Algorithm of Imbalanced Length Texts

To turn the similarity between the original light novel and adapted anime into measurable indices, the research calculates the number of words repeated in the light novel and the summary of the corresponding anime episode.

Set original light novel as *O* and adapted anime as *A*. There are O_i and A_i, which *i* is the ith of light novels and anime. *j* represents the j^{th} episode of the anime, while *k* represents the k^{th} sentence from the episode summary of the anime and *z* is the z^{th} words of the sentence.

$$\text{If } A_{ijkz} \in O_{ij}\,,\ A_{ijkz} = 1\text{, otherwise, 0}$$

$$Similarity(Oi, Aijk) = \frac{\sum_1^n A_{ijkz}}{n} \qquad (1)$$

Where,

- A_{ijkz} indicates the z^{th} word of the k^{th} sentence of the j^{th} episode of the i^{th} anime.
- O_{ij} represents the j^{th} chapter of the i^{th} original light novel.
- *n* denotes the total number of terms(words) that appears in the k^{th} sentence of the j^{th} episode of the i^{th} anime.
- Similarity(O_i, A_{ijk}) shows the similarity between the k^{th} sentence of the j^{th} episode of the i^{th} anime and the original i^{th} light novel.

Every line ending with a “.”, “?”, “!”, “......” will be considered as a sentence. Words that are between the brackets like “ “, [], and 「」 are considered as one term. The study utilizes CKIP Chinese Word Segmentation[4] to separate words and Chinese stop words[5] to parse unnecessary words(terms). After that, this work calculated the total initial number of the words, *n*. If A_{ijkz} appears in both the episode summary and the corresponding chapter of the light novel excluding the translation problem, then it is equal to 1. However, sometimes anime would have a similar object or character but a different name from the light novel. Therefore, if A_{ijkz} is a noun according to the CKIP and there’s a similar word with a similar meaning, the word counts as a repeated word. In the end, divide the sum A_{ijkz} by *n* to calculate the

[4] http://ckipsvr.iis.sinica.edu.tw/

[5] https://www.itread01.com/content/1494661154.html

similarity of the sentence and record. Formula (1) shows the calculation of the similarity of the sentence.

For instance, the terms in the sentence -“我終於知道貝爾的思想了！”- is being separated by CKIP and become

我(Nh) 終於(D) 知道(VK) 貝爾(Nb) 的(DE) 思想(Na) 了(T) ！

After parsing and deleting the unnecessary based on Chinese Stop Words, there are only three terms - 終於(D) 貝爾(Nb) 思想(Na) - left.

Then, check whether the terms appear in the corresponding chapter of the light novel. If “終於” is not in the light novel, but there’s “終于”, which is the exact same meaning but different form of word, “終於” is considered as a repeated word and is equal to 1. If “思想” also doesn’t show up in the novel and there’s a similar word like “想法”, “思想” will be a repeated word too since it’s a noun(Na). In the end, since three terms all appear in the light novel, the similarity of this sentence,(Oi, Aijk), will be 3÷3, which equals 1.

When all the sentences have their results, calculate the episode similarity by measuring the average of k^{th} sentences’ similarity as formula (2).

$$Similarity(Oi, Aij) = \frac{\sum_1^h similarity(Oi, A_{ijk})}{h} \qquad (2)$$

Where,

- Similarity(O_i, A_{ijk}) shows the similarity between the k^{th} sentence of the j^{th} episode of the i^{th} anime and the original i^{th} light novel.
- h denotes the total number of sentences that appears in the j^{th} episode of the i^{th} anime.
- Similarity(O_i, A_{ij}) indicates the similarity between the j^{th} episode of the i^{th} anime and the original i^{th} light novel.

For instance, in 加速世界(Accel World)’s episode 23, there are four sentences in the episode summary. After calculating the each sentence’s similarity, Similarity(O_i, A_{ijk}), the sentence similarity are recorded, as shown in Table 1.

Table 1. Accel World episode 23’s similarity

anime+episode	sentence 1	sentence 2	sentence 3	sentence 4
Accel World ep.23	春雪和小拓帶著小千來到加速世界與能美決鬥，能美竟然讓連接者 Black Vise 在戰場設了埋伏，這使春雪落入了陷阱。	在戰鬥力只有小拓的情況下，能美不但對他進行強力的攻擊，還用小千作為人質威脅小拓。	為了小千，小拓只能放棄了抵抗。	就在這危急關頭黑雪姬出現了，解救了小拓，此時春雪早已從陷阱逃脫，而他與黑雪姬能戰勝強大的敵人嗎？
(similarity)	0.88	0.90	1.00	0.69

In order to find the similarity(Oi, Aij),between the episode and the light novel, calculate the average of the sentence similarities.

$$\text{similarity}(O_i, A_{ij}) = (0.88+0.90+1.00+0.69)/4 = 0.87$$

Eventually, find the average of j^h episodes’ similarity for the final anime similarity when all the episodes have their episodes similarity. Formula (3) demonstrates the calculation.

$$Similarity(Oi, Ai) = \frac{\sum_1^g similarity(Oi, A_{ij})}{g} \tag{3}$$

Where,

- Similarity(O_i, A_{ij}) indicates the similarity between the j^{th} episode of the i^{th} anime and the original i^{th} light novel.
- g denotes the total number of episodes that appear in the i^{th} anime.
- Similarity(O_i, A_i) indicates the similarity between the i^{th} anime and the i^{th} light novel.

For instance, in anime 我的朋友很少(Haganai), in order to calculate the similarity((O_i, A_i), the episodes similarities,which are the orange number shown in Table 2., are added together and divided by the total number of the anime episode, which is 12.

Table 2. Haganai anime similarity

Anime + episode	sentence 1	sentence 2	sentence 3	episode similarity
Haganai ep.1	因為面相兇惡而交不到朋友的羽瀨川小鷹，在某天放學之後，目擊一名少女獨自開心與空氣朋友聊天。	少女成立了交友社團，並且把小鷹列為社員，然而校內的超有名人物來訪了。		
Adapted content	1.00	0.75		0.88
Haganai ep.2	交朋友要靠遊戲！	只要預先練習掌上型遊樂器的連線組隊模式，在交到朋友的時候就無須慌張。	由於有可能自己開局邀請朋友連線，所以記得要詳閱說明書。	
Adapted content	1.00	0.70	0.71	0.80
Haganai ep.3	開聲音玩遊戲的時候，要注意周圍的狀況。	盡可能在上鎖的個人房使用耳機遊玩為佳。	要是遊戲的聲音被聽到，或是有人忽然站在身後，即使對方是朋友也會招致反感。	
Adapted content	1.00	0.40	0.75	0.72
Haganai ep.4	不可以用外表判斷一個人。	即使看起來像是孩子，看起來像是女生，甚至是身穿白袍，內在或許也與想像中不同。	排除先入為主的觀念與他人來往，是結交朋友的重要步驟。	
Adapted content	1.00	1.00	0.57	0.86
Haganai ep.5	享受虛擬世界體驗的時候，必須注意自己處於毫無防備的狀態，頭戴式顯示器正是解決這個問題的優秀裝置，不過終究無法避免引來周圍的奇異目光。			
Adapted content	0.70			0.70

Anime + episode	sentence 1	sentence 2	sentence 3	episode similarity
Haganai ep.6	KTV 最近開始提供各式各樣的服務，成為不只是用來歡唱的休閒場所，不過上次看的電視節目有提到，能夠便宜歡唱的 KTV 還是比較受歡迎。			
Adapted content	0.80			0.80
Haganai ep.7	現代的手機分成智慧型手機與行動上網型手機，搞不懂兩者有什麼差異。	總之大部分的使用者，應該都是只要能打電話、傳郵件與上網就好。		
Adapted content	0.71	0.67		0.69
Haganai ep.8	這集難得是泳裝篇，真希望學校泳裝能夠登場。			
Adapted content	1.00			1.00
Haganai ep.9	因為自己名字太特別而感到芥蒂的人請注意，特別的名字容易令人留下印象，並且方便記憶，等到長大成人就會感受到這樣的恩惠。			
Adapted content	0.87			0.87
Haganai ep.10	經常有人說“回家之前都算旅行”，不過老實說，旅行的回程最累人。	所以就把回程當成比較長的通勤時間吧，平常總是想早點回家的人，只要這樣想就會輕鬆得多。	對不起，我說謊了。	
Adapted content	0.40	0.92	1.00	0.77
Haganai ep.11	你穿浴衣的時候，底下會不會穿內衣？	啊，"在你面前絕對不會穿浴衣"嗎，我想也是。		
Adapted content	0.60	0.50		0.55

Anime + episode	sentence 1	sentence 2	sentence 3	episode similarity
Haganai ep.12	人們會先以粗略的要素辨識他人，比方說，更換髮型會影響到整個人的輪廓，所以別人很容易察覺到不同之。			
Adapted content	0.40			0.40

Similarity(O_i, A_i)= (0.88 + 0.80 + 0.72 + 0.86 + 0.70 + 0.80 + 0.69 + 1.00 + 0.87 + 0.77 + 0.55 + 0.40) / 12 = 0.75

As a result, 0.75 is the similarity between the adapted anime - Haganai - and its original light novel.

IV. EXPERIMENTAL RESULT

1. Data Collection

The study collects 32 anime adapted from light novels, 8 for action, romance, comedy, and mystery each. Since there are 32 anime and light novels being collected, *i* ranges from 1 to 32. The criteria of experimental dataset illustrated as follows: original light novel has to be published after 2000 with its total sale recorded online, while the adapted anime has to have its release year, anime production company, rating from Bahamu, information of corresponding anime episodes with light novel chapters, and Baidu[6]'s anime episode summaries found online. All information will be collected and recorded from February, 8th 2021 to August, 6th 2021.

As the 32 anime's similarities are done calculating, they will be plotted on a graph. The x-axis is the similarity of the anime and the light novel, in which the endpoints are the maximum and minimum values of the 32 anime similarities. On the other hand, the y-axis is the popularity of the anime from Bahamu, in which the endpoints are the maximum and minimum values of the 32 anime's popularity. The coordinate of the origin is the average of the similarity(x) and popularity(y). The coordinate of each anime is shown in Table 3.

[6] https://ani.gamer.com.tw/

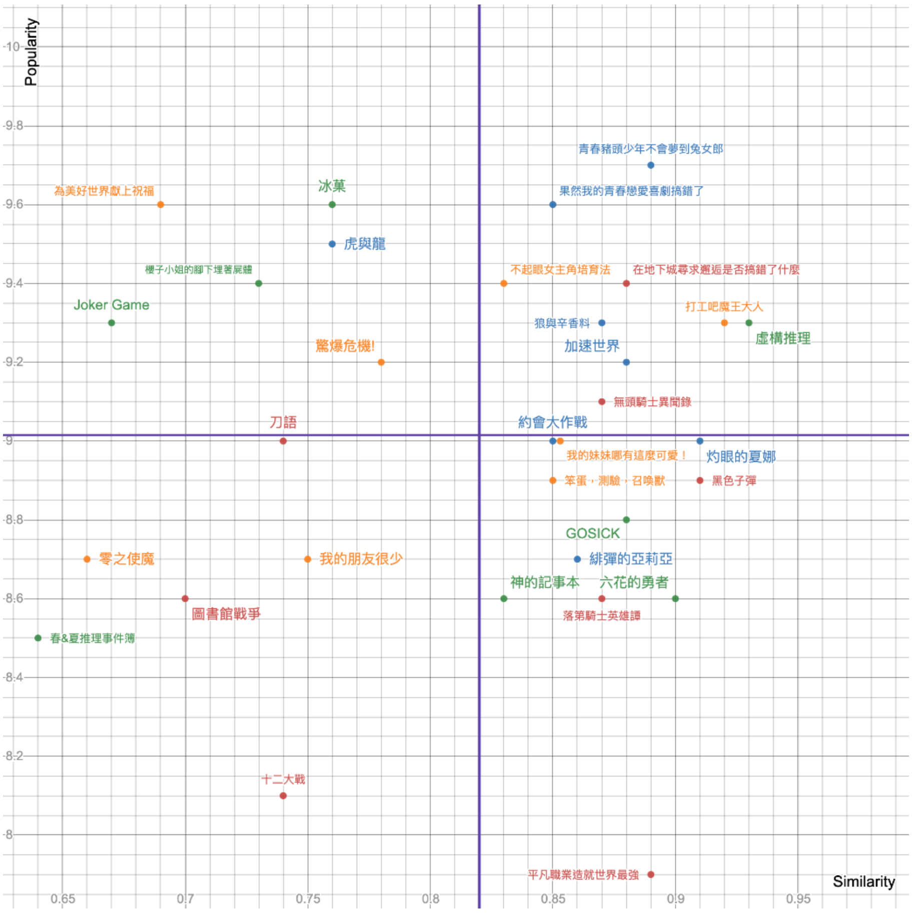
Table 3. The 32 Anime with Their Similarity and Popularity

	Genre	Anime(Chinese)	Anime(English)	x coordinate (similarity)	y coordinate (popularity)
1	action	在地下城尋求邂逅是否搞錯了什麼	Is It Wrong to Try to Pick Up Girls in a Dungeon?	0.88	9.40
2	action	平凡職業造就世界最強	Arifureta: From Commonplace to World's Strongest	0.89	7.90
3	action	黑色子彈	Black Bullet	0.91	8.90
4	action	十二大戰	12 Wars	0.74	8.10
5	action	刀語	Sword Story	0.74	9.00
6	action	落第騎士英雄譚	Chivalry of a Failed Knight	0.87	8.60
7	action	無頭騎士異聞錄 DuRaRaRa!!	Durarara!!	0.87	9.10
8	action	圖書館戰爭	Library Wars	0.70	8.60
9	romantic	青春豬頭少年不會夢到兔女郎	Rascal Does Not Dream of Bunny Girl Senpai	0.89	9.7
10	romantic	約會大作戰	Date A Live	0.85	9
11	romantic	虎與龍	TIGER×DRAGON！	0.76	9.5
12	romantic	灼眼的夏娜	Shana	0.91	9
13	romantic	緋彈的亞莉亞	Aria the Scarlet Ammo	0.86	8.7
14	romantic	狼與辛香料	Spice and Wolf	0.87	9.3
15	romantic	果然我的青春戀愛喜劇搞錯了	My Youth Romantic Comedy Is Wrong, As I Expected	0.85	9.6
16	romantic	加速世界	Accel World	0.88	9.2
17	mystery	櫻子小姐的腳下埋著屍體	Beautiful Bones: Sakurako's Investigation	0.73	9.4
18	mystery	GOSICK	GOSICK	0.88	8.8
19	mystery	冰菓	Classic Literature Club	0.76	9.6
20	mystery	Joker Game	Joker Game	0.67	9.3
21	mystery	虛構推理	In/Spectre	0.93	9.3
22	mystery	神的記事本	Heaven's Memo Pad	0.83	8.6
23	mystery	六花的勇者	Rokka: Braves of the Six	0.90	8.6

	Genre	Anime(Chinese)	Anime(English)	x coordinate (similarity)	y coordinate (popularity)
24	mystery	春&夏推理事件簿	Haruta & Chika	0.64	8.5
25	comedy	為美好世界獻上祝福	God's Blessing on This Wonderful World!	0.69	9.6
26	comedy	我的朋友很少	Haganai	0.75	8.7
27	comedy	驚爆危機!	Full Metal Panic!	0.78	9.2
28	comedy	打工吧魔王大人	The Devil Is a Part-Timer!	0.92	9.3
29	comedy	我的妹妹哪有這麼可愛！	Oreimo	0.85	9
30	comedy	零之使魔	The Familiar of Zero	0.66	8.7
31	comedy	不起眼女主角培育法	Saekano: How to Raise a Boring Girlfriend	0.83	9.4
32	comedy	笨蛋，測驗，召喚獸	Baka and Test: Summon the Beasts	0.85	8.9

2. Analysis of Similarity and Popularity

Graph 1. is the graph with 32 anime’s points plotted on it. The purple lines, which are the x-axis and the y-axis, denote the averages of all the 32 data sets(x-axis as similarity and y-axis as popularity). Red points represent action anime, blue points represent romance anime, green points represent mystery anime, and orange points represent comedy anime.



Graph 1. Different adapted anime’s similarly to their original light novels v.s different adapted anime’s popularity

(1). Action(red)

The action anime are mostly in the third and fourth quadrants, which means their anime popularity is mostly below the average. No matter how similar the anime are to their original light novels, the adapted action anime are likely to receive bad ratings. As a result, it seems like the similarity of the adapted action anime to their light novels does not have a direct effect on the popularity of the adapted action anime. This shows that action light novels would not be a popular choice for the anime production company to produce an adapted anime of theirs.

(2). Romantic(blue)

Most of the romantic anime are near and located in the first quadrant, which is the area representing higher similarity with better rating. It is clear that if a romantic anime has a high similarity to its original light novel, the adapted anime has a great chance of receiving a good rating. Since there's only one romantic anime located in the low similarity area(the third quadrant), it is unclear whether romantic anime with a low similarity would be more popular or less. For now, the experiment shows the more similar the adapted romantic anime is to its original light novel, the more popular the adapted anime is.

(3). Mystery(green)

The mystery anime are mostly distributed in the second and the fourth quadrants. This indicates that Anime with a high similarity to its original light novel has a low rating while anime with a low similarity has a high rating. Therefore, it seems like the more similar the adapted mystery anime is to its original light novel, the less popular the adapted anime is compared to the average popularity.

(4). Comedy(yellow)

The comedy anime are scattered throughout the graph. They do not show a clear relationship between popularity of the anime and the similarity of adapted anime with original light novels. In conclusion, the popularity of the adapted comedy anime may vary regardless of the amount of changes being made from the original light novel.

3. Special Cases

1. Rascal Does Not Dream of Bunny Girl Senpai/青春豬頭少年不會夢到兔女郎

As one of the romantic anime, Rascal Does Not Dream of Bunny Girl Senpai has the highest popularity among all the 32 anime. According to 不專業的動漫迷 from Medium, "The anime has a powerful, rich storyline. The foreshadowing plot is hard to guess and it often has unexpected endings."[7] Still from his blog states that

[7] https://medium.com/@takuto.novel/%E9%9D%92%E6%98%A5%E8%B1%AC%E9%A0%AD%E5%B0%91%E5%B9%B4%E4%B8%8D%E6%9C%83%E5%A4%A2%E5%88%B0%E5%85%94%E5%A5%B3%E9%83%8E%E5%AD%B8%E5%A7%8A-%E5%8B%95%E7%95%AB%E5%BF%83%E5%BE%97-%E5%BE%88%E6%B5%AE%E8%AA%87%E7%9A%84%E5%90%8D%E5%AD%97-%E8%AE%93%E4%BA%BA%E6%B7%9A%E6%B5%81%E6%BB%BF%E9%9D%A2%E7%9A%84%E5%8A%87%E6%83%85-cb0b30fde0fb

“he thought the plot of this anime was just as normal as the other anime. However, he did not expect the anime’s ending. Overall, the adapted anime has a smooth flow with thoroughgoing design of the storyboard, which included more details from the original light novel.”[8] Based on these two comments, the particular reasons for the anime to receive the highest rating are the unexpected ending with rich storyline, smooth flow, and the included details from the original light novel.

2. God's Blessing on This Wonderful World!/為美好世界獻上祝福

The comedy anime, located in the second quadrant, has the lowest similarity with the highest popularity. 拾部次元 from 壹讀 wrote a review saying that “the anime isn’t worse than the original light novel and the comic version but more interesting since the characters look more silly and dumb in the anime.”[9] ga652206 from My PTT left a comment under the article saying “the anime is relatively quick to absorb, and the novel is easy to read too.”[10] The changes made in the anime seem to have a positive effect on the anime’s popularity. They have made a better, more interesting version and clear image that the light novels can’t provide to the audiences.

3. Haruta & Chika/春&夏推理事件簿

As a mystery anime, while other mystery anime are either in the second quadrant or the fourth quadrant, only Haruta & Chika is being displaced in the third quadrant. Its position shows the anime has the least similarity to its original light novel among the 32 anime and a bad rating compared to the average popularity. In Bahamut, 閒雲 discuss the reasons why he was disappoint with the anime. He explained “the anime’s mediocre plot and unimpressive characters cause the story to be incomplete.”[11] 肉肉 down in the comment section said “the anime doesn’t put its focus on reasoning and suspense while the original light novel does a good job focusing on the reasoning and related information about music.”[12] These comments reveal Haruta & Chika’s lack of details from the original light novel and the

[8] https://wellnever2392.pixnet.net/blog/post/352606502-%E9%9D%92%E6%98%A5%E8%B1%AC%E9%A0%AD%E5%B0%91%E5%B9%B4%E4%B8%8D%E6%9C%83%E5%A4%A2%E5%88%B0%E5%85%94%E5%A5%B3%E9%83%8E%E5%AD%B8%E5%A7%8A-%E5%8B%95%E7%95%AB%E8%A9%95%E8%AB%96

[9] http://read01.com/kzRmyde.html

[10] https://myptt.cc/article/C_Chat/M.1493981497.A.353

[11] https://home.gamer.com.tw/creationDetail.php?sn=3113516

[12] https://home.gamer.com.tw/creationDetail.php?sn=3113516

incomplete storyline caused the anime to receive a bad rating while the other mystery anime with little similarity have good ratings.

4. Arifureta: From Commonplace to World's Strongest/平凡職業造就世界最強

The action anime has the worst rating among all the 32 anime. According to 寒玥 in one of the article from Bahamut, "the anime has deleted too much plot from the original light novel and deteriorated most of its plot. The production team seems to not understand the importance of the flow of the story and the amount and content of plot's adaptation. It's a pity the adapted anime is not as expected since the original light novel was a good novel." [13] The original light novel has a higher rating than the adapted anime. The production company has changed some unnecessary changes or there's limited funds. This causes audiences who have very high expectations on the adapted anime due to the original novel becoming disappointed.

IV. Conclusions

Japanese light novels have been a popular option to adapt and create anime. However, to our best knowledge, there has been no research investigating whether the similarity between original light novels and adapted anime affects the popularity(rating) of the adapted anime. This study uses 32 online original light novels that are published after 2000 and 32 corresponding online adapted anime episodes' summaries to find the key terms and calculate the similarity between. To find the similarity, text mining techniques are being used to break and compare the key terms in both anime episodes' summaries and light novels. With a graph and 32 points, which each represent an anime's similarity and popularity, the result shows different genres of anime have different relationships of similarity and popularity. The similar the romantic anime are to their original light novel, the more popular the adapted romantic anime are. Whereas the similar mystery anime are to their original light novel, the less popular the adapted romantic anime are. Comedy anime's popularity varies regardless of the similarity between the anime and its original light novel while most action anime receives bad ratings regardless of the similarity.

[13] https://forum.gamer.com.tw/C.php?bsn=46791&snA=207

In order to find whether similarity between adapted anime and light novels affects the adapted anime’s popularity, the study utilizes terms in anime episodes’ summaries and online light novels to calculate the similarity and analyze the results with each anime’s popularity. The following are the findings based on the research:

(1) Construct a new algorithm with the use of Boolean value, sigma, and average to calculate the similarity between original work and adapted work.

Since the anime episode summary is only a paragraph, which is short, while the online light novel chapter has multiple pages of words, which is long, the research cannot use cosine similarity to calculate the similarity between adapted anime and original light novel. Therefore, this study divided the terms in each sentence of the episode summary based on the CKIP system and deleted unnecessary words according to Chinese Stop Word. Then calculate the percentage the remaining terms in the sentence of the episode summary appears in the corresponding online light novel chapter. The percentage will become the similarity between the anime episode summary’s sentence and the light novel. The average of the percentage will become the similarity between the episode of the anime and the light novel. Similarly, the average of the episode similarity will become the similarity of the whole anime and its original light novel.

(2) Each genre has different relationships between the faithfulness of the anime adaptation and the series' popularity.

According to the graph, it has appeared that the similar the romantic anime are to their original light novel, the more popular the adapted romantic anime are. On the contrary, the similar the mystery anime are to their original light novel, the less popular the adapted romantic anime are. Comedy anime’s popularity varies regardless of the similarity between the anime and its original light novel while most action anime receive bad ratings regardless of the similarity.

(3) Anime production companies should determine the degree of changes being made for the adapted anime from its original light novel based on the genre of the light novel.

Based on the previous findings, since different types of anime have different relationships between the faithfulness of the anime adaptation and the series' popularity, the companies have to determine the degree of changes based on the genre of the light novel. When adapting romantic light novels, which would be the most popular choice to be adapted based on the graph, the anime production companies should preserve most of the original plot and details in the anime with a smooth storyline. Oppositely, the anime production companies should keep in mind the fewer adaptations being made in the mystery anime alone with a complete storyline, the better the popularity/rating the anime will receive. Since the popularity of comedy anime don’t have a direct relationship with the similarity, the production companies only need to be aware of the quality of the story, the flow, and other factors that might affect the rating. Finally, action anime, which is not a popular choice to be adapted based on the graph, need to include a complete smooth storyline, deeply portrayed characters, and an understandable plot.

REFERENCES

Wu, Y. T. “The Study of Fei-Yu Bi’s Novels and Adapted Movies of His”, Master dissertation, National Tsing Hua University, 2018.

Liao, F. J. “The Intertextuality Study Of Liao Hui-Ying’s Novels and Film Adaptations：The Case of Ah Fei, No Return and Drizzle Tonight”, Master dissertation, National Chiayi University, 2017.

Huang, C. J. “The Influence of Consumer Lifestyle to the Buying Behavior－the Case of Moe ACG Products”, Master dissertation, Ming Chuan University, 2010.

Tu, Y. N., & Wang, Y. S. Wang. “Use Text Similarity to Determine Whether a Movie Review is Spoiler or not - Take the PPT Movie Bulletin Reviews as an Example”, paper presented at the 31st International Conference on Information Management, Chiayi City, Taiwan (R.O.C), 2020, Dec.

Tu, Y. N., & Lin, W. Y. “Identifying the Relationship Between Characters and Characters by Using Text Dialogue in Novels-Taking Ruyi’s Biography as an Example”, paper presented at the 31st International Conference on Information Management, Chiayi City, Taiwan (R.O.C), 2020, Dec.

文字探勘技術應用於高度不平衡長度文本的相似度分析 － 以動畫改編與其原著輕小說為例

杜逸寧・李芃萱*

摘要

在文字探勘模型訓練中，相似度計算的計算方式將會對結果產生高度影響。特別是眾所周知的餘弦相似度。尤其是當所比較的文本長度高度不平衡時，詞袋將會非常大，且在詞向量矩陣中不容易具有相同的詞彙。即使存在匹配，詞彙也會受到兩個文本之間眾多不同文本的詞彙數量影響而被稀釋。本研究提出了一種高度不平衡的相似度算法並試圖解決這些問題。在實務上本研究提出了一個改編動畫與其原著輕小說之間的案例研究。新提出的相似度將取代傳統的餘弦相似度來處理這類案例研究。本研究使用了 32 部 2000 年後在網路上發表的原創輕小說，以及 32 集對應的線上改編動畫情節摘要，以找出關鍵詞並計算它們之間的相似度。結果顯示不同類型的動畫之間具有不同的相似性和受歡迎程度關係。此外，本研究也根據分析結果針對不同類型的動畫改編提供相對的應的行銷策略。

關鍵字：文字探勘、高度不平衡長度文本的相似度、改編動畫、輕小說

*作者簡介：杜逸寧，天主教輔仁大學統計資訊學系副教授（通訊作者）；李芃萱，日本早稻田大學基礎科學與工程學系(大一生，2022 年)。