

Data visualisation analysis of online novel websites

Meijuan Zhuo *

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: 1733530557@qq.com

Abstract. With the development of science and technology and the progress of the times, the online novel reading industry is also developing at a rapid pace, and online novel reading websites have also emerged. In order to make users better understand the current trend of popular novel genre popularity and give corresponding reference when users choose the novels to read. This study analyses novel creation data, novel creation genres, novel creation era, author writings and other data visualisation methods through bar charts, pie charts, radar charts, heat maps, word cloud charts and other data visualisation methods. Using web scraper, data on novel information is collected to ensure that the data is authentic and reliable. On the basis of cleaning the data using Python, the Echarts plugin is used to visualise the basic information of novels as well as the author's information to present the information of novels and readers in a more visual and vivid way. The study helps to obtain richer information about the novels and authors that one wishes to know, while making the readers' experience better and helping the website to operate and develop better.

Keywords: Web scraper, Novel reading website, Data visualization.

1. Introduction

Reading can develop people's interest in learning, enhance their language skills, increase their knowledge and develop the ability to think independently. In today's world, with the development and promotion of Internet technology, people are gradually changing from paper-based reading to electronic reading. The widespread use of Internet carriers has provided many channels for reading online novels. Powerful search engines, easy browsing methods, VIP reading systems, and rich reading resources are all attracting people to read online [1]. We often see people reading novels on mobile phones or through websites while waiting for public transport, on the underground or on other occasions. Compared to traditional paper-based reading, e-reading has many advantages, as it helps people to read books in fragments, using mobile phones or computers instead of heavy books, which is not only environmentally friendly but also improves the reading rate of books. At the same time, readers can find books in a variety of genres without having to worry about paying for expensive paper books, and they can also share their thoughts on what they are reading. As a result, online reading is becoming more and more accepted and popular. John H. Schuman, an American scholar, proposed the theory of cultural adaptation model, emphasising the importance of the degree of cultural adaptation [2]. As one of the carriers of Chinese culture, it is particularly important and necessary to analyse the data on Chinese online novels. And with the rapid development of internet technology, we have entered the digital age, which has also given rise to many social media and search engine companies.

Social media is becoming more and more powerful, both at home and abroad, and therefore generates a huge amount of data, which is part of Big Data. While traditional techniques cannot directly analyse the huge amount of data, visualisation techniques make it easier for humans to understand and accept the raw data and information transformed by computers [3]. In this context, data visualisation and analysis has become a popular direction for development and research. Data visualisation is the use of technology to display seemingly unconnected data in the form of charts or graphs on a user interface after processing by relevant algorithms, and it has a wide range of applications. In 2005, Thomas JJ et al. mentioned the use of visual analytics to discover the intrinsic connections of information and the use of interactive graphics to improve human-computer interaction [4]. Data visualisation has emerged with the advent of the big data era, and it has become

an indispensable application for big data analysis. Both at home and abroad, visualisation research has received a great deal of technical support and application. In *The Beauty of Data Visualisation*, a detailed demonstration of what visualisation can achieve and how it can be used to change the world is given in [5].

This study focuses on the analysis of novel data from 2003-2020 on the Jinjiang Literature website, aiming to be able to give some reference when users want to read or study online novels. All data was obtained through web crawling to ensure the authenticity of the data.

2. Methods

Visualisation is a technique that uses graphical images to depict the connections between seemingly unrelated data. It makes complex data more intuitive, makes it easier to communicate, analyse and transfer data, makes it more vivid and interesting to represent data in a viewable form, helps people to understand the data better, and helps in human-computer interaction. Visualisation encompasses many aspects, such as knowledge visualisation, information visualisation and also data visualisation. We will focus on data visualisation. Data visualisation technology mainly deals with data stored in databases. Through the technology of graphing and videoing data, it can represent the originally complex data with visual elements such as images, text, shapes and colours, which makes some of the laws and properties in the data that are not easy to discover emerge, and also makes it possible for data information to be conveyed more accurately and efficiently. At the same time it can bring more freshness and attract the attention of the viewer in the application of reading websites. Data visualisation has three characteristics: interactivity, visibility and multidimensionality. In data visualisation and analysis, using the right data visualisation tools can help us to be more efficient. This study uses pyecharts, which queries and retrieves the required data from processed csv files to create dynamic charts.

3. Results

3.1. Data

In this paper, a total of 268,262 raw data of novels, authors and reviews were collected from the website of "Jinjiang Literature" using web scraper. These data provide reliable data support and basis for the subsequent data pre-processing and data visualization. In order to ensure the reliability of the analysis results, the collected raw data were cleaned and filtered. Novels with less than 10,000 words were removed, and novels with unknown authors were filled in with "anonymous". The novel creation time and novel genre were standardised, and the novel genres were specified as major genres and minor genres, with minor genres being expanded under major genres, resulting in a total of 160,719 pieces of valid data.

3.2. Results of the analysis

(1) Novel major genre analysis

In the novel major genre analysis pie chart, the novel major genre and the total number of novels in each genre were used as a set of data to create the novel major genre analysis pie chart using ECharts. As shown in Figure 1, the major genres of novel are Romance novel, Pure Love novel, Lilies novel, No CP novel, and Female novel. The names of each genre and their corresponding number of novels as a percentage of the total number of novels are shown in the graph. The size of the pie chart they occupy is determined by the proportion of the number of novels in each broad genre to the total number. For example, the number of Romance novels is 94,470, the largest of all genres, accounting for 58.85% of the total number of novels, more than half of the total number of novels. Female novel, with only 1,439, is the least, and therefore the romance genre has the most weight in the pie chart, and is the most prominent and obvious in the pie chart, with the colour shown in red.

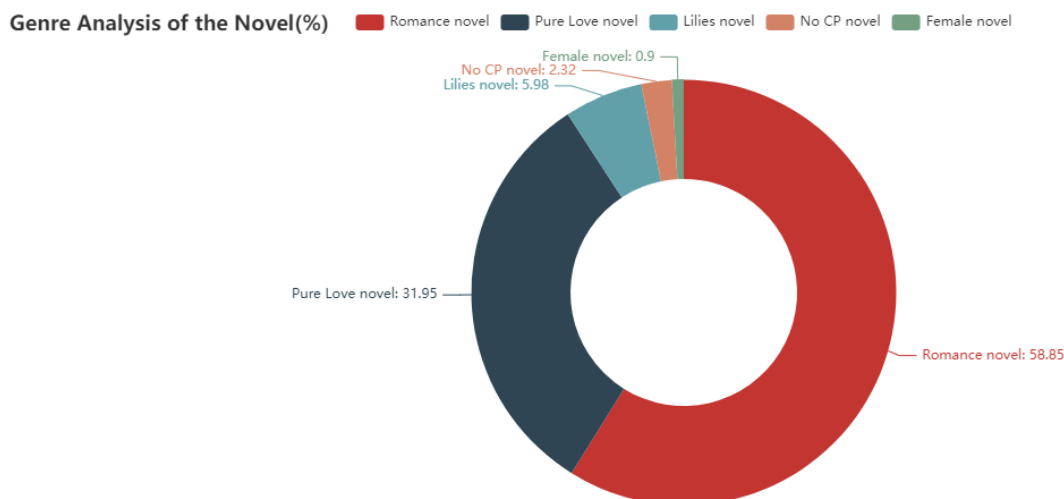


Figure 1. Pie chart for major genre analysis of novels

(2) Analysis of minor genres of novels

For the four major genres of novel in (1), each major genre can be subdivided into 12 sub-genres, including Love, Saga, Thriller, Drama, Science Novel, Fantasy, Light Novel, Fairy Tale, Wuxia, Immortal, Suspense, Game. The genre of the novel is used as the horizontal coordinate and the number of novels in each genre as the vertical coordinate. Depending on the genre division, the two bubbles of novel genre analysis in Figure 2 can be plotted. The colour of the bubbles deepens from blue to red depending on the number. The number of novels in the romance genre is the highest in both genre divisions of the analysis graph, so the genre appears as a red bubble in the graph, which is the most bright and obvious, while the genre with the lowest number of novels is shown very inconspicuously in the graph.

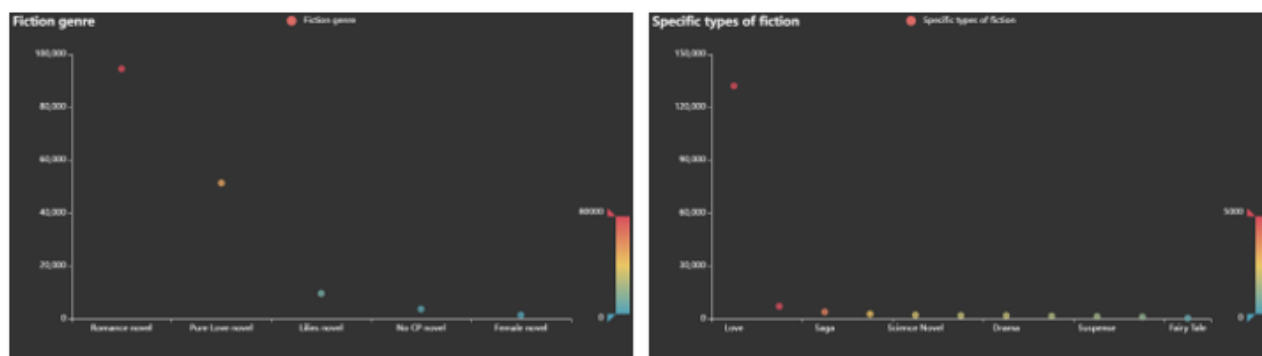


Figure 2. Bubble chart for major and minor genre analysis of novels

(3) Analysis of major and minor genres of novel

According to the major and minor genres in (1) and (2), the distribution of the number of minor genres under each major genre is plotted in a bar chart for the analysis of major and minor genres of novel. In Figure 3, the number of romance novels is the highest of all the secondary genres in the main genre of 'lilies', while the very small number of genres is not obvious in the bar chart. The same method can be used to analyse the data for the other major genres and to compare the distribution of novels in the minor genres under the different major genres. The romance genre shows the most obvious prominence under each of the major genres, with a very high number of data being created, while the number of novels in the fairy tale, thriller and science novel genres are all relatively small. Therefore, it can be concluded that most authors choose to write novels in the "romance" genre, and that users who like romance novels are more likely to find novels that suit their reading tastes on the Jinjiang Literature website. However, readers who like fairy tales, thrillers, games and other genres can try to find novels suitable for them on multiple reading sites.

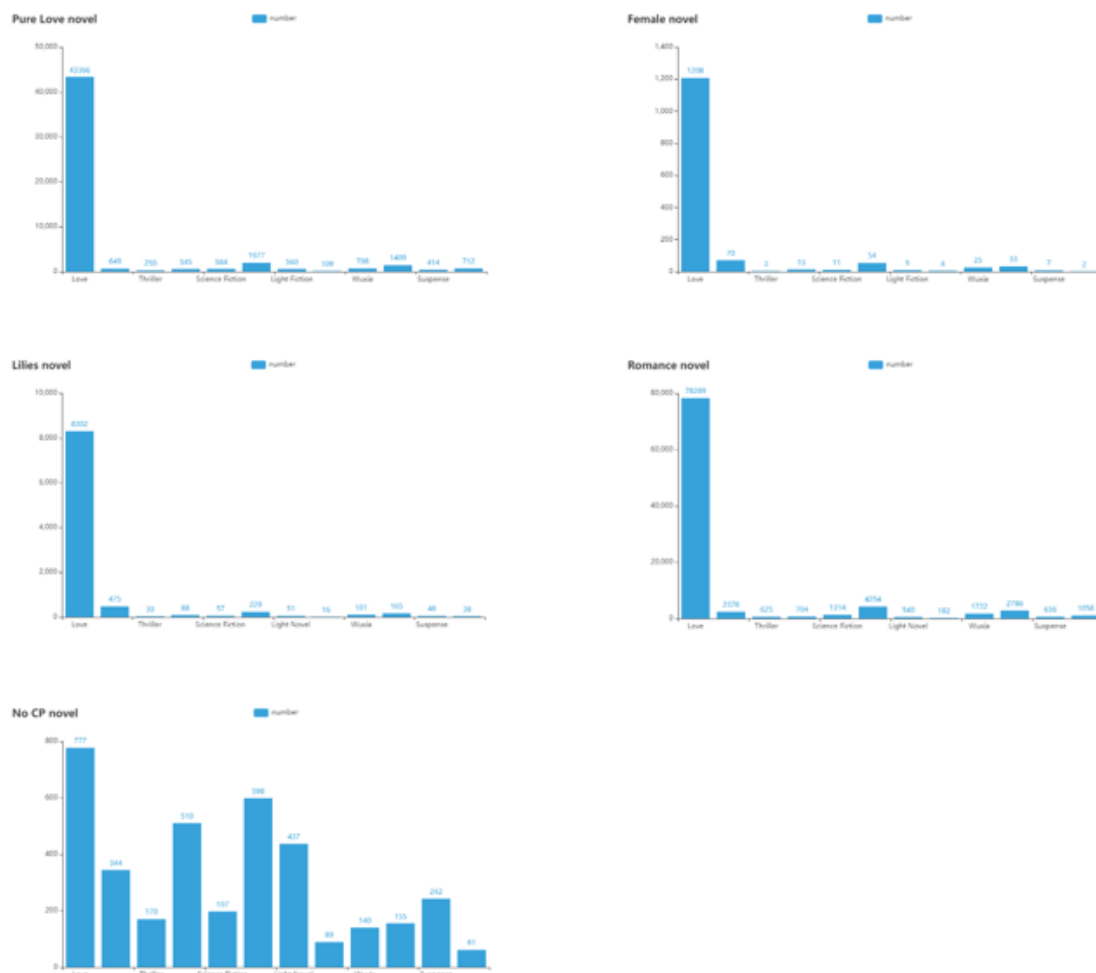


Figure 3. Bar chart for genre analysis of novels

(4) Analysis of the novel setting era

As in Figure 4, a pyramid chart of novel era analysis was created using ECharts using the novel era and the total number of novels in that era as a set of data. For example, novels are set in eras such as Ancient, Modern Modern, Fantasy Future, and Overhead History. The pyramid chart is a graph generated based on the number of these novels set in these eras, with less at the top and more at the bottom. The novels from these different eras are arranged in order of their number, with novels set in modern contemporary being the most numerous and therefore it is at the lowest end of the pyramid chart, while novels set in fantasy future are at the top of the pyramid chart due to the very small number of novels set in these eras. The scale of each part of the pyramid chart is also made in proportion to the number of novels in each different setting.

Analysis of the period in which the novel is set

Fantasy future novels Modern Fiction Novels with a hollow history Ancient genre novels

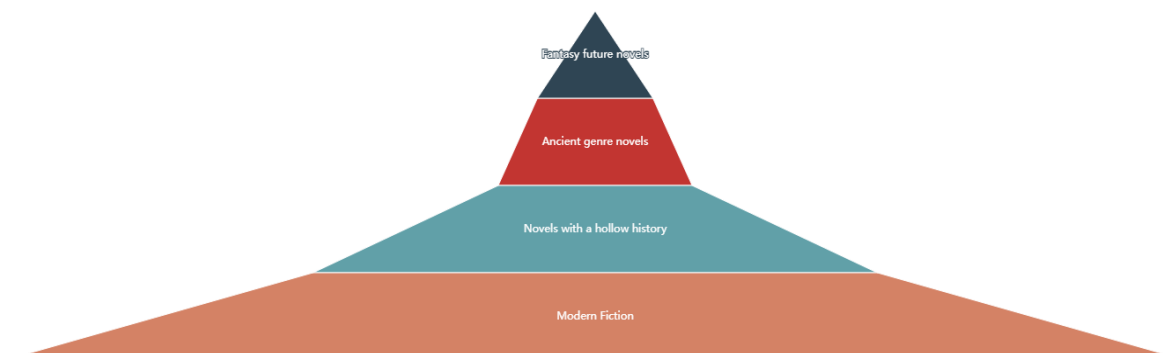


Figure 4. Pyramid chart for the analysis of the period of the novel's setting

(5) Analysis of novel creation from 2003-2020

As shown in Figure 5, a heat map for the analysis of novel creation from 2003-2020 was created using ECharts with the year of novel creation, the month of novel creation and the total number of novels in that month as a set of data. The horizontal coordinates of the heat map are the years 2003-2020 and the vertical coordinates are each month from January to December. Each coordinate in the heat map indicates the number of novels created in that month of that year, with the exact number shown above. At the same time, the background colour of the coordinates gradually changes from yellow to red depending on the heat of the novels created in each month, with the higher the number of novels created in the month, the redder and more vibrant the colour of the coordinates, and the lower the number of novels created in the year, the more yellow the colour of the coordinates. A total of 2,690 novels were created in March 2020, the month with the highest creative fever, while November 2003 and June 2004 were the months with the lowest creative fever.

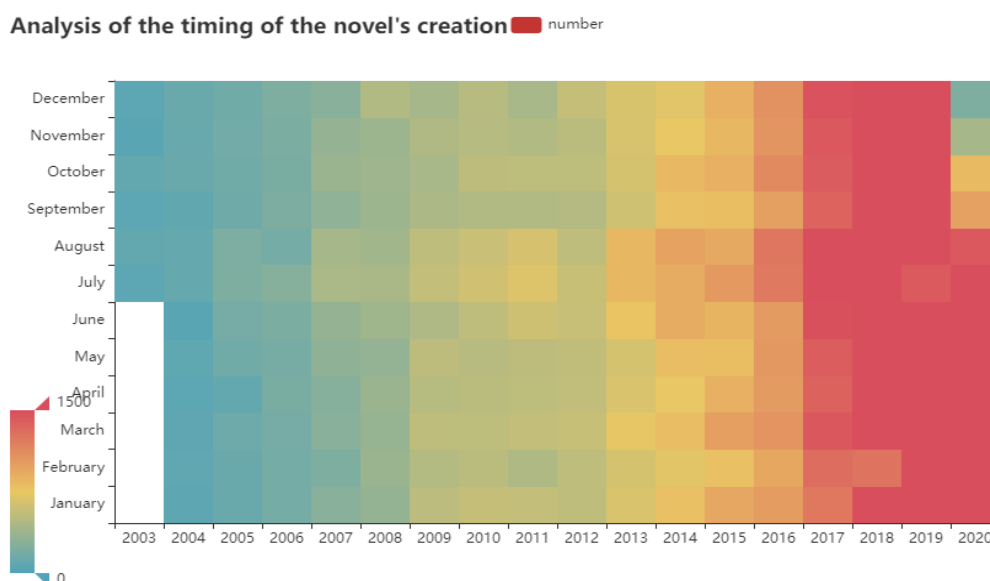


Figure 5. Heat map for the analysis of novel creation from 2003-2020

(6) Author contribution analysis 2003-2020

As shown in Figure 6, the data was processed using Python to create an author contribution analysis word cloud graph for the years 2003-2020, using the author and the number of novels created by the author as data. In the resulting author contribution graph, the more novels an author has written, the more they have contributed to the site and the more they stand out in the graph. For example, in Figure 6, the author "Nanzhi" and the author "Qiu Ling" are in larger font and more prominent, indicating that they have written more novels and contributed more to the website.



Figure 6. Word cloud diagram for author achievement analysis

4. Conclusions

This experiment is based on the data of novels and authors from the Jinjiang Literature website, and realises operations such as the collection and cleaning of novel data, and visualises and analyses the data through visual charts such as word cloud charts, pie charts and line charts. The visualisation results clearly show the current trend of novel creation and the popularity of various genres. Over 90% of all genres are pure love genre or romance genre novels, where in every year of novel creation, romance genre novels are highly popular for authors to create and readers to read. The boom in the creation of online novels increased between 2003 and 2020, and from 2017 onwards, the boom in the creation of novels continued to crescendo and continue to increase, thus also revealing an increase in people's love for online novels. The analysis of the number of assignments created, read, etc. can also help us understand the quality of the author's creations, how much they are liked, etc. This visual analysis of data from novel reading sites helps users save time in finding novels they might like, while also providing some reference for users who study online novels.

References

- [1] Ma Ji. The Course of Discourse Transformation: A Review of Ten Outstanding Online Novels of the Past Decade [J]. Social Sciences in China, 2011, (1):166-181.
- [2] John H. Schumann. Research on the Acculturation Model for Second Language Acquisition [J]. Journal of Multilingual and Multicultural Development, 2010, 7(05): 379-392.
- [3] Haber R B. Visualization idioms: A conceptual model for scientific visualization systems [J]. Visualization in Scientific Computing, 1990.
- [4] Thomas JJ, Cook K A, Electrical I. Illuminating the path: The research and development agenda for visual analytics [J]. Computer Graphics, 2005.
- [5] Julie Steele. Beautiful Visualization [M]. Beijing, 2011.