Modeling and analysis of the social network of Dream of Red Mansions based on natural language processing

Yueming Liu, Chiawei Chu*, Jiaheng Yang, Haokai Deng and Yifeng Hua

Faculty of Data Science, City University of Macau, Macau China

*Corresponding author's e-mail address:cwchu@cityu.edu.mo

Abstract. With the development of machine learning technology gradually deepening, the research and application of relationship mapping has also made great progress. Therefore, the analysis of character relationship mapping using social network model breaks the limitation of the traditional storage medium of character relationship, and the data obtained from the integration is used to illustrate the primary and secondary of the characters in the article with the primary and secondary of the data in a visual expression, so as to constitute the character relationship mapping, which is expressed in the form that is more close to the world of human cognition, and provides the readers with a convenient and concise access, and the readers can have a clearer understanding and awareness of the article's The reader can have a clearer understanding and awareness of the character relationships and the main story line in the article. Readers can have a clearer understanding of the character relationships and story lines in the article.Literary works have been one of the research focuses in the direction of digital humanities, and the relationship construction of literary works is carried out through the way of complex networks. This thesis takes Dream of the Red Mansions as the object of analysis, adopts the model of natural language processing for data extraction, and analyses the structure of the article through the method of text analysis. By studying the main relationships of the characters in the novel, the social network model is adopted for modelling and analysis, focusing on the complex network analysis method for construction. To a certain extent, it describes the characters' character as well as the storyline in the novel. This thesis finally obtains the character relationship diagram of Dream of the Red Mansions, which is analysed by further experimental results, which show that the character relationship diagram of the text of Dream of the Red Mansions has small-world characteristics.

Keywords: Natural language processing; Social network; Data Mining; Digital Humanities;A Dream of Red Mansions

1 Introduction

With the continuous progress of computer network technology in people's lives and the booming development of machine learning technology in the field of natural language processing, the relational mapping analysis technology has also made great progress in

K. Subramaniyam et al. (eds.), *Proceedings of the 3rd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2024)*, Atlantis Highlights in Computer Sciences 21,
https://doi.org/10.2991/978-94-6463-419-8_21

research and application. Natural language processing for text analysis has become one of the important means of dissecting articles. The linguistic text is analysed by natural language processing to get the corresponding relations, and the text is analysed in a computerized way. It becomes a new way of storing character relationship mapping through the establishment of complex networks. In this paper, we choose the social network in the complex network to build the model, social network is a kind of network structure based on the interaction between people, through which we can understand the interconnections and interactions between individuals or group members. Social network analysis is usually applied to deeper levels in various fields. Social network, as the most commonly used analysis method, presents people's interaction behaviour and social relationships in graphical form, providing a more realistic and comprehensive analysis platform. In this paper, we select the topics in Dream of the Red Mansions to start the research, use social network analysis to focus on the research object, and model the literary works through the method of big data comparison.

Chinese culture is as bright as the stars, and the immense amount of knowledge in the unprecedented work Dream of Red Mansions is unimaginable, so academics also set up the Chinese Society of Dream of Red Mansions as a way to specialize in the study of Dream of Red Mansions.Cao Xueqin, the author of A Dream of Red Mansions, encouraged himself with "talking about love in a big way, recording his own affairs", only according to his own affairs, following the traces, breaking away from the stereotypes, refreshingly chic, and achieving extraordinary artistic achievements[1].

2 Related Work

2.1 Natural language models

Natural language model, also known as statistical language model, is a model that describes the probability distribution of natural language in computer science. Using a language model, you can calculate the probability of a sequence of words or sentences,or give a probability distribution of possible words in the case above. In this paper, the natural language model is used to identify the occurrence frequency of the characters of Dream of Red Mansions, so as to realize the search of common frequency, which provides a basis for the construction of the following social network model[2].

2.2 Text Analysis Word Frequency Relationship Building

Chinese text analysis work is very different from English text analysis work[3]. In the text pre-processing process, one of the main differences is the step of word separation. Chinese text can not easily achieve the purpose of word separation with the help of word space, but to cut the words according to the Chinese semantics. Therefore, a special package is needed for Chinese text analysis. As this paper is a traditional Chinese masterpiece Dream of Red Mansions, so take the text analysis in the use of more packages such as JiebaR.

2.3 Interdisciplinary Application of Social Network Modeling

In recent years social relationship networks have gradually become a hotspot for interdisciplinary research, Zhao Jingsheng [4]used complex network technology and natural language processing to extract and analyze the social network of the literary work Romance of the Three Kingdoms,Tang Yi[5]used text mining, constructed character relationship matrix, and used social network method to examine the social network of 108 main characters in Water Margin: counting the number of times Ren and Li get along, and constructing character relationship network.

2.4 Social network model for the construction of Dream of the Red Mansions

In this paper, the text of Dream of the Red Mansions is data mined[5] and a new novel character relationship network is established based on its constructed characters. The constructed character relationship network is analyzed, and natural language is mainly used to deal with the characters. This thesis takes Dream of the Red Mansions as the object, that is, the network system as the basis, and it is meaningful to study the behavior of the characters under this network system, and get the word cloud extraction[6], and carry out the work of constructing the character relationship mapping on the basis of this. Since the persona relationship mapping is a huge structure rather than a simple node or edge, to understand the whole system it needs some tools to accomplish it. So the method of social network analysis is adopted, which enables us to coalesce the complex network system into a clear and explicit visualization graph.

2.5 Innovative points of social network modeling to construct the text analysis of Dream of Red Mansions

This paper analyses character relationship mapping using the social network model, which overcomes the limitations of traditional media in storing character relationships, integrates data obtained from visual expressions [7], and uses these data to collate the primary and secondary character relationships in the whole text, thus constituting a character relationship mapping, whose form of expression is closer to the cognitive world of human beings, and provides the public with a more convenient and concise way of accessing the text, which facilitates the readers to understand the character relationships and character relations in the text. relationships between characters and their connections.

Through reviewing data and literature, we found that Dream of Red Mansions does not visually show the relationships between the characters in the book, so this study adopts the social network model of character relationship analysis to analyse the character relationship mapping, and constructs a character relationship mapping based on the characters in Dream of Red Mansions.

2.6 Co-word analysis-Natural language processing

The final step in text structuring is the construction of the discourse-document matrix. Most text analysis algorithms are based on word vectors rather than word frequency statistics. In this paper, text is extracted through knowledge related to natural language processing to construct data for social network models. The text content is extracted by natural language processing. Character names are extracted from the full text of Dream of Red Mansions, and the corresponding texts are selected and processed according to the data. In this paper, we mainly use Python toolkit Pandas to process the data, build the co-occurrence matrix and perform co-occurrence analysis. Through this method, the proximity and distance relationship between two keywords in the matrix can be seen. In this paper, the co-occurrence of keywords whose letters are located in the same section is denoted as "1" and vice versa as "0".

3 Social Network Modeling

3.1 Relationships between characters through natural language processing

In this paper, the natural language model is used to extract the relationship between the characters in the novel, and some special symbols and stop words, entity alignment and attribute alignment are removed after preprocessing the text. In the establishment of the relationship between the main characters of Dream of Red Mansions, the relationship between the main characters of the novel is sorted out through this kind of relationship, and the relationship table of the main characters of the novel is obtained for further analysis[8]. Figure 1 shows the character relationship data diagram of Jia Baoyu, the male protagonist in Dream of Red Mansions. This package can complete the Chinese text preprocessing under the premise of lexical annotation. Table 1 represents the frequency of interaction between the characters of Dream of Red Mansions.

Table 1. Relationship diagram of the main characters (taking Jia Baoyu as an example)

Source	Target	Weight
Baoyu	Daiyu	3616
Baoyu	Tanchun	1186
Baoyu	Yingchun	298
Baoyu	Xichun	408
Baoyu	Baochai	2812
Baoyu	Qinshi	130
Baoyu	Liwan	620
Baoyu	Fengjie	1844
Baoyu	Yuanchun	37
Baoyu	Xiangyun	868
Baoyu	Qiaojie	47
Baoyu	Miaoyu	527

3.2 Data source of Dream of Red Mansions

This paper takes the main characters of Dream of Red Mansions as the research object to construct a social network, and uses a complex network analysis method to analyze the relationship between the characters. First of all, on the basis of the existing literature query[9], this paper does not analyze and dissect all the characters, but models and analyzes the relationship between the main characters in the book. Construct it according to the relationships of the main characters in this article. The characters in the work represent the nodes in the network. Secondly, on this basis, a network model containing all text information and attribute information is further established, which is composed of three parts: nodes, edges and domains. Secondly, the most important edge is selected among these edges and its weight value is calculated[10]. Although Dream of Red Mansions is rich in chapters, here, this paper analyzes and introduces the overall relationship between the main characters, calculates these edge weights to determine the importance of each character in the whole network, and then uses the weighting theory to connect them to form a complete model of the relationship between the characters and the characters. The digitized table of Dream of Red Mansions is shown in Table 2.

Table 2. A digital form of "Dream of Red Mansions"

Title of the work	A Dream of Red Mansions
node	13
edge	67

3.3 Character Relationship Determination

The social network graph is represented as $G = (V, E)$, where,Vdenotes the set of nodes in the graph and E denotes the set of edges of the graph. Based on this then the core roles and non-core members of each person are obtained using the complex network association structure algorithm. This thesis conducts a research on character relationship extraction. There are two steps in total, the first step is to use the characters of the work constructed in this thesis as a corpus [11], if two characters enter the chapter at the same time, there is 1 relationship between them, based on this assumption, we get the number of character relationships in each chapter. In the second step, the weight value is calculated, the higher the weight value, the closer the relationship between these two people.

3.4 Visualization Tools for The Dream of Red Mansions

Undirected weighted social network is constructed in the paper, for the dataset in the paper, first of all, Gehpi software is used to realize the visualization, Gephi has developed a set of complex network analysis software with JVM as the platform, which can realize the data visualization and the analysis of the network indexes, and it can be used to carry out the functions of data analysis, link analysis, etc., and social network analysis, etc[12].

3.5 Introduction of related formulas

There are several network calculation values involved in the social network model, and the two most important formulas in this paper are introduced here.

3.5.1 Degree.

Degree distribution is defined as ranking the degree values of the nodes in the network from smallest to largest and counting the proportion of nodes with degree value k to the number of nodes in the entire network $p(k)$:

$$p(k) = \frac{Nk}{N} \quad (1)$$

where Nk represents the number of nodes with degree k and N represents the total number of nodes in the network.

3.5.2 Average aggregation coefficient.

The formula for the average aggregation coefficient of the network is:

$$\frac{L}{L_{max}} = \frac{2L}{N(N-1)} \quad (2)$$

where L is the number of edges that actually exist in the network[12], and N is the number of nodes in the network. The Shortest Path value is the path with the shortest length between any two points[13].

3.6 Analyzing Predictions

3.6.1 Cluster Analysis.

In this paper, we mainly adopt PageRank algorithm, which is originally a method of calculating the importance of Internet web pages, any directed graph can be defined, and then used in the analysis of social influence, text summarization and many other problems.

3.6.2 Prediction of Results.

For the final result, we take multiple indicators to analyze together, and by analyzing the results of the network indicators, we explore whether the network diagram conforms to the small-world characteristics in order to further contribute to the computer discipline and literature[14].

4 Experimental Results

4.1 Visualization of Dream of the Red Mansions

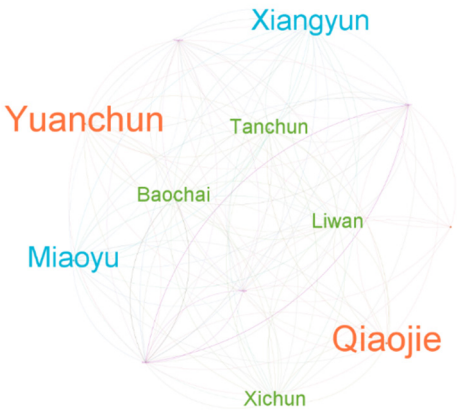

Fig. 1. "Dream of Red Mansions" visualization

As can be seen from the figure, the characters in Dream of Red Mansions, such as Jia Baoyu, Lin Daiyu, Wang Xifeng, Xue Baochai, and other nodes with large nodes correspond to a greater number of appearances in the work, and are also the main characters of the work. The main tasks depicted in the original work are associated with many other characters, as can be seen in the visualization graph, the number of nodes with edges of the main characters is much more compared to other characters. Figure 1: The relationship between people is derived from data.

4.2 Calculation of modular indicators for social networks

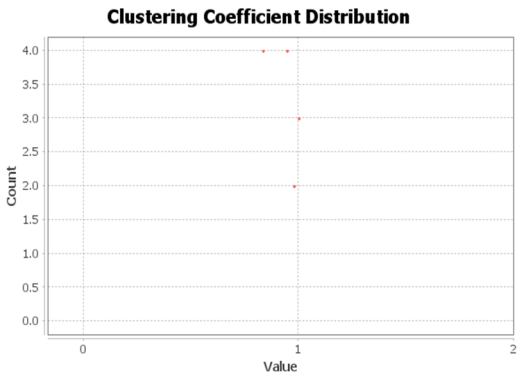

Fig. 2. Modularity of social networks

Module degreecan be used to quantitatively measure the quality of the network community segmentation, and the more its value tends to be close to 1 indicates that the strength of the community structure of the network segmented by the community structure is more and more, that is, the better or worse the division of the degree of good and bad is more and more high.

Therefore, by maximizing the modularity degree, the optimal division of the network community can be obtained The modularity index obtained in this paper is 0.929, which indicates that the modularity index is better and meets the criteria of small world. Figure 2 shows the Modularity of social network based on the data.

4.3 Calculation of the PageRank indicator for social networks.

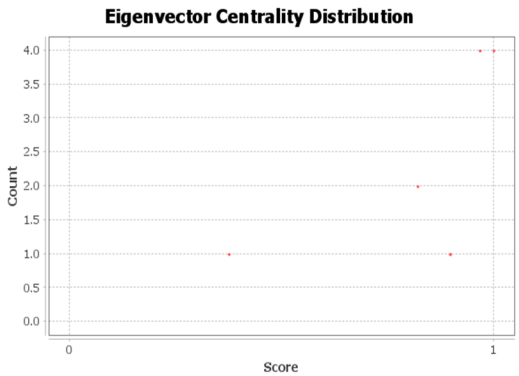

Fig. 3. Social network PageRank

PageRank is a link analysis algorithm, which assigns weights to the elements of the hyperlink collection by number, and it can be seen that the effect of PageRank in this paper is better. Figure 3 shows the Social network PageRank based on the data.

4.4 Social Network Metrics Results

Table 3. Social Network Metrics Results

Network metrics	Numeric value
Average	10.308
Average weighting	3200.000
Network diameter	2.000
Graph density	0.859
Average clustering coefficient	0.929

Table 3 provides some additional statistical characteristics calculated from the data .In this paper, we consider that the networks are all denser, and since the average path length is relatively small, only 10.308 and the average aggregation coefficient of

the network is relatively large, reaching 0.929, the character relationship network of A Dream of Red Mansions has the characteristics of a small-world network.

The diameter of the network in this paper is 2, which indicates that the character relationship of the book "Dream of the Red Mansions" and its complexity, i.e., a chain of character relationship can be passed through 2 individuals, which is also in line with the intricate characters in the writing of the "Dream of the Red Mansions" and the literary characteristics of the twisted and bizarre storyline.

In this paper, there is no clear community categorization in the book "Dream of the Red Mansions", because the characters in the book have complex relationships and the network diameter is large, so if the characters in the book make a clear community categorization instead of becoming a difficult problem. As the saying goes, "Things are grouped together, people are divided into groups", the characters in the novel can be categorized by their profiles, and different social circles can be delineated.

How can we make it easier for readers to find their favorite category? Cluster analysis solves the above problems well. Clustering makes it easy to differentiate between different categories of individuals. The book is a biographical work based on cluster analysis, in which the author analyzes various factors affecting the behavior of social interactions from a sociological and psychological point of view. In the hierarchical social environment of the book, this approach plays a role in further exploring the social relationships between the characters, excluding the status. Through this analysis, readers can learn about the characters' personalities and psychological activities, thus gaining a deeper understanding of the work. This method, in turn, can also be used for more complex social characters.

5 Summary and Prospects

With the development of the Internet and complex network technology, literary works can be analyzed with the help of complex network technology, especially social network. On the basis of social network theory, combining with the characteristics of literary texts, we construct the character relationship network model of novels based on social networks. In this thesis, Gephi is used to construct the character relationship network in the Chinese masterpiece Dream of Red Mansions, through the analysis, we can understand the relationship between characters in the work more intuitively[15], as well as the importance of the characters, which is helpful to understand the relationship of the novel characters. Calculating the network indexes in the character relationship network to determine whether it is close to the social network, through the analysis of the results of the network indexes, the study found that this network satisfies the small world characteristics[16]. In this paper, preliminary text analysis is performed through natural language processing, and co-frequency relationships are constructed through text analysis of the full text of the novel in order to achieve the construction of social networks.

By applying this network to the study of redologists, it is found to have some academic value. In the vision of literature, it gets many difficult problems that cannot be solved by exploring in the literary world, that is, how to look at the relationship of

the characters in Dream of the Red Mansions, which increases the content of the redologist research.

References

1. Fan Hongzhong, Wang Ziyue, Tao Shuang. Digital transformation and enterprise innovation: Empirical evidence based on text analysis method[J].Technology Economics,2022,41(10):34-44.)
2. Jin H O,Soowon C,Baabak A. Patterns of Skill Sets for Multiskilled Laborers Based on Construction Job Advertisements Using Web Scraping and Text Analytics[J]. Journal of Management in Engineering,2023,39(3).
3. Claire K,L. B M. Analysing community reaction to refugees through text analysis of social media data[J]. Journal of Ethnic and Migration Studies,2023,49(2).
4. Zhao Jingsheng, Zhang Li, Zhu Qiaoming, et al. Social network extraction and analysis in Chinese literary works[J].Journal of Chinese Information Technology,2017,31(02):99-106+116.)
5. Tang Yi, Wang Shuo, Hu Huan. Social Science Vertical,2018,33(04):117-120.DOI:10.16745/j.cnki.cn62-1110/c.2018.04.024.
6. Pan Debao. The spread and reception of Dream of Red Mansions in Asia[J].Dream of Red Mansions Journal,2022(06):28-40.)
7. Zhao Kai, Yao Shuhan. A review of the 2022 Academic Annual Conference of the Dream of Red Mansions Society of China[J].Dream of Red Mansions Journal,2022(06):318-337.
8. Wang Jun, He Jinrong, Ma Lerong. Computer and Modernization,2022(06):32-36.)
9. Li Shudi. Research on the text visualization design of the poetry activity of Dream of Red Mansions[D].Harbin Institute of Technology,2017.
10. Qin Guiqiu, Gu Changgui. Author analysis of Dream of Red Mansions based on sentence classification model[J].Software Guide,2021,20(04):26-31.)
11. William V,Sofia M,De V J, et al. Proposal of a Method for the Analysis of Sentiments in Social Networks with the Use of R[J]. Informatics,2022,9(3).
12. Shao Fang. Visualization of multi-dimensional character relationships in Confucianism[D].Hubei Institute of Fine Arts,2022.DOI:10.27132/d.cnki.ghmsc.2022.000132.
13. Li Zhuoyu, Ma Lerong, He Jinrong. Research on character relationship modeling based on complex network——A case study of Dream of Red Mansions[J].Modern Information Science & Technology,2021,5(03):1-4+8.DOI:10.19850/j.cnki.2096-4706.2021.03.001.
14. Wu Huiyu. Analysis of novel character relationship and social network based on Python technology[J].Computer Programming Skills and Maintenance,2020(06):61-63.DOI:10.16184/j.cnki.comprg.2020.06.021.
15. Li Xiaopei. An analysis of Xue Baozhuo's "work in calculation"——Taking the relationship between the characters in the first eighty episodes of "Dream of Red Mansions" as an example[J].Journal of Kaifeng University of Education,2019,39(07):34-35.)
16. Laya A,Ezat V. A New Fuzzy Propagation Model for Influence Maximization in Social Networks[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,2022,30(Supp02).