Automatic Extraction and Quantitative Evaluation of the Character Relationship Networks from Children's Literature works

Kun Ma*
Institute of Chinese information processing
Beijing Normal University
Beijing, China
*e-mail: 201821090021@mail.bnu.edu.cn

Lijiao Yang[§]
Institute of Chinese information processing
Beijing Normal University
Beijing, China
§ e-mail: yanglijiao@bnu.edu.cn

***Abstract*—To automate the graded reading task, we urgently need to extract and calculate the important index of the complexity of the relationship between the characters affecting the plot complexity of narrative literature. In order to realize this purpose, this paper describes a computational method for automatic analysis of the virtual social network from children's literature works. We selected the required bibliography for primary school students recommended by the Ministry of Education, then automatically extract the characters of the novel by CRF, and constructs the character network based on the co-occurrence relationship. The statistical analysis method of complex network provides a quantitative basis for distinguishing the complexity of characters' relationships in different texts. The results show that the structural characteristics of character interaction networks are similar to those of small world networks, and the selected network measurement indexes are significantly related to the complexity of text characters. Finally, we achieved effectively evaluating and predicting the complexity of the social networks from more extensive literature works some classical regression model based on machine learning.**

Keywords-Graded reading; Automatic analysis of text; Children's Literature works; machine learning; complex network

I. INTRODUCTION

The traditional vocabulary reading grading method and the evaluation of text difficulty rely too much on the evaluator's intensive reading of literary texts, and the subjective experience of literary works characters, so the efficiency is relatively low and has a strong subjectivity. In recent years data-driven analysis has emerged as a growing methodology within literary studies.

Twenty years ago, an article on the Small World Network was published in Nature and an article on scale-free networks was published in Science. Most network science researchers regard these two articles as a sign of the rise of network science. With the deepening of the theory of complex network, networks are currently being studied across many fields of science, such as Internet, social networks and even literary. The cognitive science researcher, R.Andeson shows that although the reader finally obtains the content schema by means of language schema and formal schema, once the reading is over, the deepest trace in memory is the content schema. This conclusion coincides with the graphical representation of complex networks. This paper hopes to go further on the basis of predecessors and try to use the combination of various theoretical methods to realize the purpose of automatic extraction and complexity evaluation of character interaction network of children's literature works.

This paper defines text difficulty as the subjective and objective evaluation of the difficulty of text, which combines all the quantifiable factors that affect reading difficulty. It is usually used to evaluate and sort the difficulty of reading materials. In addition to the lexical level and sentence level which are most concerned with by existing graded reading system, the textual dimension of the entire book is undoubtedly an important factor. The character relationship of narrative texts is one of the important indicators which has a significant impact on readers, especially children with poor reading ability. But even the maturest graded reading systems in foreign countries, such as: Lexile Measure, GEL (Guided Reading Level), Accelerated Reader, only achieve automatic scoring of text difficulty at the level of words and segments. The evaluation of textual level is still scored by experts. So the automatic extraction and calculation of the important indicators of the relationship complexity of characters is the one of the innovations of this paper.

Therefore, basing on the theory of complex networks, this paper is oriented to the reality of graded reading tasks and needs with the help of more mature tools and partial self-programming of machine learning. We constructed networks of characters based on co-occurrence relations for 100 narrative children's literature texts. Based on the statistical analysis method of complex network, the texts are calculated and analyzed, trying to find the network measurement index that can automatically predict the complexity of the relationships between the characters. Finally, the extraction and investigation of the relationship networks in this paper confirms the high correlation between the selected networks measurement indicators and the complexity of the textual characters through the correlation test. Finally, the classic regression algorithm of machine learning was adopted. It effectively realized the effective prediction of the complexity of narrative text characters.

Now, this article will discuss the four aspects of related work, experimental process, experimental results and analysis, conclusions and follow-up work.

II. RELATED WORKS

As early as 2002, Moretti plotted the character relationship networks of “Hamlet”, “Our Mutual Friend” and “The Story

[§] Corresponding author: yanglijiao@bnu.edu.cn

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

of the Stone" and performed a deep analysis of the plot structures (2005).Then, Sparavigna applied the method of Moretti's to plot the character relationship networks of "Harry Potter". David K. Elson and others focused on 19th-century English novels and serials, extracting networks from dialogue relationships, and drew preliminary conclusions on the simultaneous growth of social network cohesiveness, interconnectivity. Rydberg-Cox created an application to visualize and explore social networks of Greek tragedies(2011). Sparavigna and Marazzato applied the Graph Visualization Software to visual the character networks of two Shakespeare's play (2014).

With the introduction of this method, domestic scholars have gradually started to carry out relevant experiments and research. Some traditional literary researchers draw on the research methods and experiences of foreign scholars to conduct similar human relationship extraction and visual representation for a few Chinese novels, and to explore the relationship between the characters, the centrality of the characters, and so on. For example, Tang Yi studied the 108 main characters in The Water Margin. Zhu Haijun and others built the character relationship network in the TV series "Bailu Village". Liu Haiyan conducted a comparative study of Fitzgerald's four novels. The topological measurement of the four networks confirmed that they have the characteristics of "small world". Chen Bikun and others used the "Journey to the West Prequel" as a corpus to calculate and visualize the character network. Chen Lei, Hu Yiqi and others explored the social hierarchical relationship between the pairs of people who frequently co-occurred in " The Story of the Stone". In addition, some information technology researchers focused on the implementation path of automatic extraction of character relationships. The computational models used mainly include word vectors and two-way GRU nerves Network, convolutional neural network, etc.

In general, the realization of the above research and the conclusion of the preliminary conclusions have prove the similarity between the virtual world of the text and the real world to a certain degree. At the same time, it also verifies that using complex network measures to statistically analyze relationship between the characters is a completely feasible and reliable research method. But we can also see that there are obvious deficiencies in the existing research. First, the research on existing literary texts mostly focused on the network construction and text content mining of individual works, individual series of texts or individual writers. The number of texts is very limited, lacking horizontal comparison of different texts makes it difficult to get more general conclusions. Second, the use of topological measure indicators for the relationship network is more based on the qualitative research and does not have more effective application and mining of the network measurement. In addition, the relevant research in China is mostly empirical research, trying to provide a quantitative basis for the literary criticism of the novel, and thus lacks specific scenarios and fields for future applications.

III. Experimental Process

A. Data And Pre-Processing

This article selects the narrative texts recommended by the Ministry of Education in 2018 and uses the children's literature works recommended by Beijing's outstanding teachers as supplements. A total of 100 narrative children's literature texts are collected and cleaned, such as "Little Pig Lili Lulu ", "Alice in Wonderland", "Peter Pan", "Xiao Bing Zhang Ga", "Niels riding a goose travel", etc.,adding up to more than 160,000 words. Since the purpose of this paper is to evaluate the complexity of the relationship between the characters, the texts selected should have obvious differences in the complexity of the relationship of the characters. There should also be more obvious differences in indicators. The selected texts are graded by three linguistic graduate students and two associate professors based on the complexity of the relationship between the characters, with a value of 1-15. The average score of the five scores obtained from the text is used as the difficulty marker. In this paper, each literary text is divided into words by jieba tokenizer, and the total number of words in each paper is counted.

B. Extracting Conversational Networks from Literature

1) Character Identification

Character recognition and marking tasks are one of the research hotspots and emerging development directions. The name is part of the named entity. Most of the current mainstream methods are based on machine learning models. This method learns the annotation corpus and implements unregistered words and person name recognition in the form of sequence annotation. This paper uses conditional random field CRF (Conditional Random Fields) to solve the problem. By manually defining the feature template, the contextual annotation information is used to explore the named entity. The article uses the open source tool CRF++, which combines the 50,000-word "People's Daily" corpus with 50 children's literary texts collected in this paper to complete the word segmentation, part-of-speech tagging and name tagging, forming a training corpus. The non-named entity N, and the required character word entity was labeled Per, as shown:

TABLE I. The Labels of One Sentence

孔子	称赞	子产	是	对	人们	有着	惠爱	的	人	。
Nr	v	N	v	p	n	v	n	uj	N	x
Per	N	Per	N	N	N	N	N	N	N	N

Using this as a training set, the model is trained using the characteristics of the word itself, the part of speech, the position of the word. Then the name recognition was performed for the remaining half of the text to extract all the names of the remaining literary texts and save them as a character dictionary.

In addition, the name of the person should be reviewed. First, the characters with the word frequency below 2 should be excluded to avoid excessive non-important characters from interfering with the network construction. Based on these work, this article carried out the list of names. In a continuous narrative text, the person entities and the referents do not necessarily correspond one-to-one. This requires us to identify different references of the same entity and to make them consistent. Therefore, this article asks five Chinese students to participate in the proofreading work on the same name of the text person, and assign different referential methods to the same person. For example, “Confucius” is replaced by “Confucius” and “Zhongni” is replaced by “Confucius”. To ensure that the name of the person is not redundant or repeated, this is a prerequisite for ensuring accurate identification of the relationship between the characters.

2) Constructing social networks

In the predecessors' research, the character relationship network of literary texts is mostly based on the co-occurrence of characters and the dialogue of characters. This paper combines two methods to place independent paragraphs of consecutive dialogues in the same paragraph, and then adopt the method of judging the co-occurrence of characters. That is, the paragraph newline is used as the criterion, and the characters appearing in the same paragraph are recorded as co-occurrence.

The co-occurrence relationship of two characters is regarded as one side of the relationship network.We extracted the relationship of the characters based on the name dictionary ,constructed the adjacency matrix representing the co-occurrence relationship between the characters, and transformed it into a network model for visualization by the self-editing program. The construction tool used in this paper is networkx. The existing person name dictionary is used as point data, and the character relationship information is passed as side data to construct and output the network model.

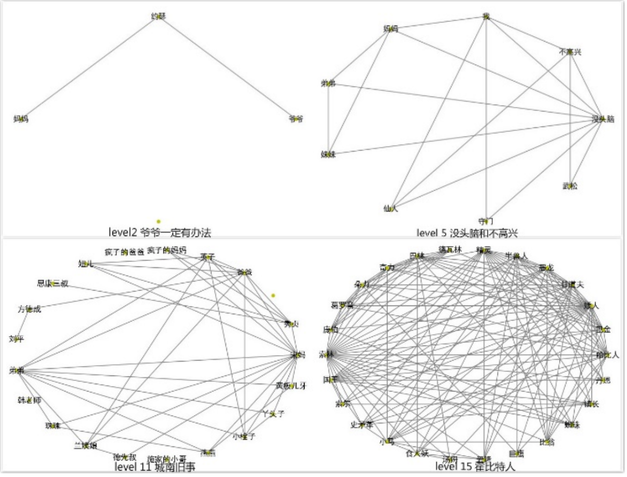

Figure 1. Examples of the networks in different levels.

In this paper, data analysis and simulation modeling of complex networks are performed on the generated 100 networks, using its built-in graph and complex network analysis algorithm. Then, the indicators are extracted and statistically analyzed.

C. Topological Measurement Calculation of Virtual Social Network

After obtaining the character relationship network in the novel text, the complex network measure can be used to analyze its structural characteristics. Here, the paper firstly defines and explains the network measurement and the calculation meaning of the corresponding person network.

1) Network: The network in this paper is actually a set of points and edges, represented by the graph G=(V, E). V is the set of network nodes; E is the relationship between the nodes on the network, which called the edge set.

2) Points and Edges: The points in this article are the abstractions of the characters, and the labels are the specific names of the characters. The side is the co-occurrence relationship of a pair of characters such as "Confucius-Mencius". In addition, the number of points of the relationship network is abbreviated as N, and the number of sides is abbreviated as M.

3)Average Path Length: In a network, the distance $\mathbf{d}_{ij}$ between two nodes, labeled i and j respectively, is defined as the number of edges along the shortest path connecting them. The average path length L of the network, then, is defined as the mean distance between two nodes, averaged over all pairs of nodes.

4) Diameter: The diameter D of a network, therefore, is defined to be the maximal distance among all distances between any pair of nodes in the network.

5) Network Density: Referring to the closeness of the connection between nodes in a network, the mathematical meaning is the ratio of the number of edges actually existing in the network to the upper limit of the number of edges that can be accommodated. And it is denoted by ‘$\mathbf{d}$’.

6) Average Degree k: The degree k_i of a node is usually defined to be the total number of its connections. Thus, the larger the degree, the “more important” the node is in a network. The average of k over all points is called the average degree of the network, and is denoted by‘$< \mathbf{k} >$’.

7) Character Density: Due to the particularity of literary text, this paper also examines the character density, denoted as P:

$$P = \frac{N}{Total\ number\ of\ words} \qquad (1)$$

In the process of selecting calculation indicators, this paper focuses on the two dimensions of the network as a whole and the internal structure of the network. On the one hand, the number of characters and relationships, the network diameter and the density of text characters are the overall portrayal of the character network of the text. On the other hand, the character network path length, the character network density, and the character network average node degree are more capable of portraying the association of the characters, and the narrative structure of the text.

TABLE II. CHARACTER NETWORK CHARACTERISTIC INDEX

Name	***Number of points***	***Number of edges***	***average path length***	***diameter***
Logogram	N	M	L	D
Name	***Network Density***	***average degree k***	***character density***	
Logogram	D	<k>	P	

Based on the above concepts and calculation methods, this paper uses self-programming program to call the networkx built-in standard graph theory algorithm to empirically analyze the structural characteristics of the character relationship network in 100 children's literary novels.

First, the data results are compared with the adult texts and the topological features of complex networks in the real world.

TABLE III. INDEXS OF THE THREE TEXTS

Book name	***N***	***M***	***L***	***D***	***d***	***<k>***	***P***	***RANK***
Grandpa have a way	3	2	0.67	2	0.67	1.32	1%	2
South of the city	17	88	1.84	4	0.65	10.44	0.29%	13
Hobbit	26	285	1.51	3	0.88	21.91	0.08%	15

IV. RESULTS AND ANALYSIS

A. The Characteristics and Analysis of the Relationship network

In network theory, "Small world theory "is a special kind of complex network structure. Most of the nodes in this network are not connected to each other, but most of the nodes pass a few cloths. The famous "Six degrees of separation" is one of the manifestations. Moreover, the diameter D and the characteristic path length L are both small, and the feature path length is proportional to the logarithm of the network scale.

In this paper, the L and D indicators of the sample are taken out and verified.More than 95% of the text L is between 1.12 and 2.00, and D is located at 2-3. Both of them are much smaller than the size of the relationship network. In addition, this paper linearly fits the feature path L with the logarithm of the number of nodes. The fitted regression equation is y=1.6093x+0.0848 and R2 is 0.6997.

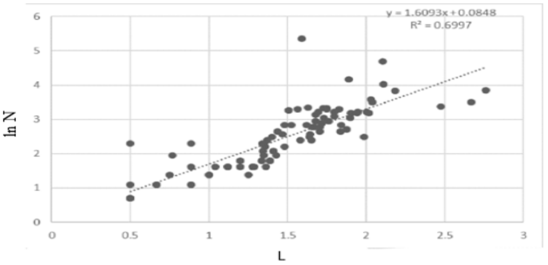

Figure 2. Linear relationship between L and ln(N).

The fitting effect is good, so it is confirmed from two aspects. The virtual social network of characters in narrative children's literature is in line with the characteristics of the small world network.

The network diameter and average distance of children's literary figures are significantly small. Compared to the regular network with the longest average distance at the same scale, the network average distance is shortened by adding a new edge to the original rule network with a small probability. That is to increase the uncertainty and randomness of the network. The connection between any two characters is uncertain before the text is formed. It is gradually developed and realized in the process of literary narration. This is not only the embodiment of language, but also the characteristics of literature. In contrast, unlike random networks, the construction of literary texts is not completely random. It has the constraints of narrative, cognitive psychology ,literature, and the writer's "deliberate action". The connection is not completely based on a certain probability, but a self-organizing network system.

The small average distance reflects the good "connectivity" of the network. One or several main characters are used as clues to connect the whole narrative process in series, so that other characters can be connected through a few parts. It reflects the common linear narrative structure of narrative literature. The frequent appearance of the protagonist is easy for children to remember and understand, and the narrative coherence is strong, which reduces the cognitive load of the reader to a certain extent, and thus is more susceptible to the preference of children's literature writers. With the length of the text, the number of characters, the number of people's relationships, the complexity of the relationship between the characters, and the network diameter of the text, it also shows an increase, from 2 to 6, gradually approaching some adult texts, such as "Beautiful Friends." It is also gradually approaching the network diameter of the real social world. This shows to some extent the development of children's cognitive ability.

In addition, this paper finds that within the same text, the connection ability and importance of different characters in the network are obviously different. Therefore, another measurement concept of complex network are defined and selected in this paper - node degree **d(v)**. The node degree of each node in the regular network is the same. The node degrees of most nodes in the random network are the same too, but the node degrees of the other nodes are very evenly distributed. However, the degree of node distribution in small world networks is not balanced. In this paper, a story of Confucius is used to calculate the node traversal of each character, and the number of nodes with the same node degree is counted as frequency. Fig. 3 shows the relationship between the node degree **(d(v))** and the frequency of nodes.

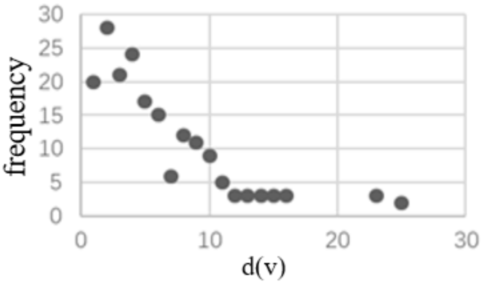

Figure 3. the relation between d(v) and frequency in story of Confucius

Fig. 3 shows the phenomenon that a small number of nodes occupy most of the edges of the entire network, and the number of connected edges of most nodes is small.Due to the narrative needs of literary works, the text focuses on the description of the protagonist. Many of the supporting roles are played around the protagonist, and more marginal characters are distributed at the outermost part of the network of character relationships. Zipf's law states that in natural language, only a very small number of words are used frequently, and most words are rarely used. From this point of view, there is a similar relationship between the character degree of the character and the number of characters of the same node degree in the narrative text. Therefore, the small world characteristics of the text reflect to some extent the detailed differences in the character characterization of children's literary works. This feature is also applicable in other narrative literary works. A large number of experimental studies have shown that real networks in the world have almost a small world effect, and scientists have also found that a large number of real network nodes obey the power rate distribution. After the mining and analysis of the relationship network, we should see that although the content of children's narrative literature is quite different from the real world, there are more virtual components, but the deep structure of the relationship between characters and objects is very similar to the real world. It shows that the beauty of literature is to reflect reality .

B. Method and indicator validity test

This paper plans to use the topological measure indicators of the extracted text person network to evaluate and predict the complexity index of the characters in the children's literature. In order to determine the validity of this quantitative method and the accuracy of the final prediction, this paper examines the differences in the ability of different topological measures to reflect the complexity of the relationship in the children's literature and the impact of each indicator on the prediction of the complexity of the relationship. This paper evaluates the degree of correlation between individual network eigenvalues and tag values, sorting from high to low, leaving a high degree of correlation. In terms of specific methods, this paper draws the correlation heat map of each index and measures the correlation between the characteristic index and the complexity of the relationship based on the Pearson correlation coefficient:

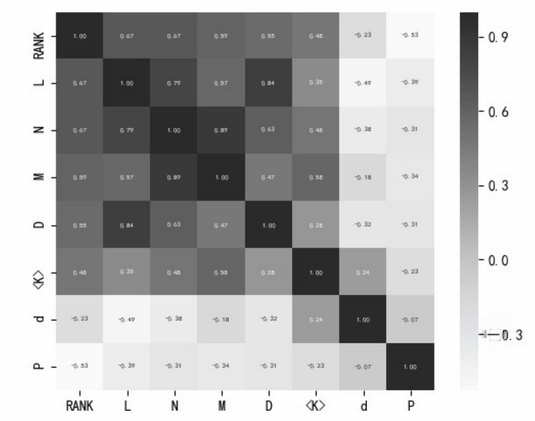

Figure 4. Heatmap of the eight indicators

It can be seen from the Fig.4 that among the selected six characteristic indicators, the absolute value of the correlation degree with the tagged classification information RANK is: the character network path length L(0.67), the number of characters N(0.67), the number of relationship pairs M(0.59), network diameter D (0.55), average degree <K> (0.53), text person density P (0.45), character network density **d** (0.2). It can be seen that there is a strong correlation between the L, N and text person relationship complexity, and the characteristics M, D, <K>, P and the text person relationship complexity are moderately related, and d is weakly related to the text person relationship complexity.

Therefore, this paper confirms the feasibility of using the topological measure index of the extracted text person network to predict the character relationship complexity index in children's literature. After the correlation test, this paper excludes the character network density **d** index, and uses the residual measure parameter as the feature to predict the complexity.

C. Regression model prediction and evaluation

Because the corpus size is not large enough, in order to ensure that the sample is used as much as possible, this paper uses the 5-fold cross-validation method to record the mean square error as the evaluation index, and selects the MAE and R2 results from the regression models of 3 commonly used machine learning.

TABLE IV. EFFECT OF REGRESSION MODEL EVALUATION

Model	*MAE(Mean Absolute Error)*	*R^2*
Linear Regression	1.487	0.595
XGBoost	1.095	0.524
Lasso	1.743	0.60

The average absolute error of the three models on the test set is about 1. Due to the ambiguity and subjectivity of the complexity of the relationship between the characters, this paper calculated the **average difference** (AD) of the scores given by the five scorers, it is 1.34. The experimental result error is within an acceptable range, thus confirming the validity of the experimental results.

The resulting regression equation is:

$$Y(\text{grade of difficulty})=0.071N+3.849L+0.018\langle k\rangle+0.156D-0.13P+0.0159M+1.267. \tag{2}$$

In order to further confirm the applicability of the method, we randomly selected another 20 children's literature texts for non-experimental data to calculate the complexity of the relationship between the characters, and used the trained regression model to predict .The input is the characters list for each text ,the network average path length L, the number of people N, the number of relationship pairs M, the network diameter D, the network average degree <K>, the text person density P six index values, and the output is the text relationship complexity prediction of each text value. The test results are shown below:

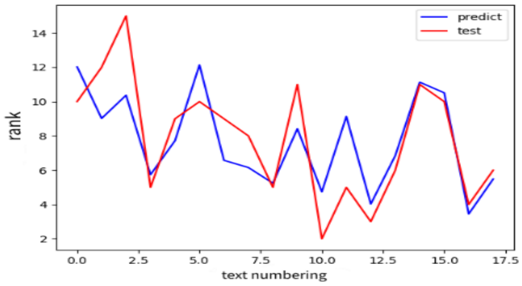

Figure 5. Result of prediction

The Fig.5 show that the average error remains at 1.42, which is within acceptable limits. But it also reflects the greater uncertainty of narrative in children's literary texts . The results of the small sample size experiment are more to prove the validity of this evaluation method. In order to improve the predictive validity, the quantity and quality of the experimental data still need higher requirements.

V. Conclusion and Future Work

Based on the methods of predecessors, this paper applied the CRF++ training model to 100 children's literature works. We automatically extracted the characters and the social network between the characters based on the co-occurrence relationship to characterize and analyze the character relationship of the literary novels. The result is consistent with the content of the novel.

In this paper, we presented a method for characterizing a text of literary fiction by extracting the network of social conversations that occur between its characters. This allowed us to take a systematic and wide look at a large corpus of texts, an approach which complements the narrower and deeper analysis performed by literary scholars and can provide evidence for or against some of their claims. After that, this study successfully applied the network attribute of some measure indicators of complex networks to the automatic analysis, mining and comparison of the character network and plot complexity of different literature texts.

On the basis of this work, we can expand the text readability measurement dimension of children's literature text from character, word and sentence to the whole book. Therefore, the score of this index will be used in the evaluation of text readability together with other indexes such as character frequency, word frequency, semantic transparency, sentence complexity, sentence length and so on.

In order to make the evaluation and prediction effect better and enhance the fitting effect of the model, this paper plans to collect more experimental data, and further consider different text characters in the data collection. The complexity of the text is as balanced as possible, facilitating the construction of subsequent machine learning models. In the construction of the personal relationship network, it is expected to try to improve the recognition accuracy based on the more characteristic BI-LSTM-CRF name recognition method. In addition, the paper finds that the diversity, ambiguity and uncertainty of the reference affect recognition of character to identify relationship. For example, "君主(prince)" as a collective noun, representing a number of prince entities, such as "Jin Wen Gong", "Chu Zhuang gong", etc. Besides, in different contexts, specific reference objects, computer automatic identification and generational digestion requires further exploration and research. In addition, in the difficulty of children's literature texts, the upper limit of the difficulty of our defense is difficult to control. This paper calculated the weight of the character and the weight of the edge in the calculation of the relationship between the characters, but it is not used in the network construction for the time being. In the later research, we need to solve how to add the weight of the edge and the character to the construction and evaluation of the network.

References

[1] Moretti, F., Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso Press. 2005.

[2] Alberich.R ,Miro-Julia.J,Rosselo F.Marvel Universe Looks almost Like a Real Social Network,.New York :Cornell University Library,2002.

[3] He, H., Barbosa, D. & Kondrak, G., Identification of Speakers in Novels. Meeting of the Association for Computational Linguistics (pp. 1312-1320). Sofia., 2013.

[4] Rydberg- Cox J. Social networks and the language of greek tragedy. Journal of the Chicago Colloquium on Digital Humanities and Computer Science, 2011,1 (3).

[5] Sparavigna A.C., Marazzato R.. Graph visualization software for networks of characters in plays. International Journal of Sciences, 2014,3 (2): 69– 79.

[6] Watts D J , Strogatz S H . Collective Dynamics of Small World Networks[J]. Nature, 1998, 393(6684):440-442.

[7] Y.Tang, S.Wang ,H. Hu, "Small World" on BailuVillage: An Empirical Analysis of the Network of Character Relations in the TV BailuVillage , China Radio & TV Academic Journal,2018,pp117-12

[8] H.Y. Liu , X.H. Yang . "Quantifying the Vicissitude of Fitzgerald's Creativity a Statistical Analysis Based on Lexical Measures." Information Technology & Artificial Intelligence Conference 0.

[9] Chen, Bikun , and Y. Wang . "Character interaction network analysis of chinese literary work- A preliminary study." Proceedings of the Association for Information Science & Technology 53.1(2016):1-4.

[10] H.P. Zhu, Z.Y.Luan, Text Mining of Character Relation Network in Water Margin, Social Sciences Review,

[11] Stiller J., Hudson M. ,Weak links and scene cliques within the small world of Shakespeare. Journal of Evolutionary Psychology, 3(1), 2005 pp.57-73