Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5553635, 8 pages
https://doi.org/10.1155/2021/5553635

Research Article

Coword and Cluster Analysis for the Romance of the Three Kingdoms

Chao Fan [1,2] **and Yu Li** [1,2]

[1]*The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China*
[2]*Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, China*

Correspondence should be addressed to Chao Fan; fanchao@jiangnan.edu.cn

Received 1 March 2021; Revised 12 March 2021; Accepted 19 March 2021; Published 1 April 2021

Academic Editor: Shan Zhong

The *Romance of the Three Kingdoms* (RTK) is a classical Chinese historical novel by Luo Guanzhong. This paper establishes a research framework of analyzing the novel by utilizing coword and cluster analysis technology. At the beginning, we segment the full text of the novel, extracting the names of historical figures in the RTK novel. Based on the coword analysis, a social network of historical figures is constructed. We calculate several network features and enforce the cluster analysis. In addition, a modified clustering method using edge betweenness is proposed to improve the effect of clustering. Finally, both quantified and visualized results are displayed to confirm our approach.

1. Introduction

The *Romance of the Three Kingdoms*, written by Luo Guanzhong, is generally considered to be one of the four great classical novels in Chinese literature. It describes the turbulent years from the end of the Han dynasty to the Three Kingdoms (Wei, Shu, and Wu) era in Chinese history. More than 1000 personalities are vividly portrayed in the historical novel.

In this research, text of original novel is divided into a number of sentences. According to coword analysis, there is a certain intrinsic relationship between the two words when they appear in the same document. Thus, we calculated the frequency of cooccurrences for two names in a sentence. The character name is reckoned as the node and the cooccurrence as the link, so that an undirected network can be established. Furthermore, various network features are computed to analyze relationships of characters in the novel. Cluster analysis is employed to explore the hierarchical structure of RTK. Finally, an improved clustering algorithm by cutting high-betweenness edges is proposed, which performs better than the common approach in clustering effect.

This manuscript is organized as follows. Section 2 gives related work of this paper. Data preparation is discussed in Section 3. Sections 4 and 5 express the network feature analysis, cluster analysis, experiments, and the analysis of results. Conclusions are drawn in Section 6.

2. Related Work

Early research about the RTK concentrates on qualitative analysis, such as the writing style, genealogy, and characters. Later, a quantitative approach was adopted to analyze the novel. Coword analysis is such a method of importance, which was first devised by French scholars and introduced into the information science field by Callon [1]. According to the theory of coword analysis, there is a close connection between two words when they appear in a sentence. More cooccurrences of the two words indicate the closer relationship between them. In this paper, we consider the cooccurrence of character names in a sentence of the RTK novel.

Numerous researches on literature analysis have been done based on the technologies of coword analysis. Ravikumar et al. [2] inspect 959 articles in scientometrics based on the coword analysis approach and find that the topics in publication are changing to new themes. As for the medical literature, there is a study utilizing this tool to process them over a span of thirty years [3]. Another work focuses on past themes and future trends in medical

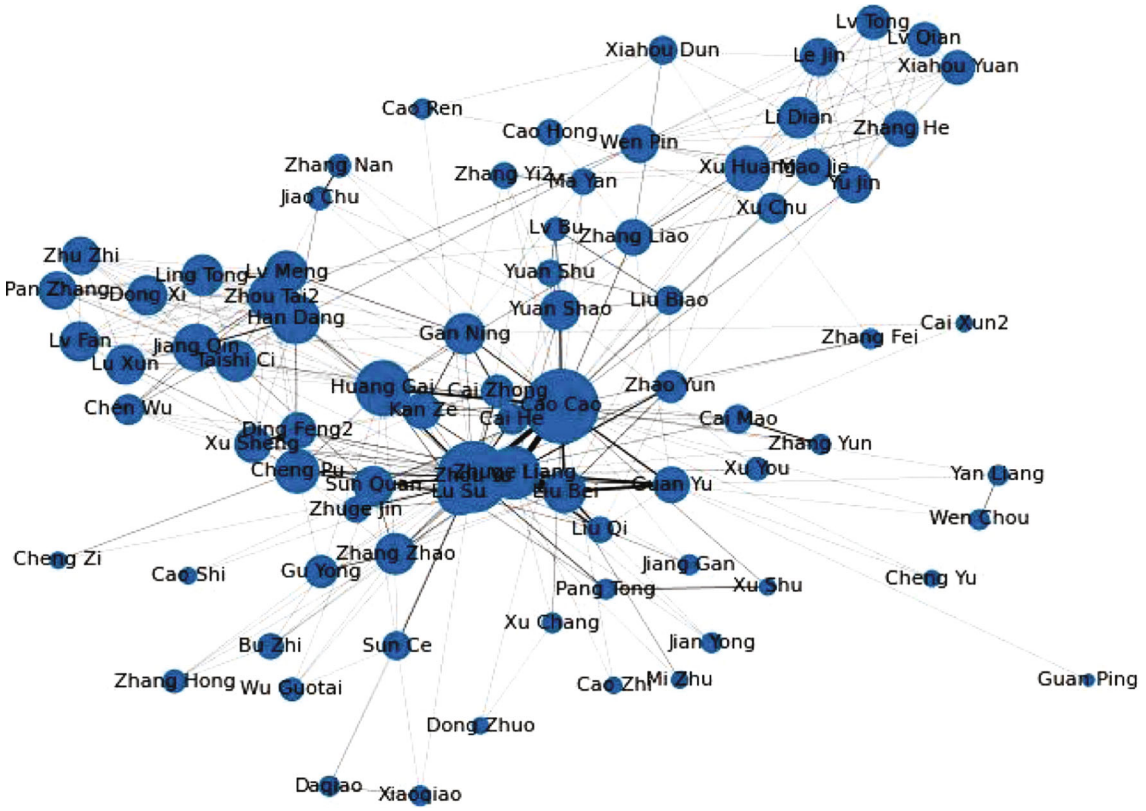

Figure 1: A network of character names (top 80 in node frequency).

tourism research [4]. Employing the coword analysis, some researchers attempt to identify the themes and trends of main knowledge areas including engineering, health, public administration, and management [5]. Moreover, a coword network is established to analyze the relationship of characters in the *Dream of the Red Chamber* [6]. Wang et al. build a similar network for the *Romance of the Three Kingdoms* [7].

After creating a social network based on coword analysis, the cluster analysis is carried out by performing a hierarchical clustering algorithm. Two types of algorithm are often implemented when moving up the hierarchy. The divisive approach of clustering reckons all data as one cluster and performs splits, which is used in many research [8]. Nevertheless, the agglomerative hierarchical clustering is a bottom-up method with many variants [9]. It merges the two most similar clusters at each time. The agglomerative method is exploited in this work because it can provide a visual expression of the clustering results.

3. Data Preparation

3.1. Building RTK Corpus and Preprocessing. As many data of the novel can be downloaded from the Internet, we selected a high-quality text document (https://72k.us/file/22215238-408791478) in Chinese character, establishing the RTK corpus by cleaning the original data. Some words with errors were modified, and the wrong punctuations were removed manually.

The raw text is preprocessed using the natural language processing toolkit ICTCLAS (http://ictclas.nlpir.org/). We acquired a name list of RTK characters through the Internet and added it to the dictionary of ICTCLAS. Then, the lexical analysis is executed to segment Chinese sentences into words where names of characters can be found.

3.2. Creation of Character Name Network. Based on coword analysis, an undirected network of character names can be created by counting the cooccurrences of two names in sentences. We treated full name, its courtesy name, and abbreviated name as one name. For example, "Cao Cao" is equal to "Cao Mengde" and "Mengde," which means the three names refer to a single person of "Cao Cao."

The final constructed network of character names has 1,133 nodes and 5,844 links. As depicted in Figure 1, the size of a node indicates the count of the character name in the novel and the thickness of a link corresponds to the frequency of two characters that appear together.

4. Network Feature Analysis

4.1. Degree Distribution. As the degree of a node is the number of links adjacent to it, the degree distribution is the probability distribution of these degrees. A power index γ can be used to describe the curve if the network's degree distribution follows a power-law distribution.

For the network of RTK characters, the top ten characters of the highest degree are Cao Cao, Liu Bei, Zhuge Liang, Sun Quan, Zhao Yun, Guan Yu, Yuan Shao, Sima Yi, Lv Bu, and Wei Yan. The average degree of the network is 10.31, and the degree distribution can be illustrated in Figure 2. It emerges

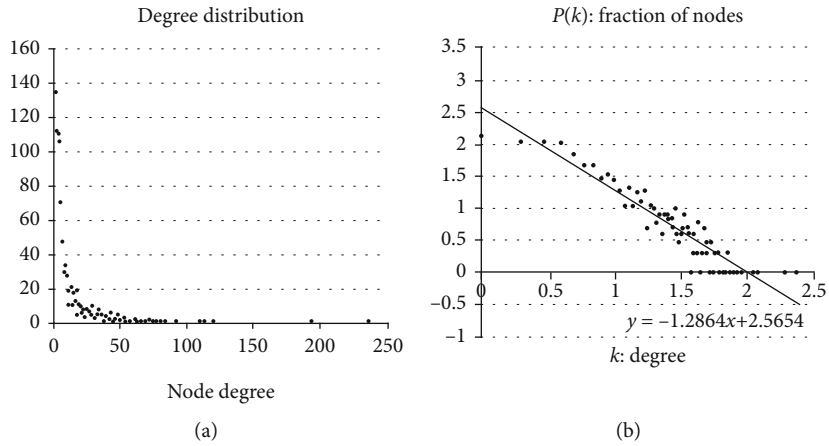

Figure 2: Degree distribution and power-law degree distribution on a log-log scale.

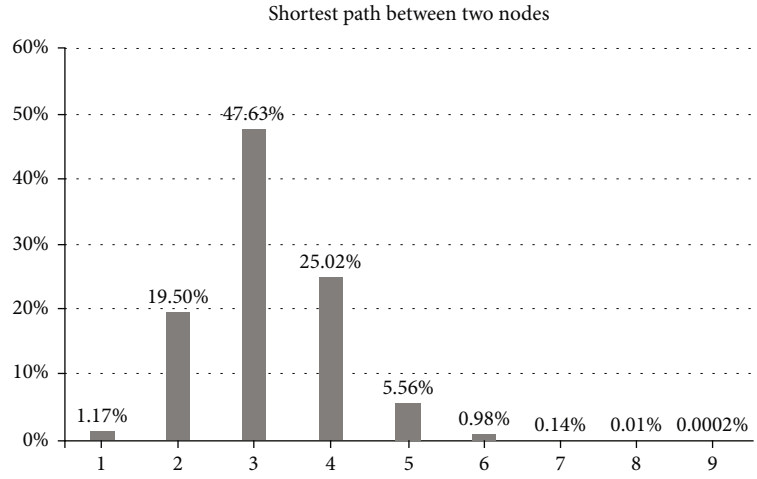

Figure 3: Distribution of shortest-path length.

to be a heavy-tailed distribution (see Figure 2(a)). As the data can be approximated with a linear function $y = -1.2864x + 2.5654$ on a log-log scale in Figure 2(b), we conclude that the degree distribution follows a power-law distribution.

4.2. Average Shortest-Path Length. The shortest path between two nodes is a path where the number of links is minimized. Accordingly, the length of the shortest path is the number of links that the path contains. A sum of all shortest-path length divided by the number of links is the average shortest-path length.

The average shortest-path length of the RTK network is 3.1743. Hence, one character can be connected to others in three steps on average, which means any two characters are "three-degree separation."

The distance of the largest shortest path in the network is called diameter. In this paper, the RTK network's diameter is 9. One path of the diameter is from Liu Ai to Zhang Shang: Liu Ai, Wang Li, Dong Zhao, Cao Hong, Cao Cao, Sima Yan, Yang Hu, Du Yu, Lu Jing, and Zhang Shang. The distribution of the shortest-path length between any two characters can be illuminated in Figure 3. According to the figure, 47.63% of the shortest-path length in the RTK network is 3 and about 92.15% is between length 2 and length 4.

4.3. Clustering Coefficient. A clustering coefficient [10, 11] measures the extent to which a network's nodes tend to cluster together. The clustering coefficient of node x can be given by

$$C_x = \frac{2E_x}{k_x(k_x - 1)}. \quad (1)$$

E_x is the existing links among neighbors of node x. As k_x is a degree of node x, $(1/2)k_x(k_x - 1)$ represents the number of potential links for node x's neighbors. Therefore, the average value for all C_x is the clustering coefficient of the whole network.

$$C = \frac{1}{N}\sum_x C_x. \quad (2)$$

A random network is produced by an Erdős-Rényi (ER) model utilizing the same number of nodes and links as the RTK network. The comparison between random network

Table 1: Comparison between RTK and random network.

	Number of nodes	Number of links	Average degree	Average shortest-path length	Clustering coefficient
RTK network	1,133	5,844	10.3159	3.1743	0.5306
Random network	1,133	5,844	10.3159	3.2702	0.0082

Table 2: Comparison of three subnetworks and the whole network.

	Density	Clustering coefficient	Average shortest-path length	Diameter
Shu	0.1652	0.6635	2.0563	4
Wu	0.1099	0.5845	2.3054	5
Wei	0.0803	0.6217	2.5953	6
The whole network	0.0091	0.5306	3.1743	9

Table 3: Top 10 characters in rank with the highest centrality.

Ranking	Degree centrality	Betweenness centrality	Closeness centrality
1	Cao Cao (0.2094)	Cao Cao (0.1751)	Cao Cao (0.4528)
2	Liu Bei (0.2085)	Liu Bei (0.1304)	Liu Bei (0.4442)
3	Zhuge Liang (0.1714)	Zhuge Liang (0.1093)	Zhuge Liang (0.4313)
4	Sun Quan (0.1060)	Sun Quan (0.0695)	Sun Quan (0.4073)
5	Zhao Yun (0.0998)	Sima Yi (0.0430)	Guan Yu (0.3969)
6	Guan Yu (0.0972)	Zhao Yun (0.0413)	Zhao Yun (0.3963)
7	Yuan Shao (0.0813)	Liu Shan (0.0402)	Sima Yi (0.3924)
8	Sima Yi (0.0742)	Guan Yu (0.0375)	Wei Yan (0.3856)
9	Lv Bu (0.0716)	Yuan Shao (0.0369)	Yuan Shao (0.3842)
10	Wei Yan (0.0707)	Jiang Wei (0.0357)	Cao Ren (0.3824)

and RTK network is shown in Table 1. The RTK network is a small-world network because it has a larger clustering coefficient as well as a smaller average shortest-path length compared with a random network.

We choose the characters who clearly belong to the three groups of Wei, Shu, and Wu and calculate the network features of the three kingdoms, respectively. The results are summarized in Table 2.

The character relationship networks within three groups have high clustering coefficients and small average shortest-path lengths. Consequently, all of the three subnetworks are "small-world" networks. From the Shu to Wu and Wei, the density and clustering coefficient of the subnetworks decrease sequentially except for the clustering coefficient of Wu. On the contrary, the average shortest-path length and diameter increase successively. This reflects a decrease in the closeness of the connections among the groups. In other words, the connections among characters in Wei are less closely than Wu and Shu.

4.4. Density. The density of a network shows the ratio of links, which can be simply calculated by formula (3). N and E are the number of nodes and links. It describes the portion of all possible links in a network that are actual connections. The value is a fraction between 0 and 1. As the density of the RTK network is 0.0091, it is a sparse network.

$$d = \frac{2E}{N(N-1)}. \tag{3}$$

4.5. Centrality. The centrality measures the importance of nodes, containing degree centrality, betweenness centrality, and closeness centrality.

Degree centrality is a measure of centrality based on degree. A high-degree node is a local center within the network. Betweenness centrality expresses the extent that the node falls on the shortest path between other pairs of nodes. A node with a high betweenness is capable of controlling the interactions between two nonadjacent nodes [5]. Closeness centrality is a measure of the average shortest distance from each node to each other node. It evaluates the closeness that a node is to all the other nodes [3].

Three centralities of characters in the RTK network are calculated, respectively. Table 3 gives the top ten characters of the highest centrality. The value of centrality is listed in parentheses. From Table 3, we can find eight names listed in three centralities: Cao Cao, Liu Bei, Zhuge Liang, Sun

Table 4: Cooccurrence matrix of main characters.

Cooccurrence	Liu Bei	Cao Cao	Sun Quan	Zhuge Liang	Guan Yu	Zhang Fei
Liu Bei	541	112	58	190	106	75
Cao Cao	112	275	39	50	58	16
Sun Quan	58	39	145	28	18	2
Zhuge Liang	190	50	28	336	43	25
Guan Yu	106	58	18	43	272	47
Zhang Fei	75	16	2	25	47	165

Table 5: Ochiai similarity matrix of main characters.

Cooccurrence	Liu Bei	Cao Cao	Sun Quan	Zhuge Liang	Guan Yu	Zhang Fei
Liu Bei	1	0.290371	0.207083	0.445641	0.276327	0.251027
Cao Cao	0.290371	1	0.195305	0.164488	0.212069	0.075112
Sun Quan	0.207083	0.195305	1	0.126854	0.090637	0.01293
Zhuge Liang	0.445641	0.164488	0.126854	1	0.142238	0.106176
Guan Yu	0.276327	0.212069	0.090637	0.142238	1	0.221856
Zhang Fei	0.251027	0.075112	0.01293	0.106176	0.221856	1

Table 6: The clustering result of the RTK network (k is the final number of hierarchical clusters).

k	Precision	Recall	F score
…	…	…	…
11	43.83%	78.90%	56.35%
12	47.08%	78.90%	58.97%
13	71.10%	75.00%	73.00%
14	71.10%	75.00%	73.00%
15	**87.66%**	**73.38%**	**79.89%**
16	87.66%	62.99%	73.30%
17	87.66%	59.09%	70.60%
18	87.66%	50.97%	64.46%
…	…	…	…

Quan, Zhao Yun, Guan Yu, Yuan Shao, and Sima Yi. They are in a significant position in the character network.

5. Cluster Analysis

5.1. Cooccurrence and Similarity Matrix. The cooccurrence matrix measures the frequency that two characters appear together. A cooccurrence matrix of main characters in the RTK network is presented in Table 4. It is a symmetric matrix, and data on the diagonal show the frequencies of characters that appear in text.

The cooccurrence of two characters cannot be used as the similarity because it is greatly affected by frequency. We normalize the cooccurrence matrix utilizing the Ochiai coefficient [12] and obtain the similarity matrix. Ochiai coefficient is defined by

$$K = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}. \quad (4)$$

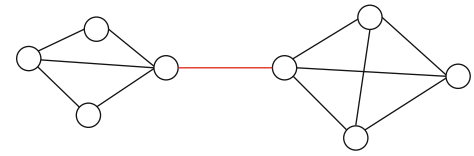

Figure 4: A link with a high edge betweenness.

Table 7: The clustering result of the RTK network (k is the final number of hierarchical clusters).

Number of removals	Precision	Recall	F score
0	87.66%	73.38%	79.89%
5	87.66%	73.38%	79.89%
10	88.31%	73.38%	80.15%
15	88.64%	74.35%	80.87%
20	88.64%	74.35%	80.87%
25	88.64%	74.35%	80.87%
30	88.64%	74.35%	80.87%
35	88.64%	74.35%	80.87%
40	88.96%	73.70%	80.62%
45	88.96%	73.38%	80.42%
50	89.94%	73.05%	80.62%
55	89.94%	72.08%	80.02%
60	47.73%	90.58%	62.52%
…	…	…	…

As A and B are sets, $n(A)$ is the number of elements in A and $n(A \cap B)$ is the number of cooccurrence. The similarity matrix calculated by the Ochiai coefficient is described in Table 5.

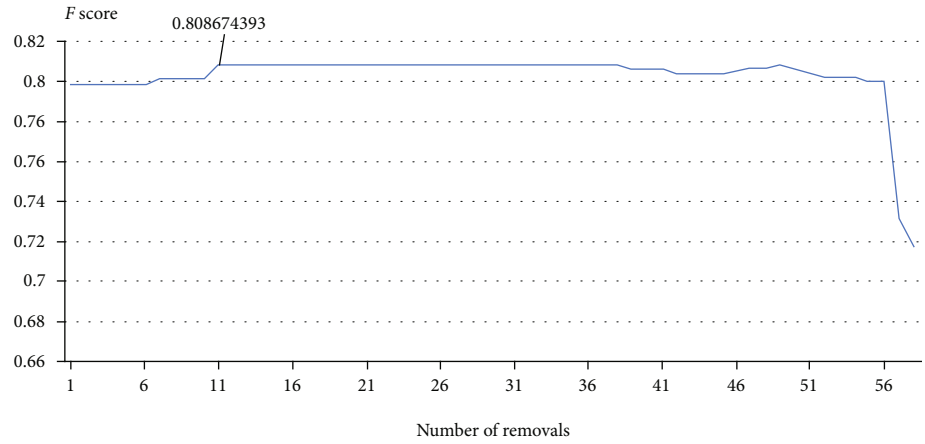

Figure 5: The change of F score according to the number of removals.

5.2. Hierarchical Clustering

5.2.1. Clustering Algorithm. An agglomerative hierarchical clustering algorithm utilizing the Ochiai similarity matrix is implemented to complete the task of cluster analysis. It is a bottom-up approach. Initially, each node is treated as a single cluster. Two clusters with the largest Ochiai similarity are combined into a new bigger cluster. The clustering algorithm stops when it achieves a setting threshold or there is only one cluster left. The similarity between two clusters is defined as the average similarity between each of their nodes.

5.2.2. Evaluation. The P-IP scores [13] are adopted to measure the clustering result. There are m character names and n clusters. Suppose C_{ij} is the number of character names marked with label j for character name i, where $j = \arg\max_{k} \max_{k=1,2,\cdots,n} \{C_{ik}\}$. The precision and recall of character name i can be given by

$$P_i = \frac{C_{ij}}{\sum_{l=1}^{m} C_{lj}}, \qquad R_i = \frac{C_{ij}}{\sum_{k=1}^{n} C_{ik}}. \tag{5}$$

Thus, the F score is calculated by

$$F_i = \frac{2P_iR_i}{P_i + R_i}. \tag{6}$$

The overall precision, recall, and F score are the averages of corresponding values. Moreover, the gold standard is built by marking the character name with a specific kingdom tag. For example, Cao Cao is tagged with "Wei" and Liu Bei is tagged with "Shu." Finally, 308 character names with definite kingdom tags are secured for cluster analysis.

5.2.3. Clustering Result. The result of hierarchical clustering is illustrated in Table 6. The F score achieves the best value of 79.89% when the number of clusters k is 15.

5.3. Improved Clustering Algorithm. In the RTK network, some characters play a vital role in interconnections of different kingdoms, like "Lu Su" between Wu and Shu, "Huang Gai" between Wu and Wei. These characters have a high betweenness according to the definition of betweenness (see Section 4.5). Further, the node betweenness can be extended to "edge betweenness" [14]. The link with a high edge betweenness is often a bridge between different clusters (see red link in Figure 4). Therefore, removing these high-betweenness links by setting a similarity of 0 will reduce the intercluster similarity and improve the clustering result eventually. The removal operation can be introduced as preprocessing before conducting the cluster analysis.

The improved clustering algorithm using edge betweenness is executed, and the result is displayed in Table 7. When the number of removals is zero, it is the baseline of the original algorithm. With an adequate removing operation, the F score reaches a peak of 80.87%. Nevertheless, removing too many links will destroy the whole network and make the F score decline dramatically (see Figure 5).

5.4. Analysis. Data visualization is also given to display the characteristics of historical figures in the RTK network. As hierarchical clustering can be depicted as a tree-based visual dendrogram, we visualize the character relationship in the RTK novel from Chapter 43 to 50, which is a period describing "the battle of Red Cliffs" (see Figure 6).

As can be seen from Figure 6, six parts can be divided manually. H1 and H3 are groups containing characters from "Wu," like Sun Quan and Sun Ce. H2 encompasses main characters from "Shu" and "Wu" in the battle of Red Cliffs: Liu Bei, Guan Yu, Zhuge Liang, Zhou Yu, Lu Su, etc. However, there are two exceptions: Cao Cao and Cheng Yu, because they are highly connected with other main characters in the battle of Red Cliffs. Further, H1, H3, and H2 merge into a bigger cluster in the hierarchical clustering because these characters are from the alliance of "Wu" and "Shu" against Cao's army.

On the other hand, H5 is composed of characters from a large group "Wei," including Xiahou Dun, Xiahou Yuan, Cao Ren, and Cao Hong. H6 includes few characters from "Shu"

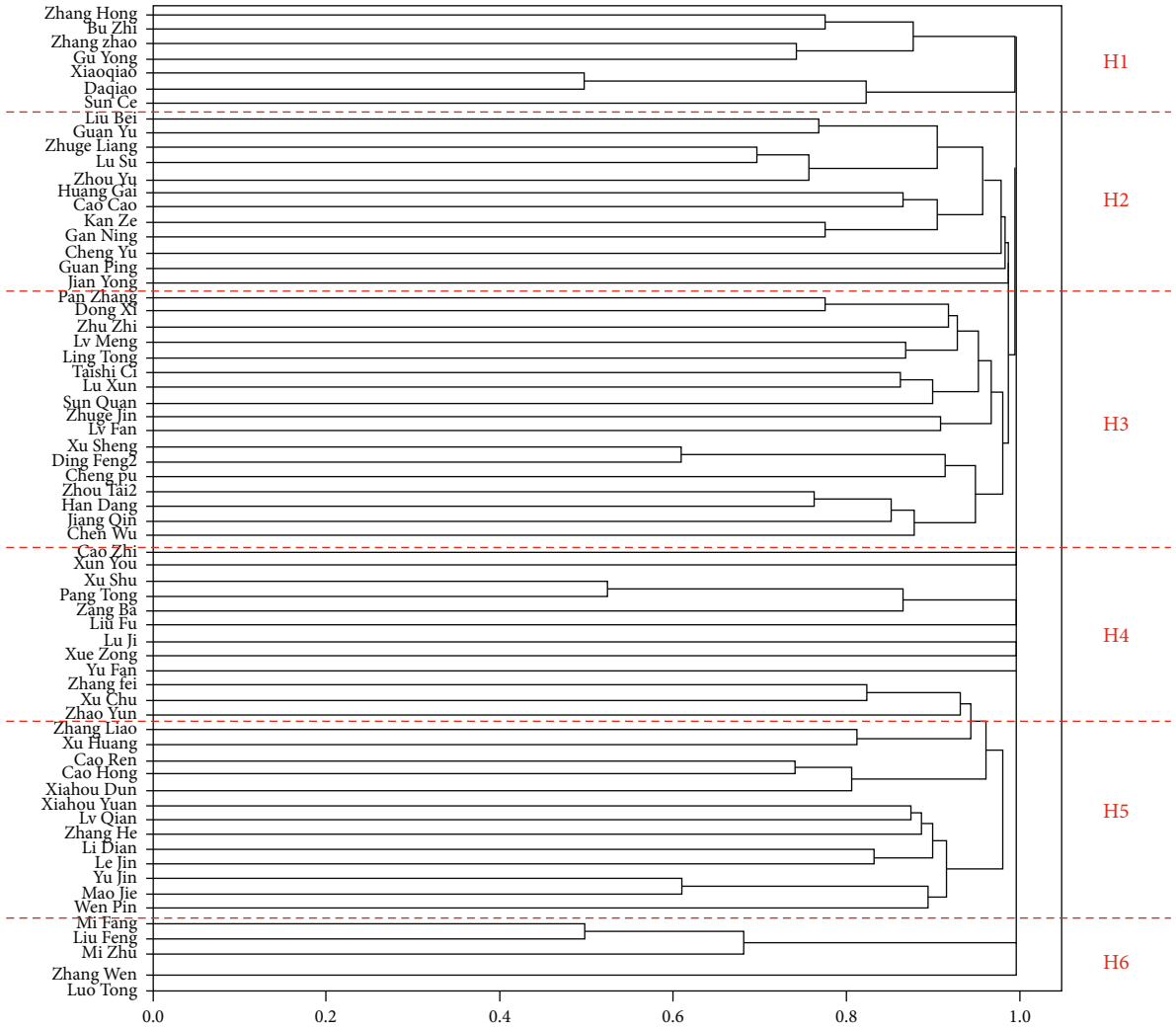

Figure 6: Dendrogram of clustering result for the period of "the battle of Red Cliffs."

or "Wu." H4 is not a cluster, and it contains a number of characters from different kingdoms.

6. Conclusions

This paper developed a general framework for analyzing the character relationship in the novel. The *Romance of the Three Kingdoms* is taken as the object of analysis. At first, the raw text of the RTK novel is processed with NLP tools and character names are recognized by lexical analysis. Then, a character name network is created based on coword analysis. After building the network, several network features are calculated such as degree distribution, average shortest-path length, and clustering coefficient. Besides, cluster analysis is conducted and it helps to better understanding of the hierarchical structure for characters in the RTK novel. A modified clustering algorithm using edge betweenness is proposed to improve the effect of clustering. Finally, visualization of results is completed to analyze the hierarchical clustering.

There are some limitations of the proposed method since coword analysis does not necessarily reflect the true meaning of character relationship. However, our approach can study the main characters quantitatively and comprehend character relationship from another perspective. Hence, it is a valuable research direction.

Subsequent work will study the meaning of pronouns because they represent different characters in different situations. Further, place names and institutions will be taken into consideration in the future.

Data Availability

The original dataset used in this work is available from the corresponding author on request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Youth Foundation of Basic Science Research Program of Jiangnan University, 2019 (No. JUSRP11962), and the High-Level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China, 2019.

References

[1] Q. Zhu, X. Peng, and X. Liu, "Research topics in social computing area based on co-word analysis," *Information Studies: Theory & Application*, vol. 12, pp. 7–11, 2012.

[2] S. Ravikumar, A. Agrahari, and S. N. Singh, "Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics (2005-2010)," *Scientometrics*, vol. 102, no. 1, pp. 929–955, 2015.

[3] D. Nguyen, "Mapping knowledge domains of non-biomedical modalities: a large-scale co-word analysis of literature 1987-2017," *Social Science & Medicine*, vol. 233, pp. 1–12, 2019.

[4] A. de la Hoz-Correa, F. Muñoz-Leiva, and M. Bakucz, "Past themes and future trends in medical tourism research: a co-word analysis," *Tourism Management*, vol. 65, pp. 200–211, 2018.

[5] D. Corrales-Garay, M. Ortiz-de-Urbina-Criado, and E. M. Mora-Valentín, "Knowledge areas, themes and future research on open data: a co-word analysis," *Government Information Quarterly*, vol. 36, no. 1, pp. 77–87, 2019.

[6] C. Fan, "Research on relationships of characters in the dream of the red chamber based on co-word analysis," *ICIC Express Letters Part B: Applications*, vol. 11, no. 5, pp. 1–8, 2020.

[7] Y. Wang, J. Yu, and C. Zhao, "Research on application of co-word analysis on relationships of characters in the romance of the three kingdoms," *Information Research*, vol. 7, pp. 52–56, 2017.

[8] A. Ishizaka, B. Lokman, and M. Tasiou, "A stochastic multi-criteria divisive hierarchical clustering algorithm," *Omega*, vol. 11, 2020.

[9] N. Liu, Z. Xu, X. J. Zeng, and P. Ren, "An agglomerative hierarchical clustering algorithm for linear ordinal rankings," *Information Sciences*, vol. 557, pp. 170–193, 2021.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[11] C. Fan and F. Toriumi, "High-modularity network generation model based on the muitilayer network," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 32, no. 6, pp. B-H42_1–B-H4211, 2017.

[12] Q. Zhou and L. Leydesdorff, "The normalization of occurrence and co-occurrence matrices in bibliometrics using cosine similarities and Ochiai coefficients," *Journal of the Association for Information Science & Technology*, vol. 67, no. 11, pp. 1–25, 2016.

[13] A. Hotho, S. Staab, and G. Stumme, "WordNet improves text document clustering," *Proceedings of the SIGIR 2003 Semantic Web Workshop*, pp. 541–544, 2003.

[14] M. Givan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.