ICIC Express Letters
Part B: Applications
Volume **11**, Number **5**, May **2020**

 ISSN 2185-2766
pp. **493–500**

RESEARCH ON RELATIONSHIPS OF CHARACTERS IN THE *DREAM OF THE RED CHAMBER* BASED ON CO-WORD ANALYSIS

Chao Fan

School of Digital Media
Jiangnan University
No. 1800, Lihu Avenue, Binhu District, Wuxi 214122, P. R. China
fanchao@jiangnan.edu.cn

Received November 2019; accepted February 2020

Abstract. *Dream of the Red Chamber is generally considered to be the pinnacle of Chinese fiction in classical novels. More than 200 personalities are vividly portrayed in this masterpiece. In this paper, social network tools are utilized to analyze relationships of characters in the novel. For one thing, names of characters are found by named-entity recognition (NER) and social network of characters is created by co-word analysis. For another thing, clustering and social network analysis are selected to quantitatively study relationships of characters. Finally, visualized results are presented and several interesting conclusions can be drawn from our experiments.*
Keywords: *Dream of the Red Chamber*, Relationships of characters, Social network, Co-word analysis, Hierarchical clustering

1. **Introduction.** *Dream of the Red Chamber*[1] (红楼梦), written by Cao Xueqin and Gao E in the middle of the 18th century, is reckoned as one of the four great classical novels of Chinese literature. It tells a story about the decline of a noble family, the Jia family, in a fictional dynasty, depicting real life of various characters in a realistic yet fairytale-like way. There exist a multitude of comparative studies between this classic work and the Japanese *The Tale of Genji* by Murasaki Shikibu. Both of them are acknowledged as celebrated works of literature in the world.

Early studies focused on qualitative research. Many scholars studied the writing style and authorship of the novel with statistical data. Karlgren is the first person who analyzed *Dream of the Red Chamber* utilizing mathematical statistics. He selected 38 words for comparison and drew the conclusion that all 120 chapters were completed by Cao Xueqin [1]. Nevertheless, Chen's experiments led to an opposite conclusion [2]. Wei chose 5 scenario indicators (flowers, trees, diet, medicine and poetry) and counted the frequency of their appearance to explore the writing style of author [3]. During the last ten years, numerous researchers have been studying the classical novel from the standpoint of machine learning. Shi classified 120 chapters with SVM algorithm by exploiting 42 function words as feature vectors [4]. Li and Liu analyzed the novel based on n-gram models and random forest classifier [5]. Liu and Xiao adopted clustering method such as hierarchical clustering and K-means, where function words, n-gram model of words and part-of-speech, all content words and the word length were taken into consideration [6,7]. Ye also researched on authorship based on clustering of statistical stylistic features [8]. Jiang examined the novel with different machine learning methods in a comprehensive way [9].

In this paper, we concentrate on characters of *Dream of the Red Chamber* and their relationships because the masterpiece has shaped a large number of characters, such as

DOI: 10.24507/icicelb.11.05.493

[1] *Dream of the Red Chamber* is also translated as *A Dream of Red Mansions*.

Jia Baoyu, Lin Daiyu, and Xue Baochai. We attempt to propose a research framework for studying relationships of the characters of *Dream of the Red Chamber* quantitatively by using social network analysis tools. Firstly, NER, a natural language processing (NLP) technique, is leveraged to identify the characters' names from the natural language text of corpus [10] because we cannot acquire list of all names from the Internet. Secondly, we build a social network for relationships of characters based on co-word analysis, including data preprocessing, word frequency calculation, and network construction. Finally, we perform a network analysis and data visualization to show the characteristics of main characters in the novel.

This paper is organized as follows. Section 2 introduces related work regarding quantitative analysis of *Dream of the Red Chamber*. Section 3 and Section 4 present the data preparation and experiment process. Section 5 reaches the conclusion of this paper.

2. Related Work.

Social network analysis is a quantitative way to study *Dream of the Red Chamber* from the perspective of character relationships. Co-word analysis is an important method to construct social network based on text contents. There is a certain intrinsic relationship between the two words when they appear in the same document. And the more times they appear together, the closer they are. This method was first proposed by French bibliographers in the 1970s [16] and introduced into the field of information science by Callon et al. [15].

A co-word analysis was adopted to study research literature on social computing [16]. Ravikumar et al. borrowed this tool and text mining to explore the intellectual structure of scientometrics [17]. Nguyen did research on the medical literature using a co-word analysis [18]. de la Hoz-Correa et al. utilized it to elucidate the thematic evolution of medical tourism research from 1931 to 2016 [19]. Corrales-Garay analyzed the knowledge areas and themes on open data with such a tool [20]. Wang et al. applied a co-word network to inspecting relationships of characters in *the Romance of Three Kingdoms* [21].

This paper tries to use the co-word analysis technique to quantitatively study the *Dream of the Red Chamber* from a social network's angle.

3. Data Preparation.

3.1. Recognition of characters' names.

How many characters appeared in the *Dream of the Red Chamber*? A researcher in the Qing Dynasty indicated that there were a total of 448 people including named and unnamed characters [12]. In this paper, we only take all characters with definite names as research objects. The corpus is based on the version published by People's Literature Publishing House in 2000 [10].

NER is used to expand the name list because only about 100 main characters' names can be found through the Internet. We took the ICTCLAS[2] Chinese lexical analysis system to segment Chinese sentence to words with tags (part of speech), and found out each word with an "nr" tag by NER ("nr" denotes an identified name). Besides, the names were sorted in descending order by word frequency. Excluding 100 main characters' names and misidentified names from high-frequency names, the rest of names were checked manually by looking into original text in corpus. Finally, we succeeded in obtaining 265 names through expansion. The actual number of characters in the original work is larger than this figure since low-frequency names that only appear one or two times cannot be found.

3.2. Construction of character relationship network.

Original novel was divided into several lines in units of sentences. Based on co-word analysis, we count the number of co-occurrences for each of the two characters in the same sentence. Taking the character name as the node and the co-occurrence as the link, a weighted undirected network of character relationships can be constructed.

[2]ICTCLAS [CP/OL] http://ictclas.nlpir.org/

TABLE 1. Newly identified names by NER

New character name	Chinese character	Word frequency
Jin Gui	金桂	57
Feng Ziying	冯紫英	56
Qin Shi	秦氏	54
Li Gui	李贵	24
Qiao Jie	巧姐	24
Xing Xiuyan	邢岫烟	21
...	...	...

As for processing strategies of names, we treated a full name and its abbreviated name as one name. For instance, Jia Baoyu and Baoyu would be regarded as one node in the network. As a consequence, the established network contained 265 nodes and 2,465 links. The weight of each link is the number of co-occurrences for two characters. Figure 1 shows relationships of characters in *Dream of the Red Chamber*.

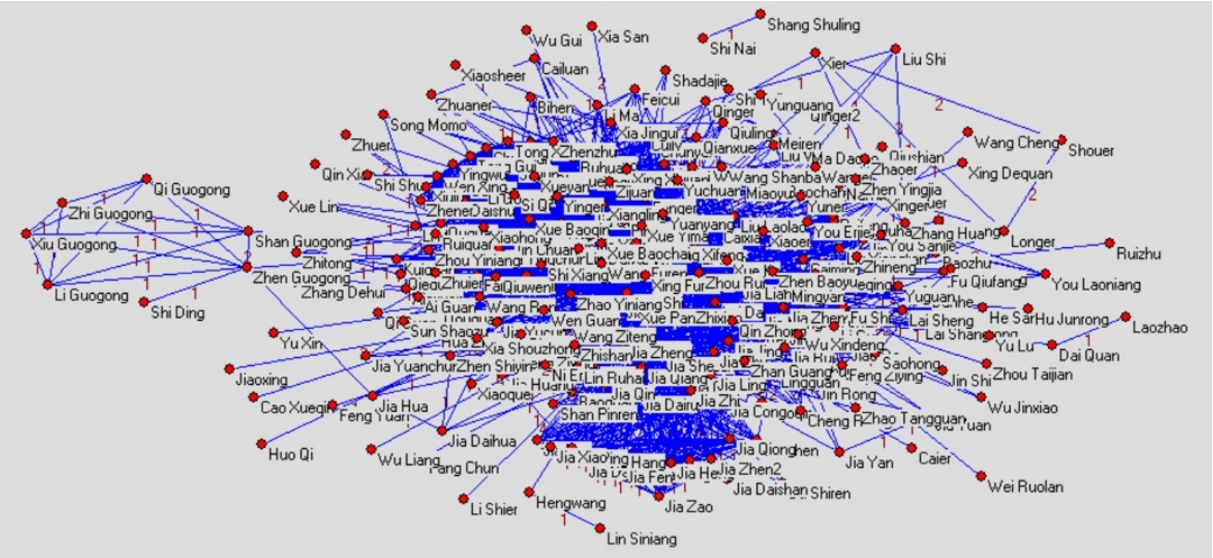

FIGURE 1. Character relationship network of 120 chapters

4. **Experiment and Analysis.** In this section, we introduce prevalent methods in social network analysis. Some tools such as Pajek are employed in this research. Furthermore, a hierarchical clustering algorithm is implemented to conduct clustering analysis.

4.1. Network features.

4.1.1. *Degree distribution.* The degree of a node is the number of links that are incident with it [11]. The average degree of constructed network is 18.604, which indicates that each person is connected with average 18 people. Therefore, nodes are closely linked with each other in the entire network.

Degree distribution $p(k)$ gives a description of probability that one node interacts with k other nodes. If the degree distribution of a network follows a power-law, it is determined by a power index γ [13]. As illustrated in Figure 2, the number of nodes with degrees 1 and 2 is much smaller than the actual number because we ignored low-frequency names when recognizing characters' names from corpus. If the name list is complete, we can boldly estimate that the degree distribution will obey a power-law.

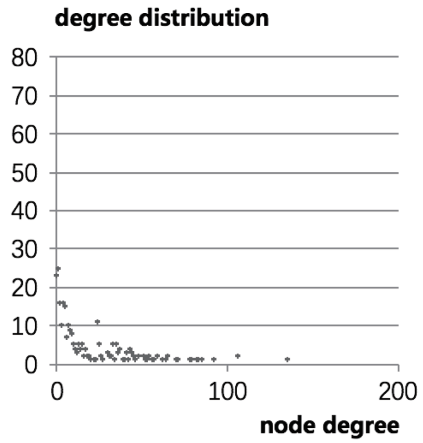

FIGURE 2. Relation between node degree and degree probability

4.1.2. *Average shortest-path length.* Shortest-path length is defined as the number of links composing the shortest path between node pairs in a network [14]. Accordingly, average shortest-path length is simply calculated as an average of all shortest-path lengths. Average shortest-path length of this network is 2.357. That is to say, one character can be linked to any other character in an average of two steps.

The diameter of the relationship network is 6, and the longest path is from Lao Zhao to Shi Ding (Lao Zhao, Dai Quan, Jia Zhen, Zhou Rui, Jia Yingchun, Zhen Guogong, Shi Ding). This result justified the theory of six degrees of separation, which means everyone is six or fewer steps away.

The distribution of shortest-path length between any two characters is depicted in Figure 3. As can be seen from the figure, 53.42% of the shortest-path length is 2 and about 85% is composed of length 2 and 3.

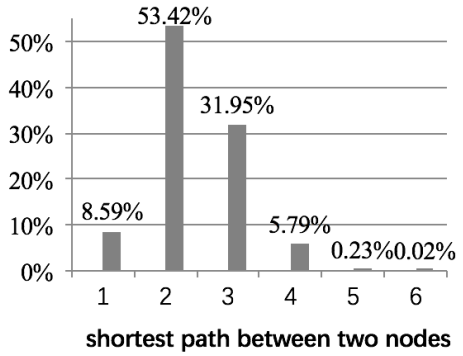

FIGURE 3. Distribution of shortest path between two nodes

4.1.3. *Clustering coefficient.* The clustering coefficient is the average probability that two neighbors of one node are connected [14]. For node i, clustering coefficient C_i is given by the ratio of existing links between its neighbors to the number of potential links. The clustering coefficient reflects the degree to which nodes tend to cluster together. A person has a high clustering coefficient if most friends of his are also friends of each other. Clustering coefficient C of networks is the average value of C_i.

Using Erdős-Rényi (ER) model [11], we created a random graph with same nodes and links as character relationship network of *Dream of the Red Chamber* (DRC network). It is shown in Table 2 that the DRC network has a smaller average shortest-path length and a larger clustering coefficient, thereby displaying a small-world property.

Table 2. Comparison between constructed network and random network

	Number of nodes	**Number of links**	**Average degree**	**Average shortest-path length**	**Clustering coefficient**
DRC network	265	2,465	18.604	2.357	0.602
Random network	265	2,465	18.604	2.726	0.037

4.1.4. *Network density.* The density of a network is the proportion of possible links that are actually present in the network [11]. It can be defined as following formula:

$$d(G) = \frac{2L}{N(N-1)} \tag{1}$$

where the number of nodes and links are denoted as N and L. The 265 nodes may have up to 34,980 links but there are actually 2,465 links. Hence, DRC network's density is 0.0705.

4.2. **Centrality.** Centrality is capable of identifying the most significant nodes within a network. There are different types of centrality as follows.

4.2.1. *Degree centrality.* Degree is a centrality measure which counts the number of links incident to a node. The top 10 characters with the highest degree are displayed in the second column of Table 3. Each character node with a large degree has many connections to other nodes in the network, and is thus in a local central position.

Table 3. The top 10 characters with the highest centrality

Ranking	**Degree centrality**	**Betweenness centrality**	**Closeness centrality**
1	Jia Baoyu (135)	Jia Baoyu (0.1711)	Jia Baoyu (0.6132)
2	Jia Mu (106)	Jia Zheng (0.0787)	Jia Mu (0.5608)
3	Wang Xifeng (106)	Wang Xifeng (0.0689)	Wang Xifeng (0.5608)
4	Wang Furen (92)	Jia Lian (0.0511)	Wang Furen (0.5439)
5	Xue Baochai (85)	Jia Mu (0.0486)	Jia Zheng (0.5305)
6	Jia Zheng (83)	Jia Zhen (0.0463)	Xue Baochai (0.5279)
7	Lin Daiyu (82)	Jia Yingchun (0.0407)	Lin Daiyu (0.5254)
8	Ping Er (79)	Wang Furen (0.0338)	Xi Ren (0.5216)
9	Xi Ren (78)	Jia Yun (0.0287)	Ping Er (0.5166)
10	Jia Lian (71)	Xi Ren (0.0270)	Jia Lian (0.5141)

4.2.2. *Betweenness centrality.* Betweenness is equal to the number of shortest paths from all nodes to all others that pass through that node. A node having a property of high betweenness might have some control over the interactions between two nonadjacent nodes [20]. The top 10 characters with the largest betweenness centrality can be found in the third column of Table 3.

As is presented in Figure 4, the distribution of betweenness appears to be a heavy-tailed distribution. The sum of the top 10 betweenness accounts for 59.48% of the total and top 20 reaches 71.18%. These people occupy important positions in the whole network and have the ability to control the interaction of other characters.

4.2.3. *Closeness centrality.* Closeness centrality is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the network. It measures how close a node is to all the other nodes [18]. The fourth column of Table 3 provides the top 10 characters with the largest closeness centrality. Each character holds a central position and can quickly interact with all others.

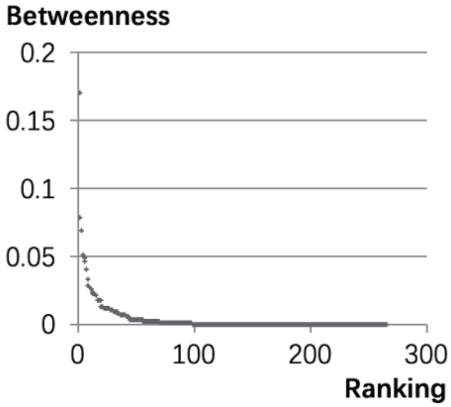

FIGURE 4. Distribution of betweenness for all characters

Seven characters appear in the top 10 for three centrality ranking, including Jia Baoyu, Jia Mu, Wang Xifeng, Jia Zheng, Wang Furen, Jia Lian, and Xi Ren.

4.3. **Clustering analysis.** Before starting clustering analysis, we need to construct two matrices: co-occurrence matrix and correlation matrix. Tables 4 and 5 represent samples of two matrices (containing some high-frequency characters). In co-occurrence matrix, data on the diagonal are the total frequencies of the characters' names. The co-occurrence is affected by the frequency of each name. In order to accurately reveal the co-occurrence relationship, this paper used Ochiai coefficient to convert the co-occurrence matrix into a correlation matrix. Ochiai coefficient can be represented as the following formula:

$$K = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}} \tag{2}$$

where $n(A)$ is the frequency of A, $n(A \cap B)$ is the number of co-occurrence.

TABLE 4. Co-occurrence matrix for some characters

Characters with high frequency	**Jia Baoyu**	**Lin Daiyu**	**Xue Baochai**	**Jia Mu**	**Xi Ren**
Jia Baoyu	2309	249	175	193	241
Lin Daiyu	249	1015	124	69	38
Xue Baochai	175	124	1075	69	64
Jia Mu	193	69	69	1446	36
Xi Ren	241	38	64	36	833

TABLE 5. Correlation matrix for some characters

Characters with high frequency	**Jia Baoyu**	**Lin Daiyu**	**Xue Baochai**	**Jia Mu**	**Xi Ren**
Jia Baoyu	1	0.1627	0.1111	0.1056	0.1738
Lin Daiyu	0.1627	1	0.1187	0.0570	0.0413
Xue Baochai	0.1111	0.1187	1	0.0553	0.0676
Jia Mu	0.1056	0.0570	0.0553	1	0.0328
Xi Ren	0.1738	0.0413	0.0676	0.0328	1

We implemented a hierarchical clustering algorithm to cluster with the correlation matrix. The algorithm can accept a cluster number or a threshold as input. Results of hierarchical clustering are illustrated in Figures 5 and 6.

The main characters in the novel can be clustered into 4 large groups and other small groups. First cluster is a big cluster containing characters from Jia family (e.g., Jia

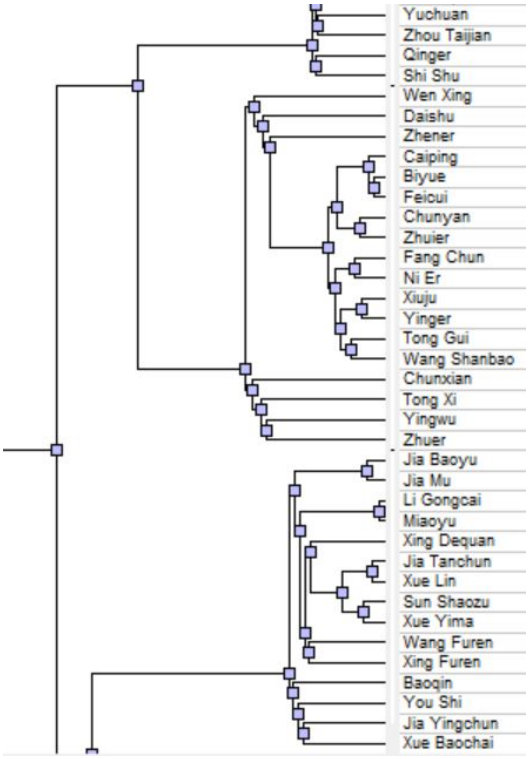

FIGURE 5. Part of result for hierarchical clustering analysis

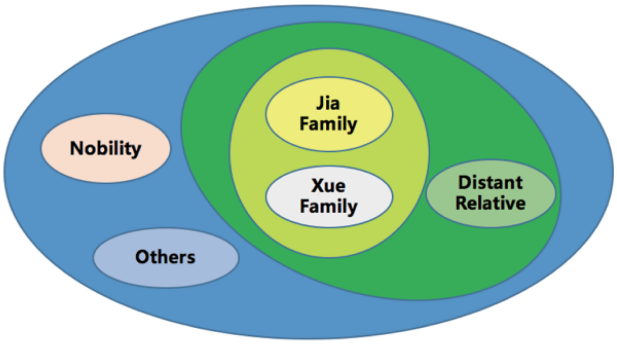

FIGURE 6. Hierarchical structure of characters in the *Dream of the Red Chamber*

Baoyu, Jia Mu). Second cluster incorporates characters mainly from Xue family such as Xue Pan and Xue Ke. Third cluster includes distant relatives, like Jia Xiao, Jia Zao, and Jia Fang. Fourth cluster is composed of noble people with titles. For instance, there are Xiu Guogong, and Zhi Guogong. Jia and Xue family merge into a bigger cluster in hierarchical clustering because they are connected by marriage. Moreover, they combine with distant relatives to form a larger group, which is also clustered with nobility and other small groups.

5. **Conclusions.** This paper has attempted to propose a research framework for analyzing relationships of characters by utilizing methodology of social network analysis. We calculated network features and centralities, performing hierarchical clustering analysis. Some meaningful conclusions were drawn. *Dream of the Red Chamber* is a small world. Besides, a small number of people are in an important position in the relationship network, such as Jia Baoyu, Jia Mu, Wang Xifeng, Jia Zheng, Wang Furen, Jia Lian, and Xi Ren. Finally, there exists a hierarchical structure for characters in this novel. Although

co-word analysis does not involve specific meaning of relationships of characters, it enriches the research content of novels. From this perspective, the ideas and methods of this paper are valuable.

In future, we will expand character name list by improving NER algorithm. Also, the novel will be divided into different parts, and each part can be analyzed respectively. Further, identification of the authorship of first 80 and last 40 chapters will be considered.

Acknowledgment. This paper is supported by the Youth Foundation of Basic Science Research Program of Jiangnan University, 2019 (No. JUSRP11962).

REFERENCES

[1] B. Karlgren, New excursions in Chinese grammar, *Bulletin of the Museum of Far Eastern Antiquities*, vol.24, pp.51-80, 1952.

[2] D. Chen, Identification of the authorship of the last 40 chapters of "A Dream of Red Mansions" from the aspect of mathematical linguistic: Discuss with Chen Bingzao, *Studies on "A Dream of Red Mansions"*, no.1, pp.293-318, 1987.

[3] B. Wei, Statistical analysis on the differences of writing style between first 80 chapters and last 40 chapters in "Dream of Red Mansions": An application of equivalent test on two independent binominal populations, *Chinese Journal of Applied Probability and Statistics*, vol.25, pp.441-448, 2009.

[4] J. Shi, The authorship research on A Dream of Red Mansions based on support vector machine, *Studies on "A Dream of Red Mansions"*, no.5, pp.35-52, 2011.

[5] H. Li and Y. Liu, Language models and classification analysis for Dream of the Red Chamber, *Proc. of the 2nd Conference on Cloud Computing & Intelligence Systems*, Hangzhou, China, pp.1459-1464, 2012.

[6] Y. Liu and T. Xiao, Studies on quantitative styles of "A Dream of Red Mansions", *Studies on "A Dream of Red Mansions"*, no.4, pp.260-281, 2014.

[7] T. Xiao and Y. Liu, Words and n-gram models analysis for "A Dream of Red Mansions", *New Technology of Library and Information Service*, vol.4, pp.50-57, 2015.

[8] L. Ye, The authorship research on A Dream of Red Mansions based on clustering of statistical stylistic features, *Studies on "A Dream of Red Mansions"*, no.5, pp.312-324, 2016.

[9] N. Jiang, *A Study of the Author of A Dream of Red Mansions Based on Machine Learning*, Master Thesis, Zhejiang University, Hangzhou, 2018.

[10] X. Cao and E Gao, *A Dream of Red Mansions*, People's Literature Publishing House, Beijing, 2000.

[11] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, vol.506, 1994.

[12] Y. Zhu, *Character's Illustration of Dream of the Red Chamber*, Baihua Literature and Art Publishing House, 1997.

[13] A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science*, vol.286, no.5439, pp.509-512, 1999.

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

[15] M. Callon et al., From translations to problematic networks: An introduction to co-word analysis, *Social Science Information*, vol.22, pp.191-235, 1983.

[16] Q. Zhu, X. Peng and X. Liu, Research topics in social computing area based on co-word analysis, *Information Studies: Theory & Application*, no.12, pp.7-11, 2012.

[17] 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.

[18] 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.

[19] 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.

[20] D. Corrales-Garay et al., Knowledge areas, themes and future research on open data: A co-word analysis, *Government Information Quarterly*, vol.36, no.1, pp.77-87, 2019.

[21] 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*, no.7, pp.52-56, 2017.