Advances in Social Science, Education and Humanities Research, volume 196
2nd International Conference on Social Science, Public Health and Education (SSPHE 2018)

Analysis of Users' Health Knowledge Requirement and Health Perception in Senior Online Community Based on Web Text Mining

Yuxing Qian*
Center for the Studies of Information Resources
Wuhan University
Wuhan, China
qianyuxing@whu.edu.cn

Huayang Zhou
Center for the Studies of Information Resources
Wuhan University
Wuhan, China
zhouhuayang@whu.edu.cn

Hao Li
School of Health Science
Wuhan University
Wuhan, China
leohao@whu.edu.cn

Meiling Ren
School of Health Science
Wuhan University
Wuhan, China
renmeiling@whu.edu.cn

Wenxuan Gui
School of Computer Science
Wuhan University
Wuhan, China
guiwenxuan@whu.edu.cn

Liqin Zhou
Center for the Studies of Information Resources
Wuhan University
Wuhan, China
zhoulq92@163.com

Abstract—Purpose/Significance: In order to use the Internet to carry out accurate health education, it is essential to study the health knowledge requirement and health perception of netizens. Method/Process: Selecting 5,296 User-shared health knowledge text as a corpus in the "health, regimen, fitness" section of the Elderly Forum "Home for the Elderly". Applying the method of web text mining to carry out the data cleaning and natural language processing of the text in the corpus. TF-IDF keyword extraction algorithm is used to extract the keywords from each text. The Co-keywords network is established to present the structure, scale, and distribution of community content. Its characteristics are analyzed to reflect the current health knowledge requirement and health perception of users. Results: The senior online community user knowledge requirements can be divided into four types: the principles and methods of traditional Chinese medicine, lifestyle adjustments and changes, disease prevention and coping with aging, nutritive value and efficacy of the food. There are interlaced relationships between different types of knowledge requirement; the demand for health knowledge revealed by the user stays at the level of physical health. The mental health and social adaptability are the potential knowledge requirements.

Keywords—online health community, web text mining, health knowledge requirement, health perception, senior

I. INTRODUCTION

With the popularity and technological development of the Internet, online health education is supposed to be an important method for health education[1]. In order to use the Internet to provide accurate education of health knowledge, popularize healthy lifestyle, improve residents' health management ability and health literacy, the study of users' health knowledge requirements and perceptions is the basic work.

The senior population has become the main source of the increment in netizens in China. As the number of senior Internet users grows, the promotion in their health resulted from the Internet and new media has drawn broad attention. Researches show that health information in the network is most attractive to older users and that they make use of the Internet for information concerning health can improve their own health condition to a certain degree[2].

The popularization of the Internet makes it feasible for the senior to socialize online, so the online communities aimed at the senior people are the gathering point for senior netizens now. These communities always set up section which focuses on health issues, providing a platform to share and exchange knowledge of health, which corresponds with their requirements and perception. Taking advantage of these user-generated content can be used to analyze users' health knowledge requirements and health perceptions.

Web text mining can be used to analyze a large amount of text data generated by these online community users. Web text mining is the mining of a large number of unstructured, heterogeneous web content and other web texts. Due to the widespread use of the Internet in various industries, web text mining involves a wide range of topics, and the content of the mining is complex[3-6]. The process of web text mining includes web text acquisition, data cleaning, word segmentation, part-of-speech tagging, removal of stop words, information extraction, and word relationship extraction, quality assessment and visualization of results. In web text mining, keyword extraction is a commonly used method of information extraction[7]. Keyword extraction refers to the automatic extraction of a number of representative words or phrases from the corpus to reflect the main semantic information of the text. It uses co-word analysis, calculates the co-occurrence of keywords in the corpus, reflects the correlation strength between keywords, and divides the keywords of different degrees of relevance in the network into different communities taking advantage of the clustering algorithm. Combined with the methods of social network analysis, it is generally applied to the subject knowledge

Funds for International Cooperation and Exchange of the National Natural Science Foundation of China. "Research on Intelligent Home Care Platform based on Chronic Diseases Knowledge Management". Project No. 71661167007

structure analysis, intense topic research and public opinion monitoring of online social media. However, the application of web text mining in analyzing user knowledge requirements is scarce, and related research is still rare.

In summary, this paper applies the method of web text mining, based on the analysis of the content generated by online community users, identifies the user's health knowledge requirements and health perception.

II. Materials and Methods

A. Research Process

According to the process of web text mining, the research procedure of this paper includes data preparation, keywords extraction, statistical analysis, co-word analysis, and based on the above process, user's health knowledge requirements and health perception are identified.

B. Data Sources

After browsing and comparing websites of the senior worldwide, the “Home for the Elderly” (http://www.lnrzj.com) was chosen as the data sources because it has a higher number of postings, more active users, greater influence, and no advertisements in the forum. Founded on May 1, 2010, “Home for the Elderly” is a professional large-scale online community for the elderly, including sections such as poetry sharing, travel strategy, photography, and health. The “health, regimen, fitness” section is a health communication platform based on this elderly online community. It mainly publishes popular health knowledge including health care and disease prevention. The contents mainly come from personal experience and online reprint, which accumulated a large number of user-generated web text data. It is a reliable data source for studying the health knowledge requirements and health perception of the elderly. The users’ ID, post text and post time were obtained by the web crawler and the time ranged from December 16, 2010 to January 19, 2018.

C. Data Preprocessing

A total of 5,296 posts are obtained and are saved in a structured form in an electronic form. The irrelevant and repeated postings are deleted. Finally, 5,099 pieces of posts are retained. The crawled data includes the image links and other web links of related content in the post content. The Python language is used to programmatically remove the links appearing in each post. The Python package of the commonly used "jieba" is used to segment the text in each post. After the word segmentation, the research combines the stopwords list of Harbin Institute of Technology, the stoppage dictionary of Sichuan University Machine Intelligence Laboratory and Baidu to remove stop words in the corpus.

D. Keywords Extraction

The paper uses the TF-IDF functions in the Python package of "jieba" to process the word segmentation and the text data after the stopwords are removed. On the basis of the segmentation, the extracted keywords are derived from the combination of candidate keywords composed of verbs, nouns, and gerunds. The number of generated keywords is set to 5. After extracting the keywords, the keywords that are irrelevant to the research purpose and have no practical significance are added to the stop word list according to the word frequency statistics result and the co-word network visualization effect. The keyword extraction step is repeated until the useless keywords no longer appear in the co-word network.

III. Results

A. High-frequency Keywords

The keyword frequency obtained by the TF-IDF algorithm is sorted and Table I shows the top 20 keywords. Keywords with the highest frequency are senior people, health maintenance, and body.

B. The Distribution of Keyword Frequency

Fig. 1 shows the frequency distribution of the keyword whose frequency is of 20 and below. As the frequency increases, the number of keywords decreases dramatically. When the word frequency exceeds 11, the number of keywords is less than 50. This shows that the content of the corpus is rich and extensive, while a few high-frequency keywords represent the main content of the corpus.

TABLE I. High-frequency Keywords

Keyword	Frequency	Keyword	Frequency
Senior people	676	Exercise	115
Health maintenance	613	Cancer	111
Body	273	Treatment	104
Food	263	Nutrition	96
Longevity	258	Skin	95
Advice	216	Sleep	90
Disease	144	Prevention	88
Patient	142	Docter	87
Diet	128	Brain	78
Sports	122	Massage	77

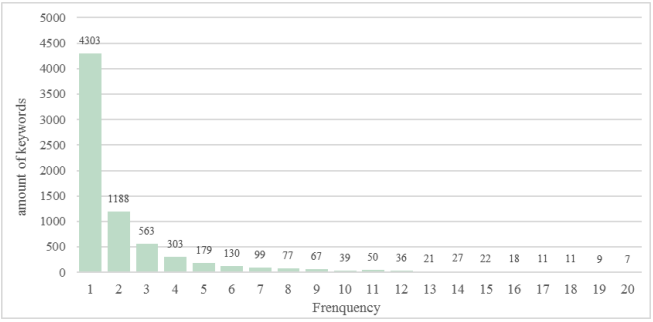

Fig. 1. The distribution of keyword frequency

C. Network Constitution

According to the generated keywords set, the keywords are paired to form the keyword pairs, that is, the pairing relationship of keywords is used to represent the co-occurrence relationship, and after the pairing process, the keyword pair is imported into the network data visualization software Gephi0.9.2, and the co-keywords network is established. This network uses keywords as nodes, and when two keywords appear in the same post, they form a connection between nodes. The average path length and the graph density of the network are calculated in the statistical function module of Geophi, and the characteristics of the network are obtained. In the modular processing, the resolution is set to the default value of 1, the role of modularity is to cluster the keywords, and different categories of nodes are marked with different colors. Due to a large number of low-frequency keywords, the network is very sparse. In order to make the network diagram clearer and the nodes representative, the filtering operation is performed in Gephi, only the nodes with the top 200 degrees are retained, and the average path length and the graph density are calculated again. The results are shown in Table II.

After filtering the nodes, the network density network is greatly improved, and the network diameter is reduced to 3, indicating that the operation of the filtering nodes has obvious effects on simplifying the co-keywords network and presenting the core content.

TABLE II. STATISTICAL INDICATORS OF CO-KEYWORDS NETWORK

Statistical indicators	Entirety	After filtering
Nodes	7285	200
Edges	37337	3138
Graph Density	0.001	0.158
Network Diameter	9	3
Average Path Lenth	3.305	1.856
Degree Range	1-1366	46-1366

TABLE III. THE DIVISION OF COMMUNITIES

Community	Representative Keywords	Main Content
Green	Yang Qi, Qi-blood, Secret Recipe	Chinese Medicine Terms
	Soak Foot, Food Therapy, Conditioning	Method of Health Maintenance
Purple	Retirement, Diet, Sleep	The daily life of senior people
	Joint, Knee, Cervical Spondylosis	Human structure and disease related to motor function
Blue	Protein, Fat, Diet	Nutritional Terms
	Tofu, Egg, Milk	Name of the Food
	Cancer, Liver Cancer.	Cancer-related
Gray	Diseases, Patients, Doctors	Medically vocabulary
	Hypertension, Diabetes, Stroke, Heart Disease	Chronic diseases and complications
Orange	Fever, Cough.	Symptoms of colds and related foods
Red	Massage, Acupoints, Relief	Massage and its related effects

D. Analysis of the Network

The network layout uses the mode of Force Atlas 2. The label size of the node is set according to its betweenness centrality and the color is set according to its community. The co-keywords network is shown in Fig. 2 which is centered on “Health maintenance” and can be divided into six communities. The communities division, the representative keywords and the main content of each community are shown in Table III.

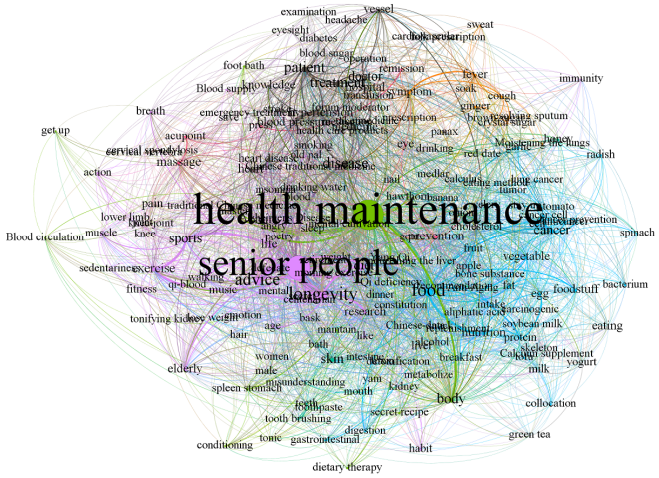

Fig. 2. Co-keywords network

The co-keywords network can clearly present the content of posts in the online community, and the clustering of keywords has a good performance. The keywords can be divided into multiple communities with obvious differences in themes. The network is readable and well interpreted. It can be seen from the figure that the content of the posts in the forum is closely related to the health of the elderly, indicating that the online community is well-targeted and can more accurately reflect the user's health knowledge requirements and perception.

Having analyzed the network structure, the knowledge requirements of the elderly can be divided into the following four types as well as health perception can be identified.

The first type of knowledge requirements is about the principles and methods of TCM(traditional Chinese medicine). The term “health maintenance” is the core of the network, and often associated with specific terms of Chinese medicine and specific health regimens such as diet, foot bathing, and massage. After the economic conditions have been improved, the elderly pay more attention to their own health status in order to further improve their quality of life. TCM is not only reliable and affordable but also enjoys great popularity among the elderly. In addition, the keywords such as longevity and centenarians reflect the expectation of the elderly for longevous life which is the ultimate goal of health maintenance.

The second type of knowledge requirements relates to lifestyle adjustments and alterations. The keyword "senior people" of the network is mainly related to daily life such as exercise, diet, and sleep quality. The senior people are attaching great importance to the knowledge related to lifestyle, such as proper exercise, the principle of nutrition, and how to

promote sleep quality. Based on this knowledge, the elderly adjust and alter their lifestyles to reach the goal of maintaining health and improving the quality of life.

The third type of knowledge requirements contains disease prevention and response to aging. In terms of prevention and treatment of diseases, cardiac-cerebral vascular and metabolic diseases such as hypertension, diabetes, stroke, and heart disease have been recognized as age-related killers. As this information presented in the network, the elderly are highly concerned about these high-risk diseases which reveals their growing demand for health knowledge. In addition, other keywords such as visual acuity, skin disease, cervical spondylosis, insomnia indicate the problems of vision, activity, sleep, aging, and other physiological malfunctions. Aging is an inevitable life course, and the elderly need more feasible knowledge about the aging process and disease preventions

The fourth type of knowledge requirements focuses on the value and utility of diets. From the perspective of food in the network, it includes TCM terms such as health maintaining and diet-related keywords, as well as western medical vocabularies such as protein and fatty acids. It fully shows that the elderly are influenced by traditional Chinese medicine thoughts and the concept of homology of medicine and food which are deeply ingrained in the elderly. Therefore, the scientific explanation of the nutritional effects of diets is equally important for the elderly. In addition, the diet belongs to the same community as the cancer-related keywords. Thus we can infer that the elderly believe that a feasible diet can prevent cancer.

According to the World Health Organization's definition of health, Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity[8]. The mental health and social resilience of the elderly have aroused widespread concern. However, there are few keywords on the mental health and social resilience in the network. The user's requirements for these two levels are not presented in the form of knowledge sharing. It probably due to the elderly's fixed understanding of health which is still remains the level of physiology. When it comes to health problems, the elderly stereotypically view that health is the absence of disease, while neglecting mental health and social adaptability, thus giving rise to inconsistencies in demand and behavior.

IV. Conclusions and Discussion

In general, the user's health knowledge requirements demonstrate the following characteristics: First, the knowledge requirements can be divided into four types: TCM principles and methods, lifestyle adjustments and alterations, disease prevention and response to aging, and diets nutrition value and efficacy. Second, there are complex interlaced relationships between different types of knowledge requirements, rather than independent and divided. For example, diet-related knowledge is both a part of the lifestyle and a measure of health maintaining and disease prevention. Third, the health knowledge requirements revealed by users stagnate at the physiological stage. The knowledge requirements for mental health and social resilience have not been revealed and it is a potential knowledge requirement.

The Online Health Community is a social networking platform that shares and exchanges health information and provides health services[9]. Knowledge sharing in online healthy communities is important for mitigating the shortage of medical resources and uneven distribution[10]. According to the user's health knowledge requirements and health perspective, targeted health education in the online health community would have a good effect.

There are still some shortcomings in this study. As the elderly Internet users may have better family economic conditions and strong social adaptability, resulting in sample bias. In the future, researchers can further study the health knowledge requirements and perception of various elderly groups by combining questionnaires, interviews and other research methods. In addition, data sources can be expanded, such as other elderly online communities at home and abroad, WeChat groups of senior users, etc., to analyze the differences in knowledge requirements and perception between the platforms and at home and abroad, and to extend the research horizon to various social platforms on the Internet.

References

[1] K. T. Win, N. M. Hassan, A. Bonney, and D. Iverson, “Benefits of online health education: perception from consumers and health professionals,” Journal of medical systems, vol. 39, no. 3, p. 27, 2015.

[2] A. Salovaara, A. Lehmuskallio, L. Hedman, P. Valkonen, and J. Näsänen, “Information technologies and transitions in the lives of 55–65-year-olds: The case of colliding life interests,” International journal of human-computer studies, vol. 68, no. 11, pp. 803-821, 2010.

[3] E. K. Ozyirmidokuz and M. H. Ozyirmidokuz, “Analyzing Customer Complaints: A Web Text Mining Application,” in International Conference on Education and Social Sciences, 2014, pp. 507-515.

[4] E. Kahya-Özyirmidokuz, “Analyzing unstructured Facebook social network data through web text mining: A study of online shopping firms in Turkey,” Information Development, vol. 32, no. 1, pp. 70-80, 2016.

[5] W. Li, K. Guo, Y. Shi, L. Zhu, and Y. Zheng, “Improved New Word Detection Method Used in Tourism Field,” Procedia Computer Science, vol. 108, pp. 1251-1260, 2017.

[6] K. K. Lai, L. Yu, and S. Wang, “Multi-agent web text mining on the grid for enterprise decision support,” in Asia-Pacific Web Conference, 2006, pp. 540-544: Springer.

[7] Y.-H. Chen, E. J.-L. Lu, and M. F. Tsai, “Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors,” Expert Systems with Applications, vol. 41, no. 2, pp. 663-670, 2014.

[8] World Health Organization, “What is the WHO definition of health?” World Health Organization, [Online]. Available: http://www.who.int/suggestions/faq/en/ [Accessed: Aug. 12, 2018].

[9] F. Liu, Y. Li, and X. Ju, “Exploring Patients' Consultation Behaviors in the Online Health Community: The Role of Disease Risk,” Telemedicine and E-Health, pp. 8, 2018.

[10] J. Shen, P. Zhu, and M. Xu, “Knowledge Sharing of Online Health Community Based on Cognitive Neuroscience,” Neuroquantology, vol. 16, no. 5, pp. 476-480, May 2018.