Advances in Engineering Research, volume 197
Proceedings of the 2020 9th International Conference on Applied Science, Engineering and Technology (ICASET 2020)

Design and Application of Multidimensional Analysis System for Medical Literature Retrieval

Gui-Min JIE[1,a], Zhen-Guo WANG[2,b,*] and Zhen-Lei YU[3,c]

[1] *Jinan, Shandong, China*

[2] *Shandong University of Traditional Chinese Medicine, China*

[3] *Qilu University of Technology, China*

[a]*amy11203@163.com*, *[b]*Wang1107@163.com*, [c]*Sunny Yu@163.com*

ABSTRACT

Based on the characteristics of clinical medical literature keywords including syndrome, disease and treatment, a multidimensional analysis system for medical literature retrieval keywords is designed to satisfy the needs of users. The system focuses on the analysis of medical literature retrieval keywords, including document information management, word frequency analysis, multidimensional statistical analysis and other functional modules, which can effectively solve the problems of batch retrieval of imported documents, data standardization, multidimensional statistical analysis according to the keyword attribute. Taking the Chinese medicine treatment of psoriasis as an example, the use and application of the system and multidimensional statistical analysis will be introduced. The system is completed to make a quick, convenient and accurate multidimensional analysis of the keywords with the aid of computers, and provides reference for the design and improvement of other medical data analysis systems.

Keywords: *Traditional Chinese medicine, Literature retrieval, Multidimensional Analysis System.*

1. INTRODUCTION

With the rapid development of computer Internet technology, in the era of big data, the traditional medical information retrieval model has been unable to adapt to and satisfy the rapidly changing needs [1]. The "2019 Research Frontier Report" also clearly stated that" with the world of scientific research continuously spreading, growing and evolving, scientific research managers and policy makers need to keep track of the progress and use limited resources to support and promote scientific development ". "The way to define a professional field called the research frontier stems from a certain commonality among scientific studies. This commonality may be derived from experimental data, research methods, or concepts and assumptions, and will be reflected in the academic behavior of scientists citing the work of other scientists in their papers "[2]. Therefore, the extraction and correlation analysis of literature information is of vital importance, which can reveal the development of the discipline and guide the research direction for researchers. Recently, our school library has been working on a project to provide information analysis on "the frontiers and hot spots of psoriasis research" . This project needs to collect a large number of journal literature, and on this basis, analyze the relationship among symptoms - syndrome type, symptom - treatment, syndrome - treatment and drug compatibility of psoriasis one by one. If traditional methods of intelligence analysis are adopted, the whole process will definitely take a long time. For this reason, we have developed an information analysis system suitable for automatic analysis of clinical medical literature. The system realizes fast, convenient and accurate multidimensional analysis of keywords and

extracts relevant knowledge with the aid of computer.

2. SYSTEM DESIGN

Database Design

The system selects the C/S application mode suitable for a large number of local operations, using the object-oriented programming language VISUAL BASIC as the development language, and adopts Sql Server2005 database as the data organization and storage object. The main information is stored in the documentation table (WXJL), keyword table (GJCB), synonym table (TYCB) and part of speech attribute table (CLSX). WXJL is employed for the imported literature records; GJCB keeps the keyword information extracted from WXJL; and TYCB is for synonyms of keywords, such as hypertension, hypertension, high blood pressure, essential hypertension, primary hypertension, early hypertension, etc., are all represented in terms of hypertension. CLSX is used to compare the attributes of keywords and classify them. For example, the blood- heat type belongs to syndrome and Keyin Decoction belongs to treatment.

Functional Design

To meet the current needs of users, the system is mainly designed with four functional modules.

keyword parameter setting. It's mainly used to set the keywords of TYCB and CLSX.

Literature information management. especially the management of basic data of academic papers, viewed as the basis of data analysis. They mainly realize the functions of inputting, editing, and deleting the title, keywords and year information of the papers.

Word frequency analysis. Calculate the frequency of different keywords in all selected papers, sort keywords by frequency and determine high-frequency keywords. On this basis, word frequency tables and graphs are generated to facilitate the researchers to analyze the research focuses and hot spots in this field.

3. MULTIDIMENSIONAL ANALYSIS

It is the core function of the system, which mainly complete the synonym combination, generate multidimensional phrase matrix, matrix word multidimensional analysis, data transfer and other functions. The specific functional structure of the system is shown in Figure 1.

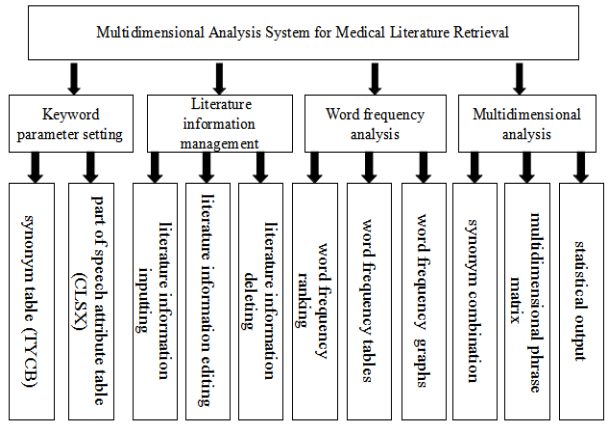

Figure 1. Specific functional structure of the system

Principles of Statistical Analysis

To make multidimensional statistical analysis of keywords is the core function of this system. The principle is to find out the relationship between statistical items of the same nature according to the measurement relationship of various data classifications, which is the measurement aggregation statistics after the dimensional analysis of the data. Among them, dimensionalization is classified according to the characteristics of the data and establish a multidimensional matrix. The specific implementation methods are as follows.

Extraction. Extract keywords or topic words from relevant literature databases, and obtain high-frequency words representing the research topic or research direction of a certain subject through word frequency analysis.

Classification. Classify high-frequency words according to their characteristics to form various classification sets , which are finally combined with each other to form a multidimensional matrix.

Analysis. Analyze the multidimensional matrix, and count the number of simultaneous occurrences of these phrases in the same article.

Here is the main idea of multidimensional analysis. In the data set, if a large number of records with feature attribute A and feature attribute B frequently appear, then feature attributes A and B constitute a frequent pattern [4], which shows the relevance between A and B , and these patterns can be observed and analyzed using association rules.

Analysis and improvement of key technology for the system program development

In order to better show the practical application effects of the keyword multidimensional analysis system, this paper takes "Frontiers and Hot spots of Psoriasis Research" as a case to demonstrate the application and principle of the system. The literature source databases are China Knowledge Network (CNKI), China Biomedical Literature Database (CBM), covering the period from 2010 to 2019. The case takes the retrieval result as the analysis object, enumerates the application of the multidimensional analysis system in word frequency statistics and synonym combination, and conducts statistics from the two dimensions of syndrome type and treatment. The search terms of CNKI include psoriasis, Traditional Chinese medicine, Chinese patent medicine, Chinese traditional medicine, Chinese herbal medicine and prescription. The search method is "theme". CBM retrieval strategy is Psoriasis OR Psoriasis AND Traditional Chinese medicine OR Proprietary Chinese medicine OR Chinese herbal medicine OR prescription.

Import of multiple data sources

According to the search terms, a total of 1389 records are obtained in this case in CBM and 1566 records are in CNKI. With the acquired data decomposed, the literature can be transformed into structured data units that can be processed by computers. Then through comparison and checking, field mapping and merging, make it standardized, accurate and orderly. A total of 1,863 records are obtained after the original data are re-checked and merged.

Keyword standardization

After the structured processing of literature data, the keyword information can be further extracted and standardized. The irregularity of the original keyword data is mainly manifested as Polysemy (the phenomenon of multiple words with the same meaning), that is, synonyms. In the operation, first all non-repetitive keywords are extracted from literature records to form a data list.Then compare and merge the thesaurus one by one. In this case, a total of 3987 keywords were extracted. Through careful discussion and analysis by clinical experts with medical records, 142 synonymous with psoriasis syndrome and treatment were screened out, and 30 standard keywords were formed after merging shown in Figure 2.

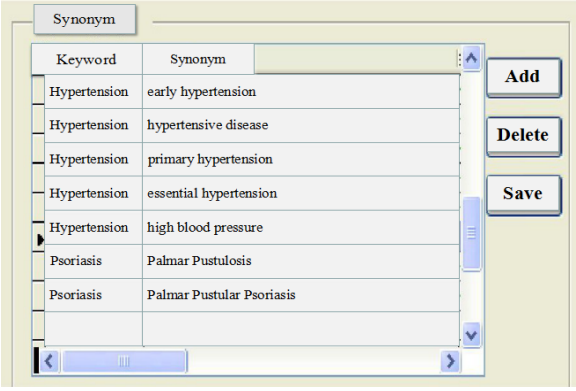

Figure 2. Keyword standardization

Keyword attribute classification

Based on the keyword database data set, frequency analysis is used to form the keyword

database with high frequency. The drop-down box is used to select the corresponding characteristic attributes from the aspects of symptoms, syndromes and treatment. Click the "Add" button to add and save the new feature. In the case, 13 syndrome type records and 21 treatment records were obtained by using 30 standard keywords according to the two characteristic attributes of syndrome type and treatment, which is shown in Figure 3.

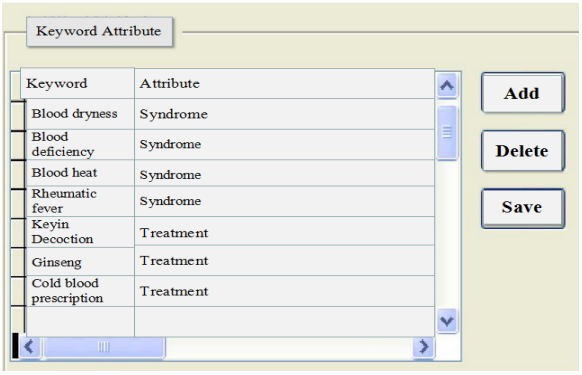

Figure 3. Keyword attribute classification

Multidimensional statistical analysis

With the two dimensions of 12 syndrome records and 18 treatment records, a 12×18 two-dimensional matrix is interactively formed, as is shown in Figure 7. After statistical analysis, the two keywords with the highest common occurrence frequency were "blood heat type" and "cold blood prescription", which appeared 48 times in total, followed by "blood heat type" and "Xiaoyin Fang" for 37 times.

The statistical results are consistent with clinical diagnosis and medical rules. The same statistical method is also applicable for other dimensions such as symptom-symptom type and symptom-treatment. In addition, by consulting the medical records of famous Chinese medicine doctors for the treatment of psoriasis,we can find that there is no fixed type of psoriasis, but blood stasis, blood heat, and blood deficiency are the most common types of psoriasis. Among them, the blood-heat type is the most frequent syndrome type. The cool blood prescription, which is based on dried Rehmannia root, Pentstemon, Radix Paeoniae Alba, Cimicifuga foetida,Glycyrrhiza, and Rhizoma Anemarrhenae, is the most commonly used and effective Chinese medicine prescription for blood-heat psoriasis in this area. Therefore, the pure TCM treatment of psoriasis is also consistent with the statistical results of this system, which increases the credibility of this system to some extent.

The multidimensional analysis of medical keywords can help us understand the research hot spots in the medical field and infer the development direction of its future research. The statistical results obtained can also be exported and saved as Excel, text files and other formats, so as to use other statistical analysis software such as SPSS for further analysis and processing.

4. CONCLUSION

With the continuous development of medical big data, users have higher and higher requirements for medical information service. The design of multidimensional automatic keyword retrieval system provides more and more convenient service support for the development of library information service. The system has been applied to the statistical analysis of relevant literature in the library reference department. The practical application shows that the efficiency of statistical analysis is significantly improved after using the system, and it meets the requirements of user retrieval statistics. At the same time, with users' experience and suggestions on the system, the system will be constantly improved and optimized to provide more accurate and faster literature retrieval and statistics services.

However, there are still some shortcomings in the function and use of this system, which needs to be improved and perfected continuously. Firstly, the function of synonym merging in the system is carried out by manual judgment. Consider adding the auxiliary suggestion function of automatic prompt to complete synonym merging work faster. Besides, how to automatically obtain high-frequency keywords from word frequency statistics and classify their attributes

to reduce the workload of manually entering information. Finally, keyword analysis is extended to automatically obtain relevant keywords from the full text, combined with medical records to achieve automatic analysis [5], so as to investigate the development and changes of medicine in a more comprehensive way.

REFERENCES

[1] Tu Xinli, Liu Bo, Lin Weiwei, Overview of Big Data Research, Computer Application Research. 31 (2014) 1612-1616.

[2] Information on http://www.baogao.com

[3] Lu Chuanjian, Zeng Zhao, Xie Xiuli, Analysis of syndrome distribution in literature of psoriasis vulgaris from 1979 to 2010, Journal of Chinese Medicine. 53 (2012) 959-961.

[4] Han J, Kamber M. Data Mining, Concepts and Technologies, second ed, Beijing, China, 2007.

[5] P. Xu; N. Na; S. Gao; C. Geng, Determination of sodium alginate in algae by near-infrared spectroscopy, Desalination and Water Treatment, 168(2019)117-122.

[6] P. Xu, N. Na, A. M. Mohamad, Investigation the application of pristine graphdiyne (GDY) and boron-doped graphdiyne (BGDY) as an electronic sensor for detection of anticancer drug, Computational and Theoretical Chemistry,1190 (2020): 112996.

[7] Yuan Feng, Design and Implementation of TCM medical case Analysis System based on data Mining, Jinan: Shandong Normal University, 2006.