PRINCIPLE-METHOD-RECIPE-MEDICINE INFORMATION EXTRACTION FROM TRADITIONAL CHINESE MEDICINAL LITERATURE

Yee-Guang Chen

Department of Traditional Chinese Medicine, Tzu-Chi Medical Center, Hualien, Taiwan

(Received 21th October 2011, accepted 15th November 2011)

Electronic text with simple information retrieval function is emerging as a trend over the area of medicine. Since, there was a large volume of Traditional Chinese Medicinal Literatures (TCML) left behind from our ancestors, it is a cost and effectiveness issue while we make use of them electronically and apply them wherever need eTCML with handy ranking functions. In this study, TCML was taken from the Traditional Chinese Medicinal Electronic Text (TCMET) database which has been created successfully and working smoothly as an eTCML retrieval system since 1997 in our laboratory. Furthermore, a cluster of thesaurus has been constructed basing upon the concept of Principle-Method-Recipe-Medicine (PMRM) and acting as the comparison terms. A conventional information retrieval tool by using the term frequency (*tf*) and inverted document frequency (*idf*) is applied to claim for the PMRM document ranking. Through *tf* and *idf* innovation, the relationship between documents and keywords can be indexed and weighted. The resulting table of the PMRM thesauri with 89,462 keywords has already been posted on the TCMET website for references. Additionally, a query with 4 sets of keyword was used to test the documental ranking function from the inverted file scorings which gave us some valuable retrieval information about the TCM concepts in PMRM and led us to rule out the others. The objective of this study is trying to integrate the TCM thesauri and TCML full-text database for promoting our eTCML projects in order to build an advanced TCMET retrieval system in future years.

Key words: Traditional Chinese Medicinal Electronic Text (TCMET), information retrieval, thesaurus, *tf* × *idf*

Introduction

The Traditional Chinese Medicinal Electronic Text (TCMET) website (www.tcmet.com.tw) has been established since 1997. TCMET is a long term project which is dividing into III developmental phases on purpose. In this Phase I study, TCMET has been installed on the website with the ability of Traditional Chinese Medicinal Literature (TCML) website online searching and Phase II gives out primitive text

*Correspondence to: Yee-Guang Chen, Department of Traditional Chinese Medicine, Tzu Chi Medical Center, No.707, Chung Yang Rd., Sec.3, Hualien 970, Taiwan, Tel:+886-3-8561825 ext. 2158, E-mail: tcmet.editor@msa.hinet.net

retrieval functions.

In Phase I development[1] - classical Chinese medicinal literature such as *Huang-Di-Nei-Jing(黃帝內經)*, *Jin-Yuan-Si-Da-Jia(金元四大家)*, and *Jing-Yue-Quan-Shu(景岳全書)* binding electronically, which contains totally about 2 million words, were installed in the Microsoft SQL database server. An end-user interface which was programmed by the Microsoft Active Server Page allows internet users to search keywords from the website database promptly. The TCMET database which was installed among these projects had been created a broad range of indexed terms dealing with classical TCM literatures. Phase I project which focused mainly on website database implementation and with primitive online searching functions has been doing well in the last fifteen years. During Phase II development [2-5], experiments were carried out by retrieving information statistically from certain Chinese medicinal literatures such as *Heijian-Liu-Shu(河間六書)* recipe and formula extractions[2], *Huang-Di-Nei-Jing* n-gram documentary segmentation[3], and *Lei-Zheng-Zhi-Cai(類證治裁)* SQLXML schema integration. The SQLXML project had the TCMET Wrapper/Mediator architecture established by using the schematic structure which given out a knowledge database of TCM diagnosis and treatment with parsing ability[4]. In 2004, a TCML material about contagious disease was extracted from twenty-two different kinds of classical TCM literatures (totally about 10 million Chinese characters)[5] which gave us some insights applying to this TCML retrieval project. Phases I and II project accomplishments allow us to fulfill future Phase III project in the aspects of TCML intellectual retrieval and/or text mining purposes. Making use of content words in TCML and keyword frequencies in texts, are given general clues to their topical importance. Actually, statistical occurrences of TCM terms in text which is one of the crucial methods applying to documentary information retrieval, extraction, clustering, abstracting and analysis[6]. In 1980, Vector-Space Model (VSM) which based on the concept of term frequency and inverted file approach to text indexing was developed by Salton[7]. In 1957, Luhn who discovered the distribution patterns of words gave significant information about the property of being content bearing[8]. Earlier, Zipf in 1949 plotted the logarithm of the frequency of a term in a body of texts against rank[9]. Thus, the constant rank-frequency law of Zipf describes the occurrence characteristics of the vocabulary while the distinct words are arranged in decreasing order of their log frequency of occurrence: log(frequency) × rank = constant. It further implies that words with low significance are at both tails of the distribution[10]. In this study, the mathematical functions were implemented to perform the term and document weighting so that the TCML and query keywords can be ranked and document concepts (PMRM) can be clustered. Nevertheless, this study can be overlooked as a pilot project of our Phase III TCML studies.

Principle-Method-Recipe-Medicine(PMRM理法方藥) these four classic TCM concepts compose the main idea of clinical diagnosis, assessment and treatment for clinical practitioner[11]. Term frequency(*tf*) and inverse document frequency(*idf*) are the two major term weighting factors ranking the document[20]. Ancient classic TCM literature authors used to quote the TCM bible---- *Huang-Di-Nei-Jing* whenever the writer wanted to illustrate an idea of their own or refer back to its earlier tradition. Therefore, the thesaurus of *Huang-Di-Nei-Jing* was presented as the principle of TCM theory. Secondly, definitions of symptoms,

diseases and the methods of treatment were summarized precisely in the book which was published in the *Qing* dynasty ---- *Lei-Zheng-Zhi-Cai* which we did a study on keyword extraction through it[4]. Terms of these definitions were presented as the method of treatment. It is not uncommon in TCM that recipes used to come along with a series of herbal drug name together as well as a formula name. Therefore, herbal drug names come up with a formula name in the same paragraph can be made use to filter out information about recipe and medicine. Formula name and their recipes thesauri were extracted from *Heijian-Liu-Shu*[2] as we did it before.

Materials and Methods

I. Computer-Mediated System

The TCMET website was documented thoroughly[1]. Web Server was built upon the Microsoft NT 2000 server as well as the Internet Information Service (IIS) and database management system was operated in the Microsoft SQL Server. Text retrieval and computing mainly were programmed under the Microsoft Visual Basic 6.0 console. The internet end-user interface and web-database managing were achieved in the Active Server Page website program.

II. Preparation of eTCML

Huang-Di-Nei-Jing was the literature preparing for the TCM principle thesaurus and *Lei-Zheng-Zhi-Cai(類證治裁)* was the resources for TCM's method of treatment and name of diseases thesauri. There were 229 different kinds of herbal medicine as well as recipe's formula names which were extracted from the *Heijian-Liu-Shu* helping to build the recipe and medicine thesauri[2]. The texts of *Jin-Yuan-Si-Da-Jia (金元四大家)* and *Jing-Yue-Quan-Shu(景岳全書)* are TCML with about 2 million words preparing for our full-text extraction in advance.

III. Organize of the PMRM thesaurus as in Fig. 1

Several thesaurus databases had been built in the Phase I and II studies among our laboratory since 1997. The following projects that we have been done, which were referred to the TCML database as in Fig.1, were applied to build the PMRM thesaurus within this study. The objective of the *Heijian-Liu-Shu(河間六書)*[2] project, which was extracted among the information of 229 drug names and 519 recipes, helped us to build the database of drug name and recipes thesauri. While in the project *Huang-Di-Nei-Jin (黃帝內經)*[3], a four-words thesauri had been finished which gave us keywords related to TCM principle. Ancient TCM doctors used to refer *Huang-Di-Nei-Jin* as the golden bible while they were editing or writing TCML. Therefore, it is reasonable for us to consider those keywords, which extracted from *Huang-Di-Nei-Jin,* were related to the TCM principles. The thesauri of 'method of treatment' and 'name of diseases' was created from *Lei-Zheng-Zhi-Cai(類證治裁)*[4]. In our previous studies, we used the term 'keywords' in the sense of generalize-phases which may be either random word combinations or specifically related to TCM, while 'thesaurus' must be defined in our TCM terms with associated meanings. For instance, 'Building an epidemic thesaurus' was a project[5] which giving us a thesauri database with specific diagnosis related to epidemic diseases. We found out that 37 epidemic-disease diagnostic terms, which were well defined and applied as the seeding thesauri, helped us to locating the epidemic-disease information in

TMCL. We may consider that each keyword occurs in the TCML is epidemic-disease thesauri related. Our previous work was really applicable to extract information from a whole bunch of eTCML, like the thesauri of drug name and recipe with extreme high precision-recall rate. In this study, we reapply these TCM thesauri to extract the PMRM information from our eTCML database and come up with rankings and scores. Making use our thesauri databases (principle, method, recipe and medicine), we extract items from our 2 million words' TCML based on keyword searching and matching algorithm. Nevertheless, locating drug name to the eTCML, we may match term frequencies occurring in the texts with exact document locations and term attributes. Therefore, principle, diagnosis-method, recipe, medicine, thesauri that we had been built as described above, could also be applied to filter out PMRM information with ranking and scores as mentioned in the blocks of Query and Retrieval ranked document which is shown in Fig. 1. In this project, we put all these keywords which were extracted PMRM information together and did not work on their hierarchy structure nor association characteristics. The PMRM thesaurus is like a general thesaurus dictionary, which put all the synonyms in a place, explains the words alike. The objective of building the thesauri database is mainly focus on PMRM classifications with the A to E attributes as shown in Table 1.

IV. Information Retrieval (IR)

Term frequency(tf) and inverted document frequency(idf) were parameters introduced by Salton in his Vector Space Model(VSM) over forty years which has been modified and still using now[7,12]. *Tf* and *idf* is a tool can be elicited to build the VSM and there are many other tools can build the VSM, too. This study we are not going to do anything with the VSM.

The term frequency (*tf*) measures the frequency of an index term in the document text:

tfi = frequency of occurrence of the index term *i* in the text.

The inverse document frequency (*idf*) weight is commonly computed as:

idf = log (*N/ni*) where

log = common logarithm

N = number of documents in the reference collection

ni = number of documents in the reference collection having index term *i*.

The specificity of a given term as applied to a given text can be measured by a combination of its frequency of occurrence inside the text (the term

Table 1. Occurrence of keywords with attributes (A – E).

attribute	Content	number of keywords
A	Principles *(Huang-Di-Nei-Jing)*	54951
B	Diagnosis of diseases *(Lei-Zheng-Zhi-Cai)*	200
C	Method of Treatment *(Lei-Zheng-Zhi-Cai)*	1327
D	Recipe *(Heijian-Liu-Shu, Lei-Zheng-Zhi-Cai)*	1771
E	Medicine *(Heijian-Liu-Shu)*	229

frequency or *tf*) and an inverse function of the number of documents in the collection to which it is assigned (the inverse document frequency *idf*). The term weighting function which is commonly use to determine the product of the frequency and the inverse document frequency(*tf* × *idf*) of the index term. Usually, the raw term frequency and the common logarithm of the inverse document frequency are computed as follows:

$tfi \times \log(N/ni)$ where:

tfi = term frequency of an index term j in the text

j = 1 …n (n = number of distinct index term in the text).

Normalization of *tf* × *idf* is that longer documents accumulate more weight in queries simply because they have more words. As such, the commonly approach of the cosine normalization of the document

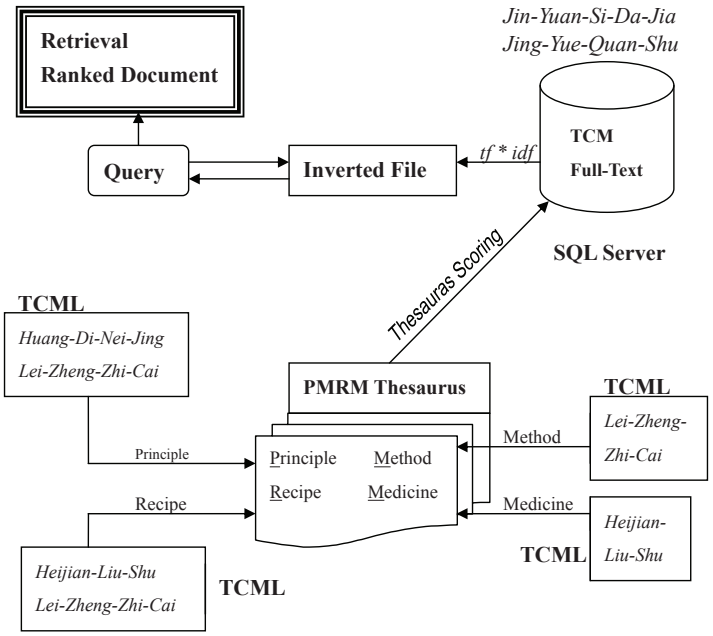

Fig. 1. Components of the system architecture make use of the PMRM thesaurus weighting on the inverted file and generating the query ranking table.

Table 2. Example of keywords match to the PMRM database with weighted scores.

NO	Class	Location	Keywords	*tf*	*Dn*	*Df*	*Cf*	*idf*	*tf* × *idf*
299	C	838	*Fen-hen (發汗)*	1	10	148	274	2.17	2.17
14	B	160	*zhu-tong (諸痛)*	1	1	59	69	2.57	2.57
15	A	7494	*xu-yuen-liu-qi (五運六氣)*	1	1	7	7	3.5	3.5
17	A	21859	*juen-huo-zhi-wei (君火之位)*	1	1	3	3	3.87	3.87
23	E	149	*ma huang (麻黃)*	1	2	321	624	1.84	1.84
240	D	1229	*ma huang tong (麻黃湯)*	1	8	64	86	2.54	2.54

tf = term frequency; *Dn* = Document; *Df* = Document frequency; *Cf* = Collection frequency; *idf* = inverse document frequency.

weight[7] as the following function:

$$\text{Cosine Normalization of Document weight} = \sum_{\text{all query terms}} \text{Weight of term in query} \times \text{Weight of term in document} \Big/ \left(\sqrt{\sum_{\text{all query terms}} (\text{Weight of term in query})^2} \times \sqrt{\sum_{\text{all query terms}} (\text{Weight of term in document})^2} \right)$$

V. Query functions of the PMRM documental ranking as in Fig. 1

As in Fig. 1, PMRM thesauri which were generated from the TCML, will be acted as the terms for the query functions. These keywords (Table 2) being on the query should be a subset to the PMRM thesauri and random keyword is not allowed which may interfere to our ranking results. Thus, the document ranking calculation was accomplished by calculating the *tf* × *idf* scores of the indexed terms.

Results

I. Classifications based on the Thesaurus:

To make use of the *tf* × *idf* function, which generating the TCM classification ranking tables, is the main objective in this study. From our primitive observation, *tf* × *idf* is a tool with cost and effectiveness by using it in eTCML extraction that we can implant those functions in a short period of time. In this study, the texts of *Jin-Yuan-Si-Da-Jia (金元四大家)* and *Jing-Yue-Quan-Shu(景岳全書)* were prepared for information extraction. It was a formatted text with 30 rows by 80 columns in each page and there were around 1,000 Chinese characters on it. Moreover, each page was regarded as a documental unit by which the *tf, idf* scores could be computed. Thereafter, we applied our pre-defined PMRM thesaurus to match the keywords among each unit. Refers to Table 1, keyword such as '*Wu-Yu-Liu-Qi (五運六氣)*' which is classified to the principle's category with attribute 'A' which were extracted from the four-word *Huang-Di-Nei-Jin* thesauri[3] totally with 54951 items, keyword such as '*fei-han*' (*發汗*) is in the category with attribute 'C' which were extracted from the method-of-treatment thesauri[3] totally with 1327 items, '*ma huang tong (麻黃湯)*' as recipe with the attribute 'D' and there are 1771 items[2] (*Heijian-Liu-Shu* 519, *Lei-Zheng-Zhi-Cai* 1252) in the recipe thesauri and finally '*ma huang (麻黃)*' as medicine with the attribute 'E' and there are 229 items[2] in the medicine thesauri. Thesaurus tables, such as (1) principles (2) name of TCM diseases (3) method of treatments (4) name of recipe formula and (5) name of herbal medicine, among these 5 topics were grouped into the thesaurus tables with the attribute-values of A,B,C,D and E. The number of keywords related to each attribute was summarized in Table 1. By using PMRM thesauri, we could apply them to calculate the *tf* × *idf* scores as shown in Table 2.

II. Query functions of the inverted files through the PMRM thesauri (Fig. 1)

After building the table with *tf* × *idf* scores, we could further apply those scores to a query that may give out the ranking weights (Fig. 1 Query block). Table 2 is an example of a term frequency and inverted files scoring with certain keywords that we may apply them to our query. After the *tf* × *idf* scores calculations, a document with the highest ranking may give us summarized information of PMRM which is being on the query. For an example, we are looking for a query with the keywords of: "*shang han(傷寒)*; wu han*(無汗)*; ma huang*(麻黃)*; shui zhing*(水腫)*", these keywords are basic concepts of

using recipe ma huang tong clinically. Their total sum of *tf* × *idf* scores which might appear in each document can be found as in Table 3. TCML document with the number 1974, got its highest *Sum(tf* × *idf) value 50.7* and most likely to concern about TCM recipe *ma huang tong.* We should take a closer look to the results as shown in Table 4 for the query interpretations. It is because *ma huang* is a term with *tf* of 18, which appear in the document 1974, gives the highest scores of *tf* × *idf.* We may interpret that from PMRM's point of view, ma huang is in class E with robust frequencies which may lead us to conclude that document 1974 is likely to concern about the medicine *ma huang(麻黃) with Sum(tf* × *idf) scoring 33.12 and about the disease shang han(傷寒) with Sum(tf* × *idf) scoring 8.30.* This information may further conclude that document 1974 concerns about the disease *shang han(傷寒)* and focuses on the medicine *ma huang(麻黃).* Reviewing document 1974 manually, it is a document with the title of '景岳全書·古方八陣·散陣' which talks about 8 recipes with herbal medicine compositions and treats the disease *shang han(傷寒)* is quite a hint from our query's interpretation. Document 944 with the second ranking, which containing the topic '景岳全書·須集·傷寒典上·麻黃桂枝辨二

Table 3. Document ranking with query grouping terms "*shang han, wu han, ma huang, shui zhing*".

Rank	Sum*(tf* × *idf)*	Document#
1	50.74	1974
2	44.94	944
3	30.66	133
4	25.34	969
5	24.64	941
6	24.38	1977
7	23.76	513
8	23.76	923
9	23.58	184
10	22.74	182

Table 4. Weighting measures of the 1st ranking document# 1974.

Keywords	Class	*tf*	*Idf*	Sum*(tf* × *idf)*
shang han(傷寒)	B	5	1.66	8.30
wu han(無汗)	B	2	2.18	4.36
ma huang(麻黃)	E	18	1.84	33.12
shui zhing(水腫)	B	2	2.48	4.96
				Total 50.74

Table 5. Weighting measures of the 2nd ranking document# 944.

Keywords	Class	*tf*	*Idf*	Sum*(tf* × *idf)*
shang han(傷寒)	B	7	1.66	11.62
wu han(無汗)	B	6	2.18	13.08
ma huang(麻黃)	E	11	1.84	20.24
shui zhing(水腫)	B	0	0	0
				Total 44.94

Table 6 Weighting measures of the 3rd ranking document# 133.

Keywords	Class	*tf*	*Idf*	Sum*(tf* × *idf)*
shang han(傷寒)	B	11	1.66	18.26
wu han(無汗)	B	4	2.18	8.72
ma huang(麻黃)	E	2	1.84	3.68
shui zhing(水腫)	B	0	0	0
				Total 30.66

十四', mainly talks about the theory of using 麻黃 in 傷寒論 as shown in Table 5. Document 133 with the third ranking, '河間劉完素·傷寒標本心法類萃卷上' which mainly talks about introductory reading of 麻黃湯 in 傷寒論 as shown in Table 6. Actually, the result in Table 6 is not only talking about 麻黃湯 but also other recipes too. Perhaps, it is the reason why has less score then document 1974. For a TCM doctor, these outcomes are worth for exploring because the filtered out documents with rankings are really affiliated TCML with PMRM concepts. Although it is just a case study, we may further apply the PMRM *tf* × *idf* scoring system for a complete TCM concept-query studies in the future which may lead us to a more informative direction.

III. Inverted tables implemented on the TCMET website

The inverted file with document weighting was showed only several rows in Table 2. The whole inverted file table in this study is too large to be shown. Nevertheless, it has already been completely posted on the TCMET website with the linkage of (http://www.tcmet.com.tw/vsm/idf.asp). It is a huge table consists of totally 89,462 rows which is exactly the same as in Table 2. We may further apply the scores of this table to calculate the document weight by using the value of *Sum(tf* × *idf)* so that we may have the query of our own TCML keywords.

Discussion

Historical documentary of TCM can be traced back to the age of Qin-Han Dynasty (B.C. 250 - A.D. 220) in China[13]. None of the original copies of TCML from that period of time can be survived till now. Nevertheless, there was a large volume of TCML with strong reputation which used to be duplicated copies or compilations from descending dynasties edited by follower scholars. Internationalization of TCM emerges as the largest alternative medicine field spreading rapidly to the Western and European countries[14]. Invaluable TCM legendary and clinical experiences have been conveyed through TCML for over two thousand years. It is crucial to preserve TCML electronically and we really did in Taiwan for more then thirty years[5]. Information retrieval technology among TCML emerged as an interdisciplinary issue. It is a subject matter which involves the area in TCM, computer technology and information retrieval. The key issue in TCM information extraction is to explore TCM physicians how they are looking for medical knowledge among the TCML. It is used to play an important role to implant the intellectual expert

system by computer scientists[15]. In this study, PMRM was extracted successfully as shown in above results. The thesaurus which bounded the information extraction scheme can be reusable repeatedly (http://www.tcmet.com.tw/vsm/idf.asp). Information given by the results from inverted documentary weighting scores as shown in Tables 2 may further be applied for building TCM knowledge database through the method of relevance feedback which was well developed in text retrieval[16]. Each keyword in the table was indexed and gave some functions to be indicative or informative of the text's content. Indexing descriptions like this, which had been installed on the TCMET website, can further be browsed and guided the user to read the full-text of the document[1-5].

Thesaurus class terms which were constructed based on the concept of PMRM played an important role in this study. A thesaurus provides a grouping or classifying of terms which used in a given topic area into classes known as thesaurus classes[17-20]. On building the herbal medicine thesauri consisting of 229 drug names, TCMET project was performed to extract recipes as well as formula name, application, and clinical usage effectively in the TCML *Heijian-Liu-Shu*[2]. Actually, we may consider that those 229 herbal names with different combinations leading to the TCM recipes with a unique formula name. Therefore, the table of herbal name thesaurus may guide us to collecting treatment prescriptions in TCMLs. The principle thesaurus, which was extracted from *Huang-Di-Nei-Jing*, was found that the 4-gram phase (4 adjacent Chinese characters together) was the most cost-effectiveness thesaurus to extract the concept of principle of TCM[3]. Surprisingly, most of the ancient TCML author gave respect to *Huang-Di-Nei-Jing* as the most important TCM theoretic resources from dynasty to dynasty. Eventually, 4-gram *Huang-Di-Nei-Jing* phase was applied to extract TCM theories from TCML. Name of disease (Diagnosis) and method of treatment are two thesauri which were exacted from the electronic TCML of *Lei-Zheng-Zhi-Cai*[4] giving powerful function to filter out recipe and formula. As shown in our results, the document ranking program really could filter out the PMRM information automatically so that it may further apply to internet TCML online searching with ranking scores. Being a TCM doctor, I hereby clarify that there is a big difference between Orthodox Western medicine and TCM in the way of building thesauri electronically. TCM has had the intuition of medicine, human body, universe, herbal medicine, as a whole[22]. This kind of universe-set medicinal concept leads us to build the PMRM Information Extraction more efficiently. This project really got the advantages from those Traditional habits in classical Chinese Medicinal writing style. The correlated PMRM thesauri had already been well-congregated since the TCML were written down by our ancestors. TCM retrieval is really functioning cost-effectively and helps binding TCM knowledge as whole. TCM doctor may find its usefulness in locating the texts with rankings as well as the PMRM concepts in a whole picture. Hopefully, we will implant the retrieval function of this study on the internet next year so that everyone can use it.

This study applies $tf \times idf$ functions to the TCMET project which integrates TCML knowledge-based thesauri and finally generated the weighting term and document table with 89,462 rows. Term frequencies and inverted document function has been widely applied to the area of information retrieval[10,17,20] but seldom applied to TCML. TCM becomes the mainstream of alternative medicine all over the

world[14] and TCML text retrieval on the internet is one of the best communication media to convey TCM knowledge publicly and globally. Since, there will be a lot of information retrieval requirements in nowadays computer era such as in the area of TCM evidence based medicine[19-22], ICD (international classification of diseases) for Chinese Medicine, Internet Robot for TCM and much more. Concurrently, the basic needs of a complete TCM electronic full-text database and its retrieval technology seem to be more urgent than ever before.

References

1. Chen YG. Word-wide-web dynamic database of traditional Chinese medicinal literature. *J. Chin. Med.*, 11:43-51, 2000.
2. Chen YG. Computer-aided retrieval of pharmacological information from the six books of Hejian. *J. Chin. Med.*, 12:1-10, 2001.
3. Chen YG. Development and application of Huang-Di-Nei-Jing concordance thesaurus. *J. Chin. Med.*, 13:49-57, 2002.
4. Chen YG. Apply SQL server 2000 XML schema to Lei-Zheng-Zhi-Cai database integration on the internet. *J. Chin. Med.*, 14:47-58, 2003.
5. Chen YG. Building an epidemic disease thesaurus retrieval from traditional Chinese medicinal literature. *J. Chin. Med.*, 15:39-46, 2004.
6. Losee RM. "Similarity and Retrieval Decisions," in *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer Academic Publishers, Boston, pp. 43-75, 1998.
7. Salton G. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Company, New York, pp. 43-75, 1989.
8. Luhn H. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Develop.*, 1:309-317, 1957.
9. Urzua CM. A simple and efficient test for Zipf's Law. *Econom. Lett.*, 66:257-260, 2000.
10. Moens MF. "The Selection of Natural Language Index Terms," in *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers, Boston, pp. 77-132, 2000.
11. Unschuld PU. "An Explanation of Why the Three Methods of Sweating, Purging, and Vomiting Should Suffice for the Treatment of Illness," in *Introductory Readings in Classical Chinese Medicine*. Kluwer Academic Publishers, Taiwan, pp. 212-217, 1989.
12. Owolabi O. Dictionary Organizations for Efficient Similarity Retrieval. *J. Syst. Software*, 34:127-132, 1996.
13. Wong CM, Wu LT. History of Chinese Medicine Book 1. The Tientsin Press, Tientsin, China, 1932.
14. Li JW. Integration of Chinese and Western Medicine and the Internationalization of Chinese Medicine. *J. Chin. Med.*, 15:137-150, 2004.
15. Hoffman RR. The problem of extracting the knowledge of experts from the perspective of experimental psychology. *AI Magazine*, 8:53-67, 1987.
16. Salton G, Buckley C. Improving retrieval performance by relevance feedback. *J. Am. Soc. Inform. Sci.*, 41:288-297, 1990.
17. Baeza-Yates R, Ribeiro-Neto B. "Text Operations," in *Modern Information Retrieval*. ACM Press, New York, pp. 163-189, 1999.
18. Cao C, Wong H, Sue Y. Knowledge modeling and

acquisition of traditional Chinese herbal drugs and formulae from text. *AI Medicine*, 32:3-13, 2004.

19. Liang SF, Siobhana S, Tait J. Investigating sentence weighting components for automatic summarization. *Info. Proces. Manage.*, 43:146-153, 2007.
20. Rokaya M, Atlam E, Fuketa M, Dorji TC, Aoe J. Ranking of field association terms using co-word analysis. *Info. Proces. Manage.*, 44:738-755, 2008.
21. Shea J. Applying evidence-based medicine to traditional Chinese medicine: debate and strategy. *J. Altern. Complement. Med.*, 12:255-263, 2006.
22. Schnorrenberger C. Epistemological evaluation of Chinese medicine and acupuncture – Part I. *J. Chin. Med.*, 22:1-17, 2011.

J Chin Med 22(3,4): 205-216, 2011

中醫典籍文獻理法方藥之資訊擷取

陳逸光

慈濟醫學中心中醫科，花蓮，台灣

（100年10月21日受理，100年11月15日接受刊載）

電子化文本具備簡單資訊擷取功能，已是醫學文獻應用在臨床之新趨勢（如實證醫學）。古代醫家留下大量中醫古籍文獻直到現今，如果我們能夠給予這些書籍電子化並具備簡單資訊擷取功能，將是花費少作用大的考量。在本研究中，中醫藥典籍文獻之原始電子文件來源，將取自於從 1997 年已開始建立之 TCMET 資料庫，類別辭典庫由中醫理、法、方、藥的觀念建構而成。將中醫藥典籍文本及關鍵詞之索引及權重，經項目頻率 (*tf*) 及反轉文件頻率 (*idf*) 計算後製成表格。一個共有 89,462 個關鍵詞的理法方藥詞庫含頻率比重，已放置在 TCMET 網際網路上供參考及閱覽。在本文的最後，我們以 4 個關鍵詞為作文獻查詢及排序進行測試，藉文獻反轉檔以計算出關鍵詞之比重，從文獻擷取排序之結果發現，詞頻權重的確能協助中醫理法方藥概念之篩選。本研究的另一個目的，是為了建立一個先導計畫平台，並將過去數年間之中醫辭典與文獻全文整合起來，可期望作為未來數年間中醫藥典籍文本擷取技術提昇之基礎。

關鍵詞：中醫藥典籍、資訊擷取、索引詞典、*tf* × *idf*

*聯絡人：陳逸光，慈濟醫學中心中醫科，花蓮市中央路三段707號，電話：03-8561825分機2158，電子郵件信箱：tcmet.editor@msa.hinet.net