

# 黃帝內經貫珠詞典庫之研發與應用

陳逸光

慈濟大學 醫學系

花蓮

(2001年11月23日受理，2001年12月27日收稿，2001年12月31日接受刊載)

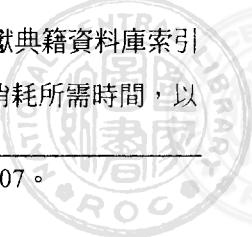
中醫經典黃帝內經之原文資料，經常被歷代醫學家引用到醫藥論著中，以闡述學術思想或醫學理論。本研究選取中醫經典資料庫 NW2001，其內容包括：黃帝內經、金元四大家著作、景岳全書及張氏醫通等中醫經典約四百萬字。所有 NW2001 典籍都交由一個文字剖析器，以 Brute Force Algorithm 切割成 N-連字串 (N-Gram) 詞典庫。結果發現 4-Gram 是最佳化的庫詞典在 NW2001 的精確率 (Precision) 及回收率 (Recall) 分別為 0.96 及 0.86；在中醫典籍網際網路貫珠集應用上效率很高，資料庫經過索引，網頁顯示的空等時間 (Downtime) 能控制在 8 秒內。由於內經詞句的專一性 (Specificity) 相當高，所以被萃取出來的內經 4-Gram 詞典知識庫，便可經由反轉檔 (Inverted File) 索引到 NW2001 資料庫有關中醫學術理論之出處。

**關鍵詞：**中醫藥典籍文獻，詞典庫，Brute Force Algorithm，N-連字串，網際網路。

## 前　　言

中國醫藥自秦漢以降二千餘年，黃帝內經被歷代醫家尊崇為中醫基礎理論經典著作<sup>1</sup>。每當一個新的中醫學論點產生時，醫家經常會引用黃帝內經條文作為立論之依據。數千年來中國醫學本源出於內經一脈相承的特色，促使本研究以文獻萃取 (Documentation Extraction) 技術<sup>2</sup>，建立自動化中醫基礎理論貫聯知識庫之研發，預期可以開發出一套經濟實用的中醫理論詞典庫，貫串在中醫典籍文獻 NW2001 資料庫 (黃帝內經、金元四大家著作、景岳全書及張氏醫通) 共約四百萬字，研究發現中醫詞典知識庫有助於專門術語中醫典籍索引之應用。

時間與空間從資訊觀點看來是一種交易，過去為了節省昂貴的儲存空間，經常要以壓縮方式來存放中醫研究資料<sup>3</sup>。相較於十年前，現今發展中醫藥典籍已經不必擔心記憶體不足，電腦運算速度也相當快速。在網際網路中醫典籍索引的研發過程中，資料搜尋必定有最佳化效應存在；換言之，對固定的電腦運算速度而言，文獻搜尋的邏輯 (Algorithm) 將決定最佳化效應，尤其在網際網路上建置中醫文獻典籍資料庫索引程序時，必須考量中醫文獻擁有龐大的資料量。文獻關聯索引量及程序直接影響到運算消耗所需時間，以



一個 56K Modem 為基準，在 TCMET<sup>4</sup>網站上每一個資料庫網頁顯示所需要之時間，皆控制在 8 秒內。若要使 NW2001 四百萬字資料，能在 TCMET 8 秒內找到內經相關資料必須有很好的索引規劃。

## 材料及方法

### 一、中醫典籍資料庫及網際網路開發環境

本研究的電腦軟硬體開發環境在 TCMET 下進行<sup>4</sup>，資料庫管理（DBMS）及程式設計（Programming）皆在 Visual Foxpro 5.0 下離線（Off Line）開發，所有經過 VF 程序篩選整理後之網際網路中醫藥典籍資料庫則存放在 Microsoft SQL Server 中供線上（On Line）使用，資料庫網際網路顯示文本（Script）為 ActiveServer Page，TCMET 現今是架設在 ADSL 之 World-Wide-Web 通訊架構上。

### 二、中醫典籍詞典庫開發

#### 2.1. 中醫藥古籍詞典知識庫建置架構（圖 1）

#### 2.2. N-Gram 文字剖析器(Parser)

資訊萃取（Information Extraction）是資訊擷取（Information Retrieval）的一個重要環節<sup>2</sup>，本研究對所有 NW2001 資料庫內之中醫典籍文獻進行字串萃取，以建立黃帝內經貫珠集索引知識庫，研究中的詞彙分割器是以 Brute Force Algorithm<sup>5</sup>進行（見圖 2），設 x 句子（String）與 y (i,i+m-1) 個子句（Substring）比較，n, m 分別是 x, y 的字數，而 i 是 x 句子的連續位置 ( $1 \leq i \leq n-m+1$ )，一直切割至句尾結束為止，切割的效率以 Big-O 運算為  $O(m+n)$ 。本研究曾對 2-Gram 到 7-Gram 詞彙（一組 2 字相鄰的詞彙稱 2-Gram, 3 字為 3-Gram, ...)，子句詞典分別記錄在索引表格中以建立反轉檔（Inverted File）<sup>6,7</sup>。

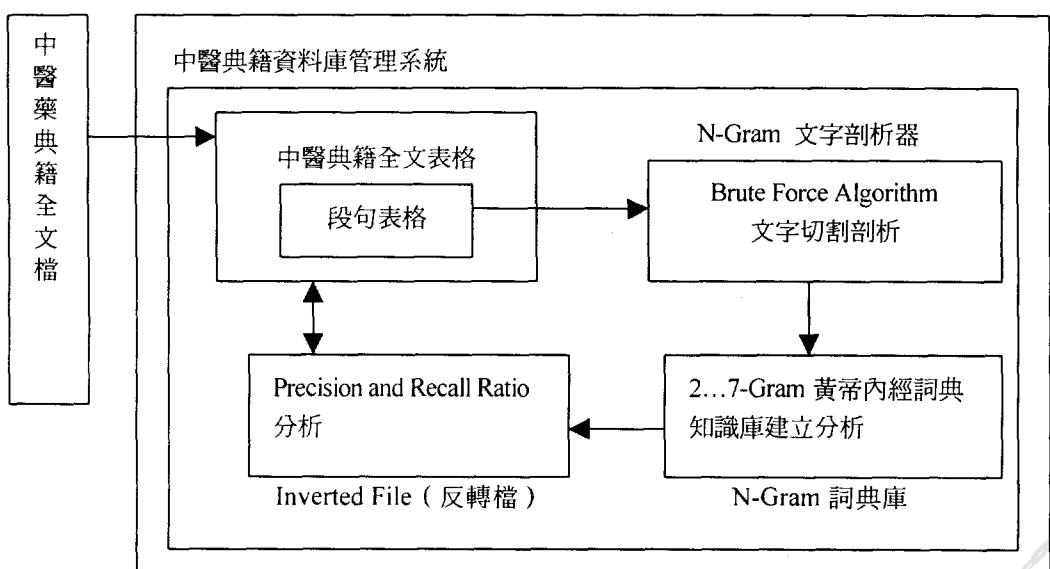
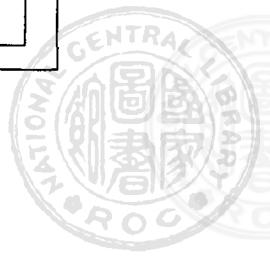


圖 1 中醫藥古籍詞典知識庫系統



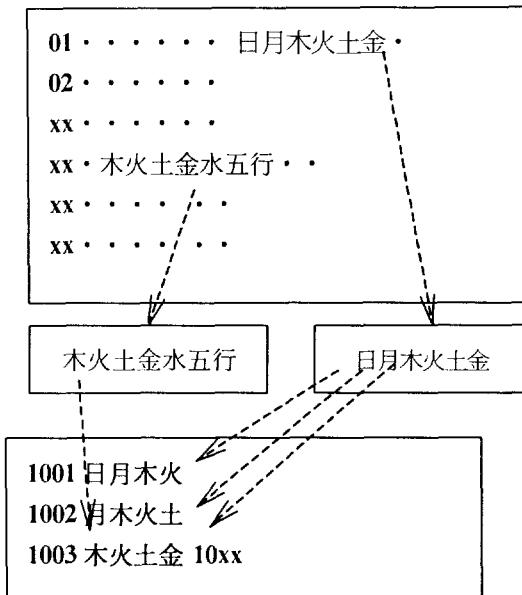


圖 2 Brute Force Algorithm 製作 4-連字串 (4-Gram) 反轉檔

### 2.3 反轉文件頻率 IDF(Inverted Documentation Frequency)

IDF 應用在文件檢索分析上已有數十年歷史<sup>8</sup>，在 2.2 節所述之詞典索引反轉檔，其內容分別記載著每個詞彙之出處、書名、頁數、行號等資料。因此，詞典索引反轉檔可以快速計算出文獻之出現頻率及相對的原文所在位置，經由 Precision (精確率) 及 Recall (回收率) 的運算，便可以藉由 IDF 在 Precision 及 Recall Ratio Curve<sup>9</sup> 中進一步了解 N-Gram 字串的文獻檢索效率。

## 三、中醫典籍文獻網際網路開發

### 3.1 詞典庫選取

從 2.1 節中可以看到中醫詞典知識資料庫的結構及建置過程，由於我們考量網際網路的傳輸速度及容易閱覽等因素，N-Gram 詞典庫資料必然要取捨，通過 2.3 節的分析後，最佳效率的 4-Gram 詞典庫送到 TCMET 之 MSSQL Server 中供網際網路 ASP 網頁開發使用，TCMET 資料庫架構見拙著論文<sup>4</sup>。

### 3.2 網頁設計

TCMET 在 ADSL WWW 通訊架構上頻寬有限，為維護網路通訊品質，網友空等時間 (Downtimes) 控制在 8 秒內，每一頁資料在 10K 以下（約 10 頁 A4 中文）。文獻貫珠集，以黃帝內經每一篇（素問及靈樞各 81 篇）為最基本的顯示頁，每一行為一個切入點 (Entry Point)，網友只要用滑鼠單擊切入點，該行的詞典庫索引表便在一個新網頁中供按選，直到網友檢索出詞典貫珠集中的相關 NW2001 典籍文獻為止。



# 結 果

## 一、詞典庫最佳化選擇

本研究對 2-Gram 到 7-Gram 詞彙分別作 Precision-Recall Ratio 比較（見圖 3）。曲線上右第 1 個黑點為 2-Gram 其 Recall 值高但 Precision 值低，可解釋為相關內經字串從 NW2001 中醫古籍文獻庫被檢索出來，被遺漏的很少；但有很多檢索資料卻並非屬於內經條文所有。反之，曲線上左第 1 個黑點為 7-Gram，被檢索出來的資料百分之九十八都是內經原文所出之處，但卻遺漏了許多未被檢索出來的資料。4-Gram 的 Precision 為 0.96；Recall 為 0.86，檢索效率最佳，本研究採用四字詞典庫作為網際網路開發的資料庫。

## 二、4-Gram 詞典庫文獻擷取

本研究對 NW2001 資料庫(黃帝內經、金元四大家著作、景岳全書及張氏醫通)共約四百字作了一次全面的 4-Gram 剖析，以 Brute-Force（圖 2）對所有 NW2001 作 4-Gram 詞典庫之建置。全部 NW2001 共收集到四字串詞句一百一十餘萬個（以內經句首為例：“上古天眞論第一”→“上古天眞”，“古天眞論”，“天眞論第”，“眞論第一”）；屬內經的四字串詞句共五萬四千餘個（見表 1）。

## 三、網際網路執行情況

以 56 Kmodem 為基準每頁顯示都可在 8 秒內完成，每一頁面的範例（見附件），共 5 個步驟網友便可獲得全文資料。

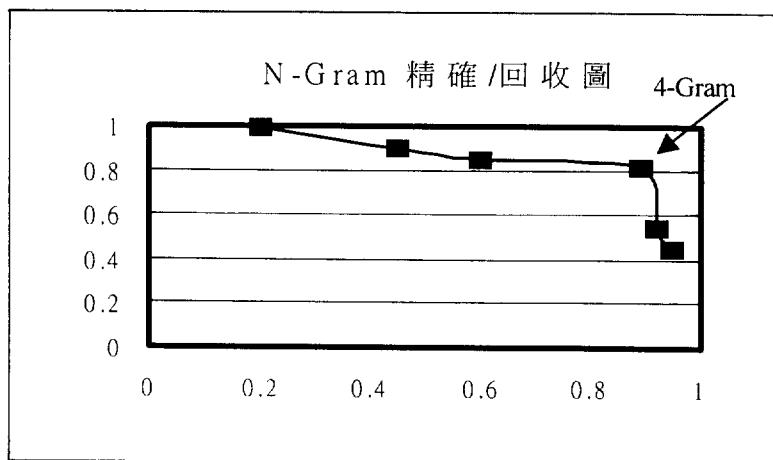
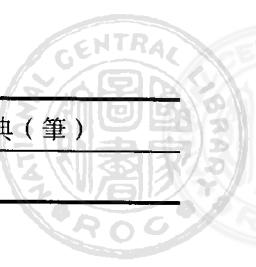


圖 3 內經 N-Gram 關鍵詞之精確/回收曲線說明：y 軸 Precision(精確率) X 軸 Recall (回收率) 曲線上黑點由右至左分別代表 2 到 7-Gram (字串)

表 1 內經及所有 NW2001 四字串詞典庫筆數

所有 NW2001 4-Gram 詞典 (筆)	文獻中屬內經 4-Gram 詞典 (筆)
1,141,259	54,951



## 討 論

本研究應用 Brute Force Algorithm (BFA) 設計了一個簡單的詞典剖析器，BFA 的操作過程以每一獨立句子為單位，按照設計程序把句子毫無保留的分解（不會遺漏一個字），這種剖析方法雖然有點慢，但在中醫藥典籍一字不漏的分割正是我們所需要的，一本黃帝內經在 Foxpro 下作四字串的切割，耗時約半個小時，由於切割資料是離線（Off Line）作業，而且只需要作一次就能永久存檔備用。本研究之剖析以句為單位，一個句字若果只有五個字，卻使用 7-Gram（七字串）去切割必定為零，自 N-Gram（見圖 3）的分析數據，便可得知 7-Gram 的 Precision（精確率）很高，而 Recall（回收率）卻很低。另外一種精確率高，回收率低的原因是古書標點符號不統一，黃帝內經及後世醫學論著所引用之條文，很多是由近代學者重新加上段句標號，但從本研究發現以內經詞典輔助擷取後世文獻資訊之成果豐碩。由於現代資訊工具之便利，我們使用電腦輔助剖析文句從 2-Gram 到 7-Gram，只要輸入合適的參數便可完成目標，因著電腦快速重複運算功能使研究分析更為精準。從圖 3 之曲線圖表數據分析得出 4-Gram 最佳檢索效率，使網際網路資料運用效益與決策之信任度更為提升。

大量中國醫藥典籍文獻在坊間便可以購得，如行政院衛生署中醫藥委員會出版的“中醫藥典籍檢索光碟”<sup>10</sup>，其內包含各朝代中醫藥典籍文獻四十餘冊，如果我們能夠以現代資訊處理技術輔助整理典籍文獻，必定對中醫文獻分類及臨床研究提供很重要的資訊。能輔助中醫文獻資訊擷取之方法相當多，本研究針對中醫詞庫知識庫建造為主題，以 BFA 作為文字剖析工具，只是資訊擷取技術冰山之一角。TCMET 的設計理念是以建立中醫文獻知識庫為導向，例如去年我們以中醫方劑自動整理分類為目標，整理出劉河間 512 方<sup>9</sup>；本研究運用黃帝內經字串貫聯歷代醫籍的特色，建造了一個內經詞典庫。以上二例說明了中國醫藥文獻索引自成體系，在建立文獻索引資料庫過程中，必須思考中醫專家及讀者們最希望得到的資訊，再規劃何者應優先發展，研究目標訂立後，便可以開始選擇適合的課題，並在中醫典籍知識範圍內，找出能進行自動整理分類的可行文獻擷取方法；畢竟人工整理典籍文獻是十分昂貴的事，倒不如先讓電腦輔助整理或分析資料，最後再交由人工做最後定案，文獻整理成本必定減少許多。TCMET 是一個在網際網路上開發的中醫典籍文獻閱覽系統，為了不要讓網友久等，線上資料庫都是經過嚴格規劃整理。本研究中，所使用的資訊擷取技術都不是最新但都很實用，能幫助我們完成“電腦輔助，人工定稿”的文獻網路化處理策略，許多文獻擷取科研課題尚待開發。

筆者使用 Foxpro 作為資料庫開發環境已有十年的時間，最早在 DOS 系統下的 FoxBase 後來被微軟收購更名為 Foxpro，Visual Foxpro 2.6、3.0、5.0、6.0 到最新的 7.0 版，其實自 5.0 版以後功能都差不多，而且 DataBase 檔案都通用。幸慶的是早期開發的 TCMET 中醫典籍檔案到現在都可以使用；反觀在當時相當出名的資料庫軟體，如 dBaseV，Clipper 等，早已不見蹤影被淘汰了。Visual Foxpro 速度快且容易使用，適合在個人電腦上作業，程式語言設計像 Basic 語言般簡單，所以 TCMET 的資料擷取、文字剖析（見圖 1）及反轉檔（IDF）都是在 Foxpro 進行，最後經過一連串的繁複運算後所產生的表格，可直接使用 Foxpro 送到 MSSQL 及 Oracle 等著名資料庫伺服器供使用。網際網路資料庫是多人多工的作業環境，Foxpro 在多工環境下表現（Performance）就比上述的資料庫伺服器遜色多了，因此 TCMET 最後還是會選用 MSSQL 伺服

器來管理網際網路資料庫，而 Active Server Page ( ASP ) 是一種伺服器端的網際網路程序控制網頁，ASP 透過 ActiveX 技術可以有效開發網際網路資料庫元件，新一代的微軟 ASP 及 XML 技術將是明年 TCMET 發展的一個新課題，相信新的工具會使中醫藥古籍文獻閱覽介面更為靈活。網際網路上雖然有很多搜尋引擎，如龍捲風、慈濟世界、中時新聞線上索引等不同類型資料庫檢索網站，可謂百花齊放，熱鬧非凡。TCMET 在專業的中醫藥文獻資料庫以中醫傳統理、法、方、藥為發展模型，會有一定的競爭優勢，但我們仍然會把重點放在資訊擷取 ( Information Retrieval ) 及網際網路系統開發軟體技術上創新。研發過程中，TCMET 網際網路資料庫系統開發技術是運用微軟推崇的 ActiveX 元件，筆者曾分別在 ORACLE 及 MSSQL 下開發網際網路資料庫元件。筆者從不成熟的經驗中發現，在 NT 網際網路環境下，ActiveX 會比 JAVA 的 JDBC 資料庫元件穩定，畢竟 NT 及 ActiveX 都是微軟的產品。資料分析可提供系統效率相關研究之科學依據，本研究的詞典剖析器 ( Parser ) 見圖 1，是以資料庫形式建造，切割資料經過 N-Gram 的分析以決定四字串使用之依據。

研發電子文獻擷取網際網路系統，除了電腦系統、資料庫及資訊擷取方法等技術外，最重要是要了解中醫知識庫的結構建置與運用。讀中醫書必定會發現當作者對談論醫學基礎觀念時，經常會引經據典，黃帝內經為中醫基礎理論經典著作，被引用的頻率必然是最多，事實亦是如此。黃帝內經四字串散見在 NW2001 文獻庫中約有五萬個，選擇內經作為中醫基礎理論詞典庫是相當正確的，因為內經用字相當嚴謹，言簡而義精。例如內經第一章第 7 行中有“終其天年”四字，句子看似常見，但在歷代中醫典籍文獻庫 NW2001 中共出現四次，而四次都是內經相關內容，因此四字串詞典庫的精確率及回收率都在 85% 以上，可見內經詞典庫貫珠集對歷代中醫著作有相當專一 ( Specific ) 的結論相當正確。因此，網友在 TCMET 網站上很容易透過閱覽黃帝內經，更深一層從 NW2001 文獻庫中找到中醫理論之相關資料，而且所有資料都整過索引整理，在網站上容易閱覽（見附件）。

## 結論

中醫文獻擷取是一項龐大工程，本研究完成了中國醫藥基礎理論醫詞典庫之建立模式，中醫沿革已久的“理、法、方、藥”學習模式，是研究中醫理論及臨床的一套法則。處方的篩選或藥物的臨床應用，都是可以在文獻擷取技術上繼續開發；Information Retrieval ( 資訊擷取 ) 開發人員，必須有足夠的中醫知識及市場需求判斷能力，配合現實環境，當使用者要求資料 ( Data ) 或資訊 ( Information ) 的時候，電腦程序便能適時提供答案或引導尋求答案。中醫文獻資訊擷取是一個真實的資訊科技領域，九十一學年慈濟大學醫學資訊 ( Medical Informatics ) 系要招生了，當人們問及何謂“醫學資訊”實在不是“醫學”加“資訊”，“文獻資訊擷取”也不等於“文獻”加“資訊擷取”的道理相同。本研究利用黃帝內經本身就是一個中醫基礎理論知識庫的特性，來進行資訊萃取，得到很高的擷取率；新科技需要更多人參予才能永續發展，並需要更多學者專家共同為明天之研究努力。



## 參考文獻

1. 李煥燊，中西病原學與病理學比觀，醫林小草，台北，國立中國醫藥研究所出版，pp. 153-182，1974。
2. Gaizauskas R, Wilks Y. Information Extraction: Beyond Document Retrieval Int J of Computational Linguistics & Chinese Language Processing 3:17-59,1998.
3. 陳逸光、汪叔游、傅式恩、陳太義、劉森，以微電腦分析中醫脈波圖形之軟體設計—並以肝炎為例，中國醫藥學院中國醫學研究所碩士論文，台中，1987。
4. 陳逸光，網際網路中醫藥典籍文獻動態資料庫研究，中醫藥雜誌 11：43-50，2000。
5. Ricardo BY, Berthier RN, Indexing and Searching, In: Modern Information Retrieval, Addison-Wesley, ACM Press New York, pp. 209-228, 1999.
6. Broglio, J, Callan J, Croft WB, Technical issues in building an information retrieval system for Chinese. CIIR Technical Report IR-86, University of Massachusetts, Amherst, 1996.
7. 曾元顯、林瑜一，模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估，中國圖書館學會會報，61:103-125,1998。
8. Salton G, Buckley C, Approaches to Text Retrieval for Structured Documents , Technical Report 90-1083, Dept. of Computer Science, Cornell University, 1990.
9. 陳逸光，電腦輔助擷取河間六書方劑專論資料，中醫藥雜誌，12：1-9，2001。
10. 行政院衛生署中醫藥委員會，中醫藥典籍檢索光碟，台北，行政院衛生署中醫藥委員會出版，1999。



## 附件 NW2001 範例說明

([www.tcmet.com.tw/nw2001/huagtemplate.htm](http://www.tcmet.com.tw/nw2001/huagtemplate.htm))

第一步：點選表中“10001”或“10002”便可讀出該書卷全文

代號	出處
10001	(素)上古天眞論篇第一
10002	(素)四氣調神大論篇第二

第二步：點選表中“10001”讀出四字詞典庫

代號	出處
10001	上古天眞論篇第一

第三步：點選表中“10001”讀出“上古天眞”四字詞典庫

代號	所尋字串	代號	所尋字串
10001	上古天眞(9筆)	10003	天眞論篇(1筆)
10002	古天眞論(9筆)	10004	眞論篇第(1筆)

第四步：點選表中“10002”讀出“張從政著濕熱門”一整頁資料

代號	字串	出處
10001	上古天眞	(素)上古天眞論篇第一
10002	上古天眞	(張)儒門事親卷十一金_考城張從政著濕熱門。

第五步：讀出“張從政著濕熱門”這裏由於篇幅所限僅提供一行資料（原 WWW 網頁上有 30 行）

行號	文章內容	所尋 Keyword : {上古天眞}
第 9 行	此法宜分陰陽，利水道，乃為法治之妙也。上古天眞論云：一陰一陽之謂道。	



# DEVELOPMENT AND APPLICATION OF HUANG-DI-NEI-JING CONCORDANCE THESAURUS

Yee-Guang Chen

*Department of Medicine, Tzu Chi University,  
Hualien, Taiwan*

(Received 23<sup>th</sup> November 2001, revised Ms received 27<sup>th</sup> December 2001, accepted 31<sup>th</sup> December 2001)

*Huang-Di-Nei-Jing*, one of the Classical Traditional Chinese Medicinal Literature (CTCML), was frequently referenced by other CTML authors to write their medical principles or ideas. In this study, a CTCML database NW2001 including: *Huang-Di-Nei-Jing*, *Jin-Yuan-Si-Da-Jia*, *Jing-Yue-Quan-Shu*, *Cheng-Si-Yi-Tong* totally around four million Chinese characters was used. A word parser, developed by a computer program, can separate words into N-Grams thesaurus, which was driven by the Brute Force Algorithm. Finally, we found that 4-Grams thesaurus indexed on the NW2001 database had Precision and Recall of 0.96 and 0.86 respectively. This optimized 4-Gram thesaurus was used to develop the CTCML world-wide-web. The database was indexed and well formed that the downtime has been controlled under 8 seconds to show each homepage on the WWW. The 4-Grams data set, which extracted from *Huang-Di-Nei-Jing*, formed a knowledge-based database. By using inverted file, the 4-Gram thesaurus of *Huang-Di-Nei-Jing* could be mapped into the NW2001 database. The mapped information had a high specificity relating to medical principles of CTCML.

**Key words:** Traditional Chinese medicinal literature, Thesaurus, Brute Force Algorithm, N-Grams, WWW.

---

**Correspondence to:** Yee-Guang Chen, School of Medicine, Tzu Chi University, 701, Section 3, Chung Yang Road, Hualien, Taiwan. Tel: (03)8565301 ext. 7207.

