南　　華　　大　　學

資訊管理學系

碩士論文

應用文字探勘技術於疾病問答系統之研究

A Study on a Disease Consultation System Using Text Mining Techniques

研 究 生：蔡 慶 鐘

指導教授：吳 光 閔

中華民國 101 年 6 月

南　華　大　學

資訊管理學系碩士班

碩 士 學 位 論 文

應用文字探勘技術於疾病問答系統之研究

A Study on a Disease Consultation System Using Text Mining Techniques

研究生：蔡慶鐘

經考試合格特此證明

口試委員：

指導教授：吳光閔

系主任(所長)：資訊管理學系 系主任 吳光閔

口試日期：中華民國　101 年　05 月　30 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：蔡慶鎮 之碩士畢業論文

中文題目：

應用文字探勘技術於疾病問答系統之研究

英文題目：

A Study on a Disease Consultation System Using Text Mining Techniques

指導教授：資訊管理學系 系主任吳光閃 博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

■ 共同享有著作權

□ 共同享有著作權，學生願「拋棄」著作財產權

□ 學生獨自享有著作財產權

學　　生：蔡慶鎮 （請親自簽名）

指導老師：吳光閃 （請親自簽名）

中　華　民　國　101　年　5　月　30　日

南 華 大 學 碩 士 班 研 究 生

論 文 指 導 教 授 推 薦 函

資訊管理 系碩士班 蔡 慶 鐘 君所提之論文

應用文字探勘技術於疾病問答系統之研究

係由本人指導撰述，同意提付審查。

指導教授 吳光陽

101 年 5 月 30 日

誌　　謝

又到了鳳凰花開的日子，自從大學畢業踏進職場也將近十年，想想當初五專的時候只要能畢業就好，從來也沒想過能繼續進修到研究所，一直到家人及朋友的鼓勵下，決定試試看，而最後也能如期所願。

回想三年前，雖然第一次落榜，但仍澆不熄求知、求學的渴望，就在第二次的時候順利過關，而也開啟了研究之路，這兩年當中，都能盡量把握每堂課的精華而所學習，不過自己也發現好像比以前求學的時候更認真努力，也或許是有某種程度的壓力吧，正所謂人要在壓力下成長，確實是如此。

在這段期間，非常感謝班導師及指導教授以及每位老師的不吝指導與教誨，以及很高興能認識班上的所有同學，因為有了大家的互動及鼓勵，所以才讓全班一起感染了讀書氣息，更也激起了打拼的精神，除此之外，更要感謝家人的體諒及包容，同時也要感謝我的好朋友、好同學及好同事，在各方面的贊助、鼓勵及支持。

時間過得很快，也許這只是人生的一小段時間，但我相信這是很精彩的一段，畢業後各奔前程，而能將所學貢獻於社會並造就下一個未來，或許大家的相遇不知何時，但不會忘記「那些年，我們一起讀的南華」。

應用文字探勘技術於疾病問答系統之研究

學生：蔡慶鐘　　指導教授：吳光閔

南　華　大　學　資訊管理學系碩士班

摘　　要

資訊科技的進步，許多科技的文明病因此而增加，而一般在就診過程中，也因為人數眾多問診的時間卻被壓縮得更少，且得到的病情狀況卻往往不如期望。

由於網際網路的發達，搜尋引擎的進步，通常只要把大略知道的病因輸入即可列舉出許多相關的病情，但大部分都是模糊而無法找出相似的病情，許多相關的醫療機構都會提供給病患所謂的衛教資訊，但這些資訊都是透過預先制定及條列式的病況需求，也因為這樣的機制讓問診者無法有效率獲知相關資訊。

本研究藉由從線上網路醫院之醫病回覆內容中收集大量資料，應用文字探勘技術找出非結構化的回覆內容並利用中央研究院提供的中文斷

詞系統，將有用的資訊回傳並歸類成相關詞彙。

因此本研究將建構一套疾病問答系統，將問診者所提出的問題分析、比對相關詞彙再依權重比例挑選出相似度較高的答案，來幫助問診者更有效率地進行疾病問答，更可讓問診者在就醫前的參考值。

關鍵字：文字探勘、中文斷詞、問答系統、相似度

A Study on a Disease Consultation System

Using Text Mining Techniques

Student：Ching-Chung Tsai　　Advisors：Dr. Guang-Ming Wu

Department of Information Management
The Graduated Program
Nan-Hua University

Abstract

With the advancement of information technology, many diseases of civilization have appeared and increased the demand for outpatient services. As a result, patients are allowed less time for consulting physicians and usually cannot obtain as much information as expected.

Due to the advent of the Internet and fast development of online search engines, it is now easy for patients to search for sick conditions and causes by symptoms. However, the search results are not entirely accurate. Many medical institutions will also provide health information to patients, but such health information is usually arranged based on a predefined list of sick conditions and cannot effectively help patients obtain necessary disease information.

This study collected a large amount of information from replies of an online hospital and used text mining technique to find unstructured content of the replies. Later, this study applied the Chinese Word Segmentation System provided by Academia Sinica to convert useful information into phrases.

The objective of this study was to build a Disease Question Answering System that can analyze questions brought up by patients and find answers to the most similar questions by weight from the database to help patients obtain necessary information efficiently and use it as a reference before seeking outpatient services.

Keywords: Text Mining, Chinese Word Segmentation, Question Answering System, Similarity

目　錄

表目錄

圖目錄

第一章 緒論

第一節 研究背景

網路發展迅速，搜尋引擎的進步，正帶領著人們對於知識追求的動力，然而現今搜尋引擎所提供的資訊往往都不是使用者所需要的，而是以未經過整理且搜尋條件方面需要以專有名詞或關鍵字來讓準確率提高，因此在獲得知識前，都需要尋找許多相關主題才可能得知所需資訊，而資訊檢索技術也因現況的局限下蓬勃發展，所扮演的角色愈形重要。

以資訊檢索為基礎的系統及相關研究也慢慢受到重視，例如問答系統，由於問答系統與搜尋引擎最大的不同在於前者將大量的文件分析後進行比對並找出最接近的答案，而後者則將所查詢的資訊且未經過處理並以模糊化的方式呈現，對查詢者來說則需要以更多時間瀏覽。

醫療品質及科技的進步，透過搜尋引擎，人們逐漸提高對於醫療保健的認識。闕瑞紋（2001）研究發現，網路族群著重上網尋找健康資訊，以醫療保健網站為主要的健康知識取得途徑佔有 58.9%，如果涵蓋以其為次要的途徑，則醫療保健網站成為取得健康資訊重要途徑佔 84.1%。

因此，許多醫療機構紛紛推出民眾與醫師的互動網站平台，除可以提供醫療資訊及諮詢外，更可以增進醫病關係。

第二節 研究動機

國內許多醫療機構皆有提供醫療資訊及諮詢的互動平台，但往往所提供的資訊都是比較屬於制式性的問答或選項，民眾只能依照網站所提供的方向而選擇適合自己的問題。

因此，本研究將建構一套以文字探勘技術為基礎的問答系統，期望藉由文字探勘的相關技術，以大量的醫療諮詢問答紀錄進行分析及分類，再以關鍵詞彙的權重進行排名，以利挑選出最接近的答案並呈現查詢結果。

第三節 研究目的

本研究主要以文件的關鍵詞的擷取及文件的相似度比對，而查詢者僅輸入簡短的問句即可被系統以資訊檢索的技術分析並將計算過後的詞彙權重作為排序的依據，以將最接近結果的關鍵字呈現。

第四節 研究限制

一、本研究僅針對資料來源所收集的醫療諮詢問答紀錄進行分析。

二、本研究僅探討以文字探勘及資訊檢索等技術建構問答系統。

三、本系統僅於斷詞後的關鍵詞與醫療專有名詞比對，以提高準確度。

第二章 文獻探討

第一節 知識探勘

知識探勘的步驟大致分為：資料蒐集、資料清理、資料轉換、探勘技術運用、結果呈現與解讀。知識探勘可分為資料探勘（Data Mining）與文字探勘（Text Mining)。前者處理結構化資料，即每筆資料有共同欄位可記錄於資料庫者，而後者處理非結構化資料，即每筆資料沒有共通的結構性可言，經常為長短不一、記載訊息的自由文字，也是本研究所採用的探勘方法。

壹、資料探勘

資料探勘是一大量自動化的過程，其運用統計分析從大量資料庫中挖掘出潛在、非顯然的、未知的、潛在的「可能」有用資訊之過程(Frawley et al.,1991)。而 Grupe & Owrang (1995)學者則認為資料探勘是指從已經存在的資料庫當中挖掘出專家仍未知的新事實。 Fayyad(1996)則定義知識發掘(knowledge discovery)為從大量資料中 選取合適的資料，進行資料處理、轉換等工作，再進行資料探勘與結 果評估的一系列過程，也就是說資料探勘只是知識發掘過程當中的一個步驟。定義資料探勘為使用自動或半自動的方法，對大量資料作分 析，找出有意義的關係或法

則（Berry & Linoff，1997)。Akaka(2004）提出資料探勘是一種應用資料庫的技術，像是統計分析與建立模型， 用以發現資料中隱藏的模式與隱約的關係，並進行推論以預測未來結果。AAAI(American Association for Artificial Intelligence，2006）近期指出資料探勘是一種很強大的人工智慧工具，它可以自資料庫中 發現有用的資訊，並
可用來改善行為。

資料探勘我們可解釋為資料庫之知識發掘（Knowledge Discovery in Databases，簡稱 KDD)。也就是說可以從一個大型資料庫裡頭所 儲存的大量資料當中萃取出有趣知識，這個大型資料庫有可能是線上作業的資料庫，也有可能是資料倉儲。

一、資料探勘架構

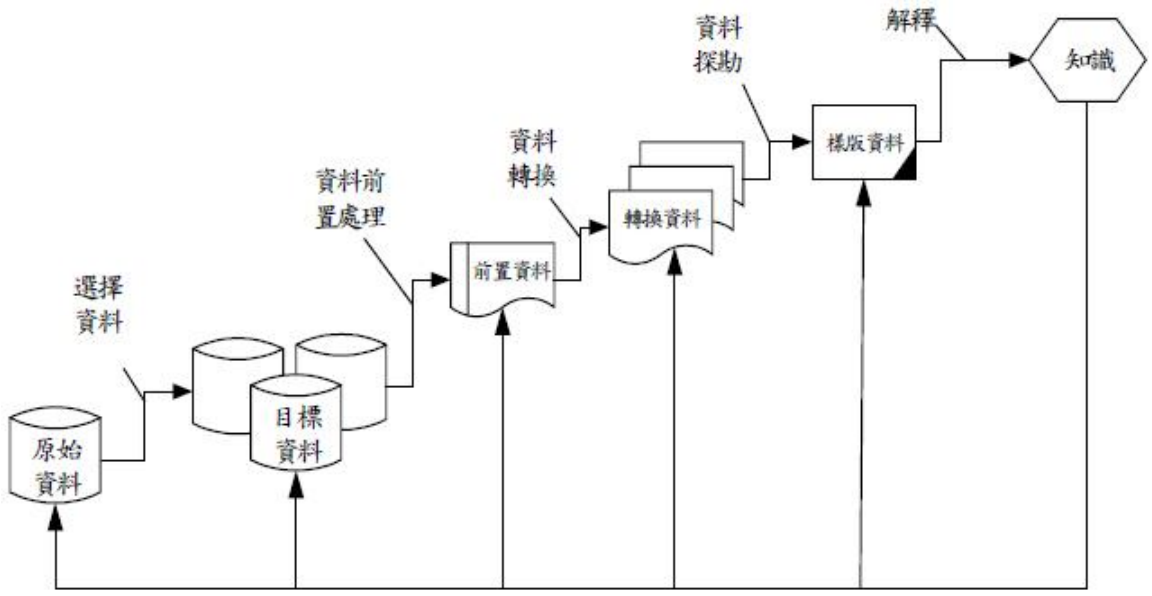

圖 2-1 資料探勘架構
（資料來源：Piatetsky-Shapiroet al.,1991;Fayyad et,al.,1996；姚吉峰，2002）

1.資料選擇 (Selection)

瞭解該領域的知識，挑選與分析工作相關的資料，用以建立目標資料集，在資料探勘的過程中專注於選擇的資料子集合。

2.前置處理 (Preprocessing)

資料集中的資料會包含錯誤、遺失及不完整的資料內容，必須將其去除，如此一來才能夠排除干擾和不一致資料的影響，並將格式不同的資料進行處理，使其具備一致化。

3.資料轉換 (Transformation)

進行資料的簡化及轉換工作，從大型資料集中進行分析找出有用的資訊，牽涉龐大的運算量， 必頇適時減少資料量， 例如降維(Dimension Reduction)、轉換或編碼等方式。

4.資料探勘 (Data Mining)

KDD 過程中最重要的步驟，透過演算法分析資料找出資料潛藏的特徵及規則，包括了資料分群、資料分類、關聯規則、決策樹、統計回歸等演算法。

5.解釋或評估 (Interpretation/Evaluation)

經過資料探勘找出的特徵或模式，可用圖形工具轉換為容易理解的圖表，供決策支援之用；另外評估資料探勘產生的模式之

正確性也是極重要的,用以判斷產生的模式是否可作為未來商業決策上的應用，例如預測顧客的消費金額。

二、資料探勘分類

1.分類（Classification）

按照分析資料已知的事實及其屬性加以定義， 來建立類組。在分類問題中,除了提供預測的分類結果之外,亦可提供 發生這個分類結果的可能預測機率。使用的技巧有決策樹、記憶基礎推理等。例如：信用卡申請者之風險性程度分類。

2.推估（Estimation）

依據目前已有連續性數值之相關屬性資料，以獲致 某一屬性未知之值。使用之法有迴歸分析及類神經網路方法， 例如：顧客價值預測。

3.群集化（Clustering）

面對大量的資訊，我們將相似的事物分群，如此可以使得複雜的資訊變得大幅簡化。群集化於商業上最常見的能用即是市場區隔。例如：顧客分群，以依顧客屬性分類(根據看電影的品味將 觀眾分組)。

4.同質分組（Affinity Group）

將一個異質母體，分隔為一些具相同性質 的群體，即是從所有物件決定那些相關物件應該放在一貨。例如：型錄的編排方式、貨架的擺置方式;例如大賣場相關之電器用品(電話、傳 真機、電話線)，放在同一個貨架上。

5.關聯（Association）

關聯法則在資料探勘的技術中佔有很重要的地位,它主要是找出物品與物品之間的關聯性。在商業上的應用常用來尋找客戶購買行為上，例如：消費者在購買麵包時也會購買牛奶，這就是一種關聯法則。然而對於已知的事實規則，我們並不需要透過資料探勘來確認,真正有價值的關聯資訊是我們所未知的而卻又存在消費者的行為之中,這種令人感到興趣的的資訊正好可以提供經營業者修正銷售策略的參考，例如：辦理促銷活動、舉辦買A商品送B商品的計劃，具有關聯性的商品可以擺設在鄰近的區域，以提升買氣。

6.時序（Time Series）

使現有數值來預測或分析可能值，不過主要是與時間有關聯。

7.迴歸（Regression）

使用現有數值來預測或分析可能值，與時序最大的不同是與時間沒有關聯。

8.序列(Sequential)

在同質分組中找出哪些事物會相伴發生，透過序列找出事物先後發生的順序，我們有時稱這樣的規則為時序規則。例如：網頁瀏覽序列分析。

9.預測（Prediction）

當分類的工作偏向於插入的資料、預測資料分類或發展趨勢時，此時的工作即為預測分析。所有用來進行分類及推定（Estimation）的技術，都可以經過修正之後，透過已知變數數值的訓練組資料來求得，其中歷史性資料是一個很好的來源。歷史性資料可以用來建立模型，以檢視近年來觀察值的變化，若運用最新資料，作為輸入值，可以獲得未來變化的預測值。像是「購物籃分析」 就可以預測在量販店中，哪些商品總是會被同時購買，若經過修正後，也可以透過最新的更新資料，來預測購買行為。

經過相關學者的定義後，可以更了解到資料探勘是一種將資料中未知、潛在的資訊，利用自動或半自動的技術將有意義、有價值的資訊擷取之技術。此一資訊可以幫助決策者作決策 Hui and Jha(2000)與林文修(1998)。資料探勘目前已經成功應用在醫療、零售、電信、金融、保險、工程、製造等行業、以及，如行銷、財務、生產、品管等不同領域之決策(鄭滄祥，2008)。

但是一般的資料探勘僅能從結構化的資料庫中萃取出資訊，因此對於半結構化和非結構化的文字等資料不易處理，半結構化之問題特徵為沒有明確解答、但有條件或者目標可供依循，而非結構化資料則無法可依。因此衍生出文字探勘的方法(Text Mining)，來處理此種類型之資料。

貳、文字探勘

根據 Tan(1999)對文字探勘的定義：「文字探勘是從文件中粹取有趣和不平常樣式(Pattern)或知識的過程」。而 Losiewicz(2000)對文字探勘定義：「文字探勘是從文件集中獲得確切、潛在有用的及最終可理解知識的過程」。且另一位學者 Dan Sullivan(2001)的對於文字探勘的定義則是「編輯、組織及分析大量文件的方法和過程，為了可提供特定使用者特定的資訊，以及發現特定資訊的特徵之間的關連」。由於文件資料大多

不具結構性，所以，無法直接進行分析，必須先對資料預先做處理，擷取出適當的資訊後才能進行，也因此，文字探勘需整合一些傳統的資訊檢索技術，如：關鍵資訊擷取、文件自動分類、全文檢索等，經由對文字資料提供更多的處理，讓使用者更能方便地的從文件資料中取得其所需的資訊。

綜合上述，文字探勘的定義的共通性為非結構化和半結構化文件中萃取出有效資訊和理解知識的過程。

文字資料是最常見資料格式，而文字的資料通常是以半結構化或非結構化的形式儲存。而文字資料中多半含有隱藏之資訊，Hearst(1999)提出文字探勘的定義是從文獻中擷取隱含的知識，以簡要的格式呈現資訊給使用者。文字探勘包含了資訊檢索(Information Retrieval)、資訊萃取(Information Extraction)以及資料探勘(Data Mining)三項主要活動。

文字探勘技術在使用資料前必須先將文字資料經過事先的處理手續，如文字切割、字根還原、同義詞比較、關鍵字比對、未知詞判斷等，而後發掘文字中所呈現的特殊樣式（pattern）、項目（Term）關連（Association Degree），尋找關鍵字（Keyword）以及關鍵詞組(Key Phrase）、使用概念階層（Concept Hierarchy）、分類和推論規則

（Classification or Prediction Rule）、語意網（semantic networks）等技術（丁怡婷，2009）。

文字探勘勘技術應用之領域如自然語言的處理、人工智慧、機器學習、文件分類等問題，應用十分廣泛，市面上亦有數種探勘軟體，但是皆以英文語系為主，英文語系以及中文處理最大之差異在於英文語系文字之間隔明顯，電腦斷詞容易，而中文語系無明顯分界，因此中文語系之文字探勘研究更增加一層困難度，此外共通之問題是有關於語言中歧義、未知詞以及語意辨識問題，仍是目前文字探勘中學者研究的主要重點。

一、文字探勘架構

（Losiewicz, 2000）提出一般的文字探勘架構分為如以下說明：

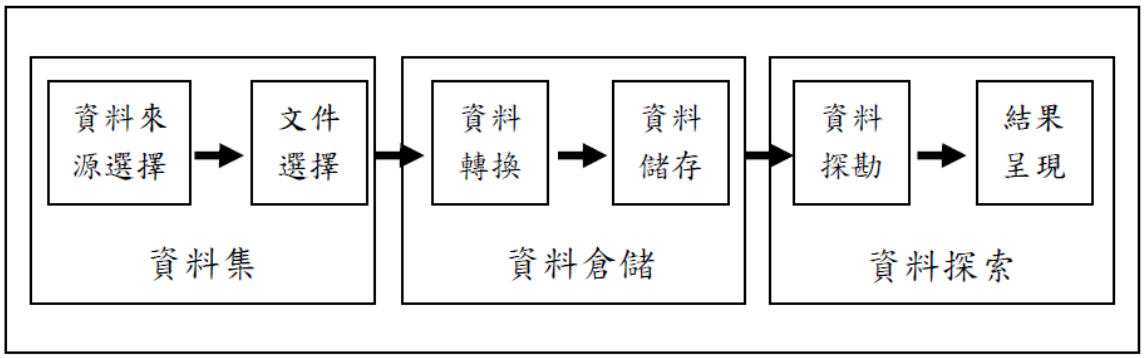

圖 2-2 文字探勘架構
（資料來源：Losiewicz,2000）

（一）資料集（Data Collection）：包括資料來源的選擇及文件過濾。

1.資料來源的選擇：選擇所要探索資料來源的過程，包含察覺可利用的資料、知識領域、了解最終目的。

3.文件選擇：從資料來源選擇、獲得每一文件的過程。此過程可能完全自動化或是由該領域專家來完成。

（三）資料倉儲（Data Warehousing）：含資料轉換和資料儲存。

1.資料轉換：將各個文件的資料表示成資料儲存及資料探勘所要求的特定格式。相當於在 KDD（knowledge Discovery in Database）的預處理步驟。

2.資料儲存（Data storage）：將資料以適當格式儲存以利存取，儲存的格式要能明確說明資料間的關係，使後續處理容易進行。

（三）資料探索（Data Exploitation）：包括資料探勘和探勘結果呈現。

1.資料探勘：找出合適資料模式的過程。

2.結果呈現：將資料探勘的結果加以解釋和視覺化過程，

用來評估資料品質和估計是否選擇的模型和模型的解釋是恰當的。

第二節資訊檢索

在傳統的資訊檢索相關的研究上，ACM 99 年出版的「Modern Information Retrieval」[Baeza-Yates & Ribeiro-Neto, 1999]裡把目前研究領域大致分為以下三種模式。

（一）布林模式（Boolean Model）

布林模式係根據集合理論和布林代數發展出來的簡易資訊檢索模式。布林模式的缺點是它是以二元比較，當文章中包含使用者的查詢時，則其相似性為 1，若不包含則為 0，這種方法的優點是清楚、簡單，缺點是缺乏「程度上」的比較，也就是無法進行相似文章的查詢。因此，有些學者則提出擴充布林模型(Extended Boolean Model)建議加權每個關鍵字以提昇檢索的效果。

（二）機率模式（Probabilistic Model）

機率模式最早由 S.E.Robert 所提出，機率模式（Robertson & Jones, 1976）。主要優點在於能夠計算相似度的機率值，缺點是必須要猜測一堆文章中相關及不相關的集合未考慮到字串在文

件中出現的頻率。對索引字串須假設相互獨立。

（三）向量空間模式（Vector Space Model）

向量空間模型(Vector Space Model，VSM)的概念最早由Gerard Salton所提出，Salton(1983)認為，在資訊檢索的過程中，必須對資訊本身進行分析以幫助檢索之進行。這個分析的過程稱為建立索引，索引主要在表徵文件的內容，同時給予索引詞彙一定的權重(Weight)，以反應該詞彙在文件內容識別的重要性與價值；建立索引的方式為：針對系統中所有的文件所構成之集合D，找出一組屬性（A_1，A_2，…，A_k），使得在D中的某一文件D_i能有一組屬性值（A_{i1}，A_{i2}，…，A_{ik}）具有足夠的資訊以代表該文件，該組屬性值就稱為D_i的索引向量。在目前大部分使用向量空間模型的文件資訊檢索系統中，屬性大多代表某一個詞彙或概念，而屬性值則為該詞彙或概念在文件中的統計資訊（Salton et al., 1983)。它有幾個主要的優點：(1）以Term-weight 的方法改善了資料粹取的效率；(2）它能允許相關文章的查詢；(3）它能計算文章間相似程度，以找出最大相似度的文章。

在向量空間中，系統檢索文件是以相似度值做為依據，查詢句也是以向量表示，可獲得求出文件向量 $\vec{d}_j = (W_{1,j}, W_{2,j}, W_{3,j}, \cdots,$

$W_{n,j}$）與查詢向量 $\vec{q}$=（$W_{1,q}$，$W_{2,q}$，$W_{3,q}$，…，$W_{n,q}$），並經內積相乘則可以計算出二者之間的相似度，如公式 1。

$$sim(d_j,q)=\frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \bullet \left|\vec{q}\right|}=\frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w^2{}_{i,j}} \times \sqrt{\sum_{i=1}^{n} w^2{}_{i,q}}}$$

公式 1

壹、中文詞庫小組

中研院資訊所、語言所於民國七十五年成立一個跨所合作的中文計算語言研究小組共同合作建構中文自然語言處理的資源與研究環境，為國內外中文自然語言處理及其相關研究提供基本的研究資料與知識架構。代表性研究成果包括中文詞知識庫、語料庫及中文處理技術等。

由於網際網路產生大量資訊但缺乏有效的自動化分析方法及技術足以快速處理。為了達到智慧型的資訊處理，知識為本的訊息處理成為目前研究的核心焦點，中文詞知識庫之計畫主要進行三個主要研究方向：知識擷取，知識表達及知識應用。

（一）知識擷取：建構知識本體、語言及常識知識庫

知識建構是一件耗時費事的大工程，在過去二十多年發展了中文處理基礎建設為未來的自動化知識建構打下基礎，這些基礎建設包含標記語料庫、句結構樹資料庫、詞彙庫、中文語法、詞彙分析系統及句剖析器等，且利用完成的基礎知識與技術來自動抽取網路文件中隱含的訊息，擴充現有知識架構並建立領域知識庫及詞彙知識庫，並將連結不同的知識庫形成一個完整的概念網以提高計算機推理及語言了解能力。

（二）知識表達：廣義知網

在知識表達研究方面，知識本體架構的基礎理論及細緻語意的表達模型的研究，藉由分析近義詞的細微差別，並找出細緻語意的表達方式，同時也對知識表達模型及語意合成機制有更多的瞭解，同時也整合了當下最重要的一些知識本體架構，如詞網、知網及事件框架網，得到一個較佳的知識表達系統，稱為「廣義知網」。

（二）知識應用：知識為本的中文語言處理技術

以概念為中心的中文處理技術，所發展的技術將利用自動抽取得到的統計、語言語法及常識訊息作為基礎知識用於分析文件的概念結構並瞭解文件的意義，進而抽取新的知識，然而應用相關中文語言處理技術之概念及步驟以形成一個自動化的學習系統，語文處理系統可經由自動分析學習新知逐日更新知識庫，同時也藉由知識庫的更新增進了語文處理的能力。

貳、中文斷詞處理

在文字的語句裡面，英文詞和詞之間是以空格隔開，而中文以字為基礎，由字的排列組合形成詞，再由詞的排列造成句。中文分詞技術屬於自然語言處理上是最基礎的工作，每一個斷出來的詞都可能是關鍵字，但是詞的長短並不固定，一字可為一詞、另有雙字、三字甚至多字都可能成為一個詞，而詞與詞之間並沒有明顯的區隔，因此若進行斷詞時，句中斷字的位置不同，即可能造成完全不同的語意甚至產生錯誤。現有中文斷詞演算法可分為三類：

（一）詞庫式斷詞法

詞庫式斷詞法為目前最普遍之斷詞方式，其斷詞之品質幾乎憑其詞庫之品質來決定，因此使用此法來斷詞，首要就是加強斷詞資料庫之品質。後續學者亦參考詞庫斷詞法，發展出規則式斷詞法，提昇斷詞的品質(陳克健，1986)。

（二）統計式斷詞法

統計式斷詞法(Sproat, 1990)是參考一大型語料庫（corpus）上的統計資訊，透過相鄰字元同時出現頻率作為斷詞之依據，並以其統計數據加以斷詞，其缺點資料庫不具共通性，由於沒有詞庫或語料庫的輔助，因此有可能擷取出無意義或不合法的詞彙，優點則不易受到語系或文法的限制，可以擷取出詞庫內不存在的專業用語、新生詞彙及專有名詞，出現頻率較高的詞彙較具有代表性。舉例說明，假設有兩個字詞 a 及 b，在文件中發生的次數分別是 f(a) 與 f(b)，a 與 b 同時發生且 b 緊接於 a 之後的次數是 f(a,b)。則可以判定 ab 為一個可能的有效詞，如公式 2。

$$C = \frac{f(a,b)}{f(a) \cdot f(b)}$$

公式 2

（三）混合式斷詞法

混合式斷詞法為同時使用詞庫以及統計斷詞法之方式。(Nie, 1996)先使用詞庫斷出不同組合之詞組，再利用各詞彙之統計資訊，發現最佳的斷詞組合。此法仍需要大型語料庫提供統計資訊。在國內有關於文字探勘之研究中幾乎都是使用中研院中文詞知識庫小組(CKIP)所發展的線上中文斷詞系統 (CWSS)，來進行處理，本研究則用 CWSS 將研究所需之資料進行斷詞處理。

一、中研院中文自動斷詞系統

目前國內中文斷詞的研究中，中研院中文詞知識庫小組(CKIP)發展出的中文自動斷詞系統，已能對中文作精準的斷詞；該系統具備下列功能（中研院詞庫小組，2008)：

1.自斷詞功能。

2.詞類自動標記功能：

該系統將中文詞性分成：形容詞(A)、連接詞(C)、副詞(D)、名詞（N)、感歎詞（I)、介詞（P)、語助詞（T)、動詞（V）八大類，可再細分成 46 類，如表 2-1 所示。

3.選擇不同的詞典，做為斷詞及詞類標記的參考。

表 2-1 詞類標記

精簡詞類	簡化標記	對應的 CKIP 詞類標記	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa、Cbba、Cbbb、Cbca、Cbcb	/*關聯連接詞*/
ADV	Da	Daa	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	Dab、Dbaa、Dbab、Dbb、Dbc、Dc、Dd、Dg、Dh、Dj	/*副詞*/
N	Na	Naa、Nab、Nac、Nad、Naea、Naeb	/*普通名詞*/
N	Nb	Nba、Nbc	/*專有名稱*/
N	Nc	Nca、Ncb、Ncc、Nce	/*地方詞*/
N	Ncd	Ncda、Ncdb	/*位置詞*/
N	Nd	Ndaa、Ndab、Ndc、Ndd	/*時間詞*/
DET	Neu	Neu	/*數詞定詞*/.
DET	Nes	Nes	/*特指定詞*/
DET	Nep	Nep	/*指代定詞*/
DET	Neqa	Neqa	/*數量定詞*/
POST	Neqb	Neqb	/*後置數量定詞*/

（資料來源：中央研究院資訊科學所詞庫小組）

二、詞彙權重計算

TF（Term Frequency）與 IDF（Inverse Document Frequency），是由 Salton、Buckley（1983）所提出來的概念。提出該架構的理由，是因為每個文件中的詞彙佔整篇文件的重要性，是不太相同的，因此 TF 和 IDF 的概念就由此產生了，這兩者的結合，就可以衡量詞彙在文章中的重要性。TFIDF 以詞彙出現在文件中的頻率 TF，乘上詞彙出現的文件佔所有文件中的篇數比值倒數 IDF。由 TF 可以獲得詞彙 ki 在文件 dj 中的重要程度；IDF 則可估算 ki 在各文件中出現的情況，若 ki 出現較多文件，IDF 值會很小，表示 ki 到處可見，對 dj 的重要性似較小(Salton and McGill,1983)。計算方式如下：

（1）$W_{i,j} = TF_i \times IDF_i$

（2）$TF_{i,j} = \frac{freq_{i,j}}{Max(freq)}$

（3）$IDF_i = \log \frac{N}{df_i}$

其中，$W_{i,j}$ 為詞彙 ki 在文件 dj 之權重，$freq_{i,j}$ 為詞彙 ki 出現在文件 dj 的頻率，$Max(freq)$ 為所有在文件 dj 中出現的詞彙的

最大頻率。N 表示文件集中所有文件篇數，dfi 表示文件集中至少出現 ki 一次的文件數目。

第三節 正規表示式

正規表示式英文譯名 Regular Expression，簡稱 Regex 最早是由數學家 Stephen Kleene 於 1956 年提出，後來在資訊領域廣為應用，現在已經成為 ISO（國際標準組織）的標準之一。它易於訂定對字元字串即文字檔處理規則的特性，使得文字處理的規則更具彈性。而規則運算式的大量模式比對標記法可以更迅速的剖析大量文字並將其擷取、編輯、取代或刪除，對於許多處理字串的應用程式（例如 html、Log 記錄檔、Meta Data 或 http header 等），然而規則運算式是不可或缺的工具。

在面對多樣的且不定的格式，建立及篩選這些擷取規則就成為繁重的負擔，而應用 Regex 可有效降低及解析文件的複雜度。本研究使用的資料來源為國家網路醫院網站的資料，由於網頁的原始檔是由文字組成的文件，很適合 Regex 的應用。

一、在 Regex 的應用，Hetal C. Shah（2003）將 Regex 分為四類：

（一）資料的解析

（二）資料的驗證

（三）字元序列的處理

（四）資料的擷取與報表產生

二、Requlat Expression 語法

Jeffrey E. F. Friedl（2002）將 Regular Expression 表示式內容的組成分為兩種：

（一）一般字元（Characters）：其代表意義與欲找尋字元之字面意義相同。

（二）運算（操作）元（Operators）：用以表示某一規則的意義，主要用以作為找尋比對時的特殊控制字元。

三、常用之 Regular Expression 部分語法規則如下表。

表 2-2 正規表示式

規則	說明
^	限制字串必須出現於行首，例：^a 表這串字必須以 a 開頭; 如果 a 出現在其它地方則都不算數。
$	限制字串必須出現於行末，例:a$ 表這串字必須以 a 結尾; 如果 a 出現在其它地方則都不算數。
\	將特殊字元還原成字面意義的字元，例：\(代表 (這個符號，\\ 代表 \ 這個符號，這種表示法適用於 (,),[,] 等在 Regex 有特殊意義的字元。
+	前面的字元或字元集合出現一次或一次以上，例：a+
-	字元集合中可使用 - 來指定字元的區間, 必須包在中括號裡面。例如 [a-z] 表示從 a 到 z 的英文小寫字元, [1-3] 表示從 1 到 3 這三個數字之一
?	前面的字元或字元集合可出現一次或不出現，例：a?
*	前面的字元或字元集合可出現任何次數或不出現，例：a*
{n}	重複 n 次
[]	其中之一字元可出現可不出現，例：[abc] 表示不論出現 a 或 b 或 c 都算符合。
|	代表「或」，例：(Sun|Mon|Tue|Wed|Thu|Fri|Sat)、(日|一|二|三|四|五|六)，必須以左右括號括住。
.	句點符號，代表除了換行符號 (\n) 以外的任一字元，如果要包括換行符號，請使用 [\s\S]
\w (\W)	代表任何英文(以外的) 字元，數字字元也被承認。
\s (\S)	代表空白 (以外的) 字元
\d (\D)	代表數字 (以外的) 字元
\b (\B)	代表位於文字邊界的 (以外的) 字元，例：\bA 可以檢核出 AB，A\b 可以檢核出 BA，\bAA\b 可以檢核出 AA
\r	代表換行字元 (或稱 CR, Carriage Return)
\n	代表換行字元 (或稱 LF, Line Feed，通常和 \r 一同出現，一般以 \r\n 代表換行。
\t	代表 TAB 字元 (或稱 HT, Horizontal Tab)
\(	代表左括號
\)	代表右括號

(資料來源：MSDN、點部落)

第四節 問答系統

壹、問答系統介紹

現今許多網站都有提供相關的 FAQ 讓使用者查閱，依網站的性質提供不同的 FAQ 內容。FAQ 系統也因此被應用於不同領域，如：醫療諮詢、商業客服、特定領域 FAQ 查詢系統……等，至今問答系統之發展主要有兩個方向：一是依循 TREC 比賽中對於問答系統之相關規則，其主要重點為必須使用競賽所提供之資料集作為測試文件集，開發自有的文件搜尋及答案擷取技術應用於該文件集上。而另一個則是利用網際網路上龐大的資源作為回應答案之來源。本節簡述此兩個方向之描述。

（一）TREC QA

TREC 於 1999 年的第八屆大會中，首次舉辦問答系統的競賽（Question Answering Track），其主要目的在鼓勵更多的文字資訊處理領域的學者，能參與自然語言問答的研究，將過去資訊擷取系統以文件回應為主的方式，進階到以更精簡的答案作為系統回應之資訊。TREC QA 競賽的目標在於建立一個開放領域問答系統（Open Domain Question Answering System），亦即所開發之系統必須要能處理不同學科領域之知識詢問，因此需要有龐大且

涵蓋領域多之後端知識庫或文件庫，作為答案萃取的來源。針對後端知識庫這項需求,每年 TREC 競賽皆提供了一個龐大的新聞文件集，提供參賽者開發之系統能夠從中找尋答案，該龐大文件集亦稱為「Answer Corpus」，主要是由 The New York Times（NYT）, Associated Press Worldstream(APW), and Xinghua English (XIE) 等新聞媒體之報導集結而成的，由於整個文件檔的大小多達數 Giga Bytes，因此使用者通常須要對文件作預先處理，包括對文件作主題分類及索引表建立等，以建構文件搜尋引擎並加快系統回應速度。

（二）WEB QA

雖然 TREC 所提供的 Answer Corpus 文件數量已相當龐大，但相對於網際網路上持續劇增的文件資源,TREC 文件數量等級顯然無法相比，且所能回答的問題種類亦受限於其後端文件集之新聞主題，而使得系統的限制性也相較為多，因此有愈來愈多研究者，將相關問答處理技術轉移至網路的資源上作應用，目前已有許多運用網路資源作為答案來源之自然語言問答系統開放線上使用，這些系統有些是只針對某特定學術領域而設計的所謂有限領域問答系統（Close Domain Question Answering System)，但也

有不少開發類似 TREC 的開放領域問答系統。

採用網路文件資源開發之問答系統，雖然看似坐擁龐大的文件庫，但網路文件有許多令人詬病的缺點，像是充斥許多錯誤資訊且文件規格不一等，皆使得處理上增添許多困難。因此，許多採用網路文件作為答案萃取來源的問答系統，都會藉由對問句中的字詞作適當的檢索詞轉換，將大量不相關文件濾除以有效的減少處理量，並加快系統回應速度。

貳、問答系統處理架構

每個問答系統所採用的引擎及處理方式皆不同，而在探討問答系統的技術與架構中，Light（2001）曾提及一個完善的問答系統必須包含下列四個模組，其中也整合一般問答系統處理的流程及步驟：

一、文件收集（Document Collection）

收集大量的文件或相關資料。

二、語句檢索（Sentence Retrieval）、問句分析（Question Analysis）

當自然語言問句輸入到系統時，需要針對問句中語意資訊作分析，並且根據問句中的重要字詞建構查詢，再以此查詢從大量的文件中擷取出可能含有答案的相關文件。關於字詞重要性的決定以及

如何擴充查詢以提高文件擷取之精確度，在資訊擷取（Information Retrieval）領域中有許多相關研究。

三、文件檢索(Document Retrieval)、資訊檢索（Information Retrieval）

為加快文件的擷取速度，後端龐大的文件集必須先做一些前處理，包括對文件作主題分類以及建立索引表（Index Table）等，以提昇問答系統之整體處理速度。

四、答案擷取（Answer Extraction）

答案擷取所觸及的技術層面包含了自然語言處理（ Natural Language Processing）及資訊萃取（Information Extraction）等，分別利用於處理文件中句子之語句結構分析，及進行答案之萃取。最後擷取出的所有候選答案還需要依其正確性作排序（answer ranking）。

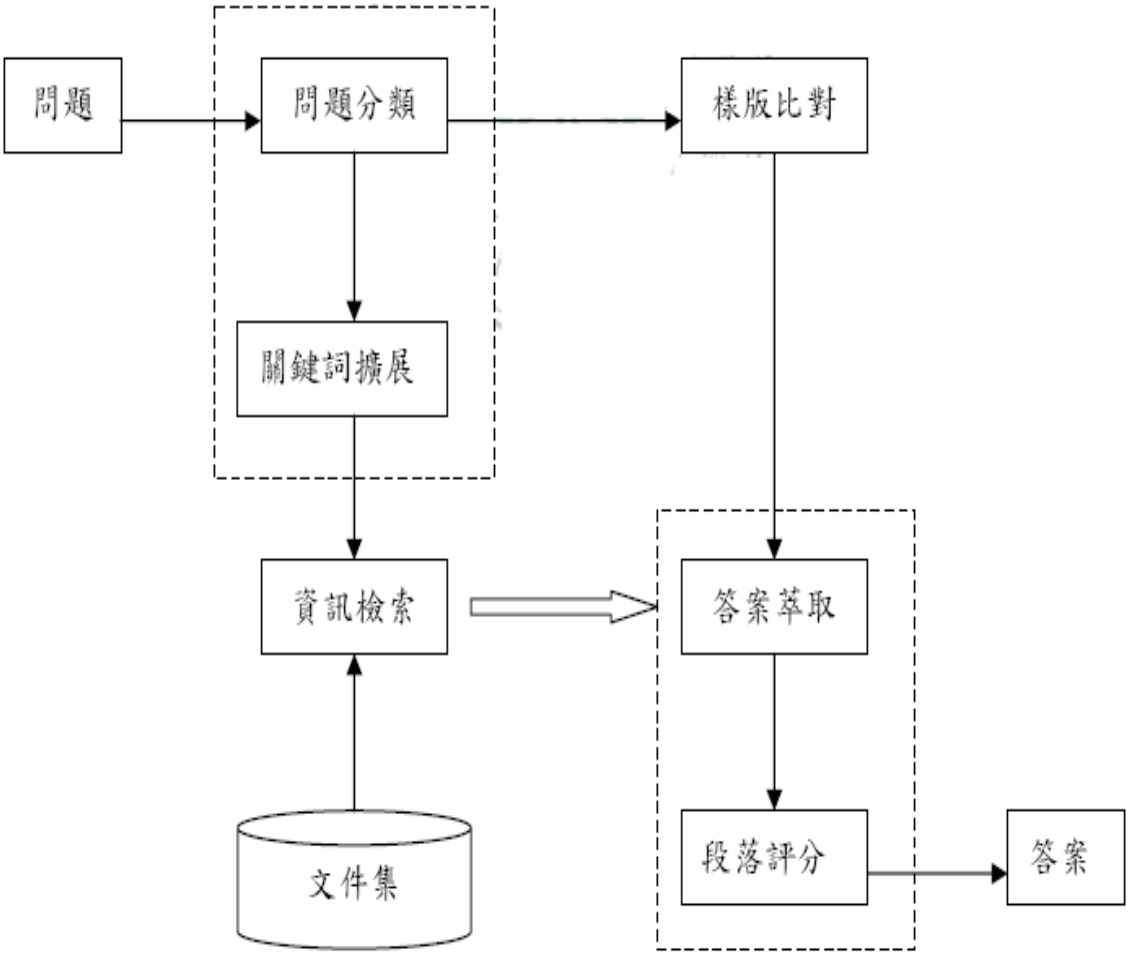

圖 2-3 問答系統架構
（資料來源：Light（2001））

第三章 研究方法

第一節 研究流程

本研究是以文字探勘之資訊檢索技術為基礎設計系統及實作的方式實現疾病問答系統，其研究流程如圖所示。

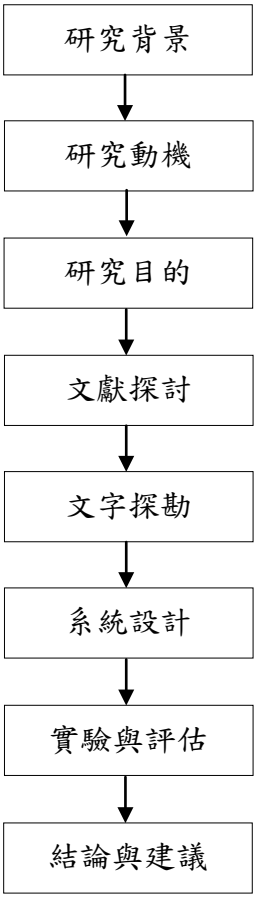

圖 3-1 研究流程圖
（資料來源：本研究整理）

第二節 系統架構

本研究將以資訊檢索技術及自然語言處理建構問答系統並以實驗探討兩者之間的關係，而為了可以讓使用者可以迅速的操作及提高系統相容性，本系統將以 Web-Based 的方式呈現並以架構功能的不同分為三個階段。

一、資料處理

1. 資料來源：主要由網頁所提供的資訊當作資料來源並收集。

2. 資料剖析：排除干擾資料並將不同格式的資料處理及一致化。

3. 關鍵詞萃取：由剖析後的資料萃取資料中的關鍵詞，並轉換成一致的內部表示關鍵詞。

4. 資料轉換：將剖析及萃取後的資料進行資料的簡化及轉換工作並儲存至資料庫。

二、前端應用

1. 使用者輸入：以網頁方式提供使用者輸入及查詢的介面。

2. 問句剖析：將輸入的問句進行剖析。

3. 資訊檢索：以資訊檢索技術將剖析後的問句處理。

三、結果呈現

1. 搜尋答案：將前端所處理後的資料與後端的資料庫進行比對。

2.結果呈現：整合查詢及剖析資料並呈現查詢結果。

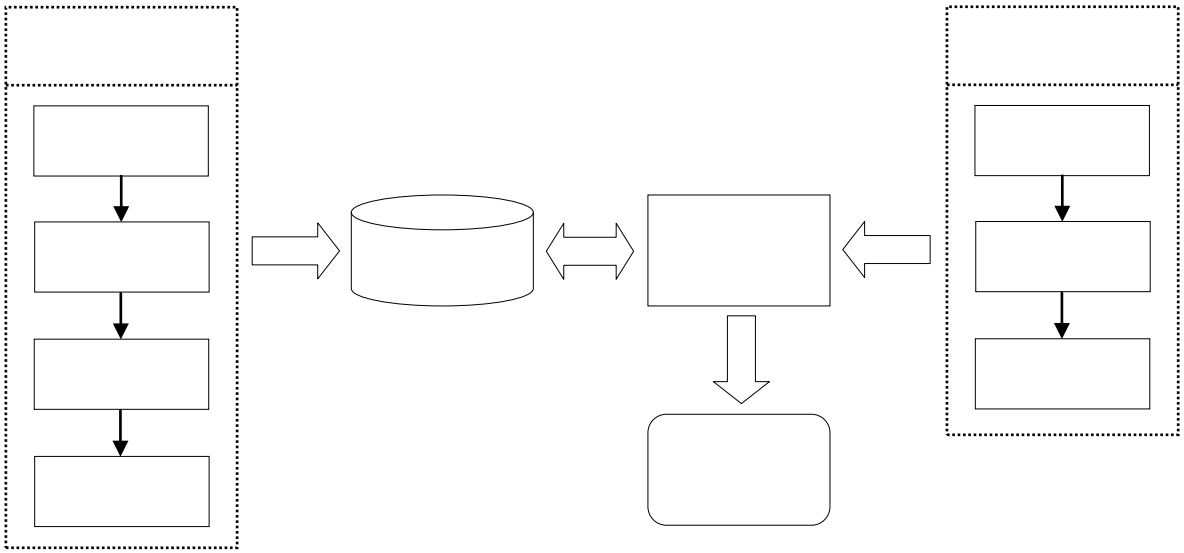

圖 3-2 系統架構
（資料來源：本研究整理）

以整體架構及流程而言，本系統利用網頁擷取技術將網頁資料內容，進行剖析、萃取及轉換並將去蕪存菁的資料儲存到資料庫，而前端的使用者則是將問句輸入並經由資料庫比對以提供合適的解答。

第三節 資料處理

由網頁擷取的資料需要先經過處理後再將其有用的資料儲存，其過程包含資料來源及收集、斷詞處理、關鍵詞萃取、資料轉換及資料儲存。

一、資料收集

由於是疾病相關的資料，本系統將以國內最知名的網路互動平台網頁的網站國家網路醫院當作資料來源，此網站包含了 90 個科別以上且超過上萬筆有關醫病問答的紀錄，而這些資料都是經由 1000 位以上合格專業醫生進行回答，因此為本研究之資料來源的基礎。

本系統以 ASP.NET、JAVA 混合開發，在網頁資料的處理上也大量的運用 Regular Expression 技術將有用的資料過濾並儲存，並於國家網路醫院 1996 年開站以來至 2012 年 3 月止，擷取 20 個科別，共 81068 筆問答紀錄，其分類如表 3-1。

表 3-1 科別統計表

項次	科別名稱	筆數	項次	科別名稱	筆數
1	一般內科	3479	11	牙科	3096
2	一般外科	3316	12	皮膚科	8524
3	一般婦科	7928	13	血液腫瘤科	660
4	醫學美容	1696	14	泌尿科	6667
5	小兒科	159	15	過敏風濕免疫科	1095
6	家庭醫學科	5321	16	神經內科	4408
7	骨科	5490	17	神經外科	1250
8	肝膽腸胃科	6139	18	精神科	3502
9	耳鼻喉科	5137	19	整形外科	2915
10	眼科	4622	20	婦產科	5664
共計 81068 筆					

（資料來源：本研究整理）

（一）前端網頁資料擷取

資料擷取的部分如圖 3-3 以 JAVA 開發，並結合中央研究院之中文斷詞系統如圖 3-4 為斷詞後的資料連結，再由本研究所撰寫的程式進行資料交換、擷取與斷詞，最後將處理並分析過的資料存進資料庫，以利前端語句問答時所需比對之資料。

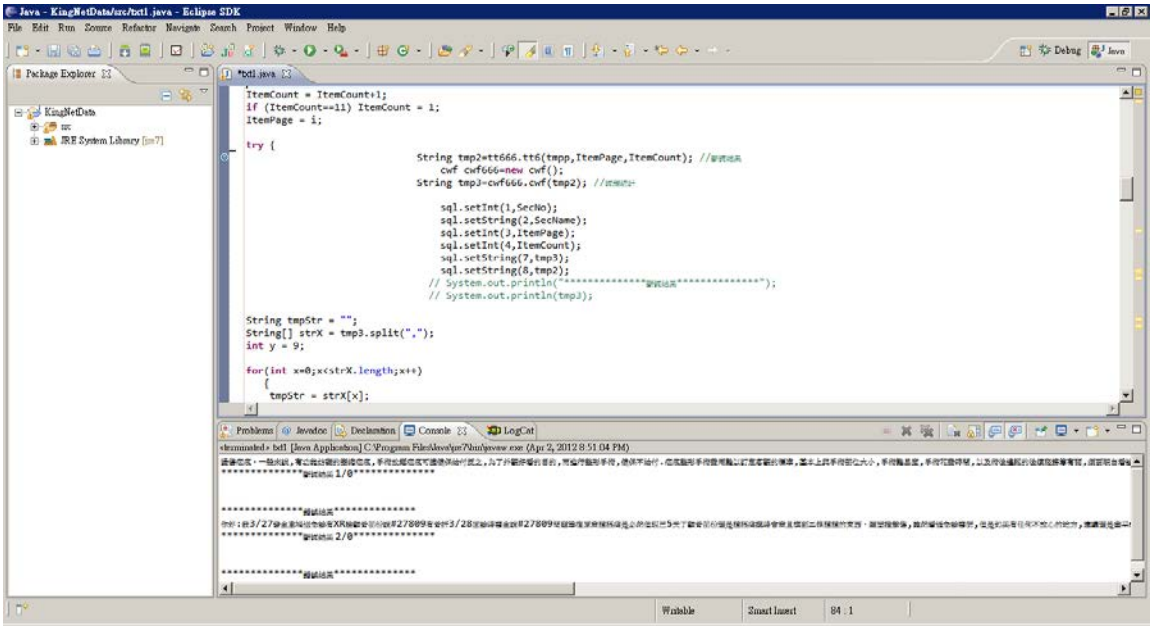

圖 3-3 網頁資料擷取程式碼

（資料來源：本研究整理）

圖 3-4 中央研究院中文斷詞系統

（資料來源：中央研究院）

（三）前端使用者問句查詢

前端使用者提供問句輸入的查詢如圖 3-5，當輸入時將會透過前端程式碼如圖 3-6 先行處理後再傳送到後端進行樣本比對，而程式中也提供科別的選擇，可針對該科別查詢以提高準確率。

圖 3-5 前端使用者問句查詢
（本研究整理）

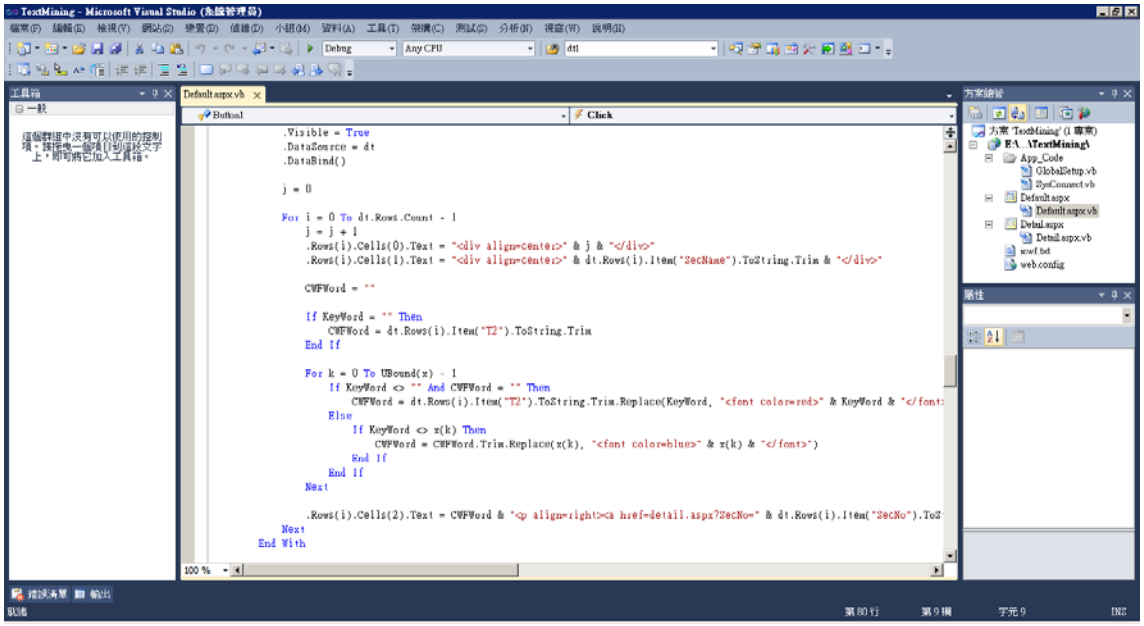

圖 3-6 網頁資料處理程式碼
（資料來源：本研究整理）

表 3-2 問答集範例

問句	問題描述	醫師回覆
坐久了臀部會有灼燒的感覺？	睡覺及坐著的時臀部大腿下方皮膚像被扭曲一樣的壓迫感(立即)，久了會有灼燒的感覺...	神經傳導有其限制,一些週邊表皮神經壓迫並無法偵測出來...
如何改善油性皮膚？	請問因為我前幾天動手術好幾天都沒洗澡和頭，我頭一推頭皮屑，臉也油油的，請問怎樣才能改善油性皮膚...	因為有脂漏性皮膚炎的體質提供網頁文章給您參考何謂脂漏性皮膚炎呢脂漏性皮膚炎並不是屬於油性肌膚，也不是皮膚油脂溢漏...
手或腳會酸痛或刺痛？	有時候我的手或是腳會出現不知道是要用酸痛還是刺痛來形容的感覺...	根據妳的描述，我覺得是循環不順暢所致，而你的循環不順暢應該是心無力所引起...

（資料來源：本研究整理）

二、中文斷詞處理

以程式自動擷取資料後，由中央研究院之中文斷詞系統分析並進行斷詞、詞性標記、詞性分類及詞頻統計，其處理並解析後之結果如表 3-3 範例。

表 3-3 斷詞處理

PatientQuestion（病患問句）
最近經常忘記事情，倦怠感很大，早上起來不想上班。想睡覺，但也睡不久，起來以後又很累。對於屋外的車子聲音，鳥叫聲，家裡的電話聲，日漸無法忍受。每天都覺得不太開心，也不想講話。
QuestionDoctorReply（醫師回覆）
遇到煩惱的事情,除了想辦法處理事情,也記得先處理好自己的心情！生活的壓力，可能會導致情緒、身心、睡眠、精神、體力等多方面勿失調症狀，好比持續施加外力的彈簧，可能出現彈性疲乏一樣的道理建議妳可以直接到門診，接受進一步的評估與適當的診療，情況可以逐漸改善的！
QuestionCKIP（問句斷詞）
最近(Nd) 經常(D) 忘記(VK) 事情(Na) ,(COMMACATEGORY) 倦怠感(Na) 很(Dfa) 大(VH) ,(COMMACATEGORY) 早上(Nd) 起來(VA) 不(D) 想(VE) 上班(VA) 。(PERIODCATEGORY) 想(VE) 睡覺(VA) ,(COMMACATEGORY) 但(Cbb) 也(D) 睡(VA) 不久(Nd) ,(COMMACATEGORY) 起來(VA) 以後(Ng) 又(D) 很(Dfa) 累(VHC) 。(PERIODCATEGORY) 對於(P) 屋(Na) 外(Ng) 的(DE) 車子(Na) 聲音(Na) ,(COMMACATEGORY) 鳥叫聲(Na) ,(COMMACATEGORY) 家(Nc) 裡(Ncd) 的(DE) 電話聲(Na) ,(COMMACATEGORY) 日漸(D) 無法(D) 忍受(VK) 。(PERIODCATEGORY) 每(Nes) 天都(Na) 覺得(VK) 不(D) 太(Dfa) 開心(VH) ,(COMMACATEGORY) 也(D) 不(D) 想(VE) 講話(VA) 。(PERIODCATEGORY) ,(COMMACATEGORY) 遇到(VC) 煩惱(Na) 的(DE) 事情(Na) ,(COMMACATEGORY) 除了(P) 想(VE) 辦法(Na) 處理(VC) 事情(Na) ,(COMMACATEGORY) 也(D) 記得(VK) 先(D) 處理好(VC) 自己(Nh) 的(DE) 心情(Na) ！(EXCLANATIONCATEGORY) 生活(Na) 的(DE) 壓力(Na) ,(COMMACATEGORY) 可能(D) 會(D) 導致(VL) 情緒(Na) ,(COMMACATEGORY) 身心(Na) ,(COMMACATEGORY) 睡眠(Na) ,(COMMACATEGORY) 精神(Na) ,(COMMACATEGORY) 體力(Na) 等(Cab) 多方面(A) 勿(FW) 失調(VH) 症狀(Na) ,(COMMACATEGORY) 好比(VG) 持續(VL) 施加(VC) 外力(Na) 的(DE) 彈簧(Na) ,(COMMACATEGORY) 可能(D) 出現(VH) 彈性(Na) 疲乏(VH) 一樣(VH) 的(DE) 道理(Na) 建議(VE) 妳(Nh) 可以(D) 直接(VH) 到(P) 門診(VA) ,(COMMACATEGORY) 接受(VC) 進一步(D) 的(DE) 評估(VE) 與(Caa) 適當(VH) 的(DE) 診療(VC) ,(COMMACATEGORY) 情況(Na) 可以(D) 逐漸(D) 改善(VC) 的(DE) ！(EXCLANATIONCATEGORY)
QuestionCWF（去除停用詞並統計字頻）
想(4),事情(3),起來(2),倦怠感(1),壓力(1),外(1),外力(1),天都(1),導致(1),屋(1)

（資料來源：本研究整理）

三、關鍵詞萃取

將每個斷詞後的詞彙並與疾病專有名詞配對並儲存到詞彙庫。

表 3-4 關鍵詞萃取

問句	扁桃腺發炎有什麼症狀？
斷詞	扁桃腺、發言、有、什麼、症狀
關鍵詞	扁桃腺、症狀
專有名詞	扁桃腺

（資料來源：本研究整理）

四、資料轉換及儲存

將來源資料經由斷詞處理步驟存到資料庫，其相關資料如下表。

	SecNo	SecName	ItemPage	ItemNo	PatientQuestion	DoctorReply	Qu
1	24	整形外科	0	1	您好 近期冬有眼皮下垂 就是天生的多餘眼皮遮到睫毛 朋友並且有醉翁亭美尾文也曾打過內毒桿菌但是次後效果很...	眼皮整形雖然是最常見的美容手術之一 但是每一位患者的條件與需求皆不同 再加上眼...	
2	24	整形外科	0	2	右眼角內側置痣眼角皮膚拉開才看得到長了一顆表皮實體，約0.1~0.2公分大小的圓形，	眼皮切口縫合過程中 大約於手術後12~14天 倒口縫合被硬 手術後約3個月左右 疤痕開...	
3	24	整形外科	0	3	請問醫是骨盆過大現在有無整形手術比方說削骨盆的手術讓骨盆變小呢，	如果真的是骨盆過大 臨床上沒有理由 網路有結構破壞來縮骨盆腔縮小 不過門診患者中...	
4	24	整形外科	0	4	醫師你好想再請問一下我因腳關節一公分的手術，塊狀線後，醫師建議我運動時避免4~6個月即可，請問指也...	1.理論上來說 貼紙都要比除痘效果更好 2.紙都有與除痘效果又互使用 會使雙方的效果...	
5	24	整形外科	0	5	請問經合後幫吃拿捧抽針等，那如果是二手菸呢，因工作環境會接觸到很多二手菸會對於經合後的恩底有影響嗎，	二手菸也是很玩 對於倒口縫合都會有不良影響 不正確的進行恩底換藥 會使恩底腫張...	
6	24	整形外科	0	6	您好 我是代我老婆詢問的因為我老婆的乳頭天生凹陷有考慮動手術處理但是現今健保好像不給付這類的手術但由...	這個問題看實不是很容易回答 筆者的看法是 1.乳頭及乳量部位有許多種疾病 外觀上看...	
7	24	整形外科	0	7	我家去整形外科割除腳踝上一個約1.5公分的皮下囊腫，請問術後該如何保養避免恩底，另外腳踝會有聲足腫嗎	1.一般來說 手術後前腳馬出聲足腫機率極低 2.外傷或是手術後恩底照顧 須要注意之處...	
8	24	整形外科	0	8	小朋友約4跳出在前羊車鴨蹄舶部和手部有受到壓保因為最近發現她在哭鬧時舶部會出現瘀紅(當時出車鴨蹄擠保...	討論年齡對肥底治療影響的文獻太多 如果您對專業知識有興趣 不防自行閱讀參考 http:/...	
9	24	整形外科	0	9	每當天氣冷的時候 我的乳頭就會很痛且會脆起 是因為我的乳頭內凹的因素嗎最近看到經絡有教透乳頭痛可以幫...	乳頭痛有可能可以改善乳頭內陷問題 但是須要多久才可達成效果 有多少比例患者有效...	
10	24	整形外科	0	10	小朋友約4跳出在前羊車鴨蹄舶部和手部有受到壓保因為最近發現她在哭鬧時舶部會出現瘀紅(當時出車鴨蹄擠保...	1.受保已經經過一年 如果肥底沒有突起 只存在瘢紅問題 可以考慮打KTP或是染料雷射...	

圖 3-7 資料轉換及儲存

（資料來源：本研究整理）

第四章 實驗結果與分析

第一節 系統開發環境

本系統以三層式（Three-Tier）架構為基礎，並以 Java 為網頁資料擷取及 .NET Framework 4.0 ASP.NET 為網頁開發，再結合 IIS Server、SQL Server 等服務，以利整合系統之運作。

一、三層式（Three-Tier）架構

Client 與 Server 間的中介架構，其三階層（Three-Tier）為：

（1）Presentation：負責顯示使用者介面與資料查詢。

（2）Business Process：負責提供應用程式規則。

（3）Database Server：主要是 DBMS（Database Management System,DBMS,資料庫管理系統），用以管理及儲存資料。

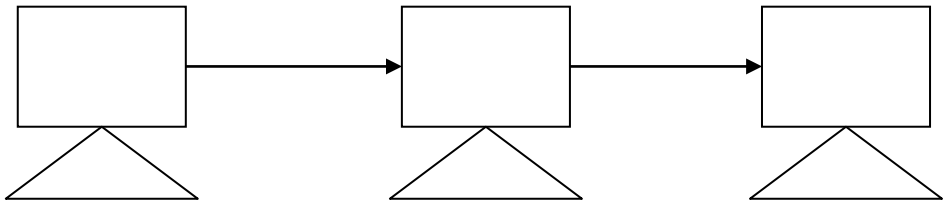

圖 4-1 三層式架構 Three-Tier

（資料來源：本研究整理）

二、系統配置

表 4-1 系統配置

類別	名稱	版本
硬體	處理器	Intel Core 2 Quad Q8400 2.67GHZ
	記憶體	4GB
軟體	作業系統	Microsoft Windows 7 Service Pack 1
	網頁伺服器	Microsoft IIS 7.5
	資料庫	Microsoft SQL Server 2008 R2
	程式開發平台	Microsoft Visual Stuido 2010 Ultimate Eclipse 3.7.0
	程式語言	.NET Framework 4.0 ASP.NET Java JDK 1.7.0

（資料來源：本研究整理）

第二節 實驗設計與分析

本研究資源來源為國家網路醫院網站之醫療問答紀錄，並設計一個實驗，其主要是評估系統整體搜尋答案的效能，以文件相似度計算精確率（Precision）、召回率（Recall）及平均分數（F-meansure）。

實驗設計首先定義 6 個科別，以各科問答集之紀錄各抽出 5 筆，共 30 筆測試問題，再依每一筆測試問題回傳的結果實驗及分析，其說明及公式如下：

一、實驗公式

（1）精確率（Precision）：檢索出來的筆數有多少筆是正確。

$$精確率（\text{Precision}）=\frac{正確筆數}{回傳檢索筆數}\times 100\%$$

（2）召回率（Recall）：資料集中找出來的相關筆數有多少比例。

$$召回率（\text{Recall}）=\frac{正確筆數}{資料集主題相關筆數}\times 100\%$$

（3）平均分數（F-meansure）：召回率及精確率的值越高時，平均分數也會提高，則系統效能也越高。

$$平均分數（\text{F}-\text{meansure}）=\frac{2(\text{PR})}{\text{P}+\text{R}}$$

二、實驗結果

表 4-2 文件相似度

項　目 / TFIDF	回傳檢索筆數	正確篇數	資料集主題相關篇數	精確率	召回率	平均分數
0	264	145	250	55%	58%	56%
0.2	259	137	224	53%	61%	57%
0.4	312	187	296	60%	63%	61%
0.6	295	242	281	82%	86%	84%
0.8	193	137	187	71%	73%	72%

（資料來源：本研究整理）

第三節 系統操作

當使用者輸入問句時，系統會先擷取分析並回應最後的結果，而如果符合問句中的關鍵字或專有名詞則會以紅色字體顯示，以凸顯字句的權重。

以圖 4-2 所示，使用者輸入問句「我常常感冒耶」，首先系統會將不必要的詞類（我、常常、耶）排除，最後留下關鍵字「感冒」，並以權重的分配與後端資料集比對並呈現相關的結果。

圖 4-2 系統操作

第五章 結論與未來研究

第一節 研究結論

本研究以網際網路上所收集的資料為基礎並建構一個問答系統，透過網頁資料擷取的技術及中央研究院之中文斷詞系統將前後台的資料處理後再分類，讓使用者可以使用簡短的自然語句查詢，再以以資訊檢索的方法將相關詞彙依字詞權重及相似度呈現。

由於本研究只探討如何以文字探勘及資訊檢索等技術並建構系統，而再以實驗結果來驗證系統的可行性，但在資料由後端先經過收集及處理分類後，前端使用者輸入並查詢的過程中，可以讓系統迅速提供相關資訊。

第二節 未來研究方向

本研究應用相關技術所建構的系統，主要探討其方法及過程，對於實驗之準確度，雖有用醫療專有名詞比對，但以整體的準確度或許會稍微偏低，在此提出幾點可於未來研究之方向。

一、可將醫療專有名詞及萃取的關鍵詞經由轉換後再更口語化，以符合使用者的習性並紀錄常用的問句，在比對上可以提高準確度。

二、本研究僅探討以文字探勘及資訊檢索等技術建構問答系統，未來可應用語意網技術，可透過屬性的定義及關係來輔助相關詞彙，以提高準確度。

三、由於斷詞結果可能會把專有名詞的關鍵詞斷開，如果把已分開的詞彙透過專屬的資料庫比對及合併，使用者查詢時可將一般詞彙及專有名詞分開並獨立搜尋專屬資料庫，以提高專業性及準確度。

參考文獻

一、中文部分

1. 柯禹伸，使用文字探勘技術預測股票漲跌之研究，北台灣科學技術學院電子商務研究所碩士論文，2011。
2. 黃國政，運用文字探勘技術於人才招募推薦系統之研究，靜宜大學資訊管理學系碩士論文，2006。
3. 江柏勳，基於自然語言處理技術之網路文件問答系統，國立成功大學資訊工程學系碩士論文，2004。
4. 楊勝源，新一代智慧型網路資訊系統 FAQ-master，國立臺灣科技大學電子工程學系博士論文，2005。
5. 曾元顯，專利文字之知識探勘：技術與挑戰，現代資訊組織與檢索研討會，111-123頁，2004。
6. 陳克建，陳正佳、林隆基，中文語句的研究－斷詞與構詞，技術報告 86-006，1986。
7. 巫啟台，文件之關聯資訊萃取及其概念圖自動建構，國立成功大學資訊工程學系碩士論文，2002。
8. 丁怡婷，文字探勘技術應用於中醫診斷腦中風之研究，輔仁大學應用統計學研究所碩士論文，2009。
9. 闕瑞紋，台灣網路族群醫療保健網站使用行為初探，國立陽明大學衛生福利研究所碩士論文，2001。
10. 施威宏，結合分群法和關聯性法則之資料探勘 – 以 104 家教網為例，國立彰化師範大學資訊工程學系碩士論文，2009。
11. 陳治宸，植基於個人郵件之雙層垃圾郵件過濾方法，國立臺灣科技大學資訊工程學系碩士論文，2007。
12. 朱怡霖，中文斷詞與專有名詞辨識之研究，國立臺灣大學資訊工程學研究所碩士論文，2001。
13. 江志銘，應用問答系統技術於電腦領域論壇之研究，國立雲林科技大學資訊管理學系碩士論文，2005。

二、英文部分

1. Hearst, M. A. (1999). Untangling Text Data Mining. Paper presented at the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.
2. Robertson, S. E. & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information.
3. Nie, J., M. Briscbois and X. Ren, 1996,On Chinese Text Retrieval, Conference Proceedings of SIGIR, pp.225-233.
4. Light, M., Mann, G., Riloff, E. & Breck, E. (2001). Analyses for Elucidating Current Question Answering Technology. Journal of Natural Language Engineering 7.
5. Baeza-Yates, R. & Ribero-Neto, B. (1999). Modern information retrieval. Addison Wesley.
6. Frawley, W. J., Paitetsky-Shapiro, G. and Matheus, C.J., 1991, Knowledge Discovery in Databases: An Overview Knowledge Discovery in Database,AAAI/MIT Press, California, 1-30.
7. Grupe, F. H., & Owrang, M. M. (1995). Database mining discovering new knowledge and cooperative advantage. Information Systems Management, 12(4), 26-31.
8. M. J. A. Berry and G. Linoff, Data mining Technique For Marketing, Sale, And Customer Support, Wiley Computer, 1997.
9. Akaka, D. K. (2004). Data mining: federal efforts cover a wide range of uses.Washington Federal Government General Accounting Office (GAO) Report(GAO-04-548).
10. American Association for Artificial Intelligence (AAAI). (2006).
11. Hui, S.C., and Jha, G., 2000. Data Mining for Customer Service Support, Information & Management, 38, pp.1-13.
12. Salton, G., and McGill, M. J., “Introduction to Modern Information Retrieval,” McGraw-Hill Book Company, New York, USA, 1983.
13. Tan, A. H. (1999). Text Mining : The state of the art and the challenges. In Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining(PAKDD’99), Beijing, pp. 65-70.
14. Losiewicz, P., Douglas W. O., and Ronald N. K. (2000) Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information System, Vol. 15, pp. 99-119.
15. Dan Sullivan(2001), Document Warehousing and Text Mining , (Wiley, 2001).

三、網路部分

1. 中央研究院中文詞知識庫小組，http://godel.iis.sinica.edu.tw/CKIP/
2. 曾元顯，文字知識探勘與自動化資訊組織研究成果，
   http://blue.lins.fju.edu.tw/~tseng/
3. 曾元顯，關鍵詞自動擷取技術與相關詞回饋，
   http://blue.lins.fju.edu.tw/~tseng/papers/feedback.htm
4. MSDN，http://msdn.microsoft.com/zh-tw/
5. 點部落 - IT 技術知識社群，http://www.dotblogs.com.tw/

附錄一 問答集

本研究以 30 個問句為實驗之問答集，問答結果如下表內容，而問句內容以簡短及口語化方式表達。

問答集

問句	問題描述	醫師回覆
坐久了臀部會有灼燒的感覺？	睡覺及坐著的時臀部大腿下方皮膚像被扭曲一樣的壓迫感(立即)，久了會有灼燒的感覺...	神經傳導有其限制,一些週邊表皮神經壓迫並無法偵測出來...
如何改善油性皮膚？	請問因為我前幾天動手術好幾天都沒洗澡和頭，我頭一推頭皮屑，臉也油油的，請問怎樣才能改善油性皮膚...	因為有脂漏性皮膚炎的體質提供網頁文章給您參考何謂脂漏性皮膚炎呢脂漏性皮膚炎並不是屬於油性肌膚，也不是皮膚油脂溢漏...
手或腳會酸痛或刺痛？	有時候我的手或是腳會出現不知道是要用酸痛還是刺痛來形容的感覺...	根據妳的描述，我覺得是循環不順暢所致，而你的循環不順暢應該是心無力所引起...
為何會有瘀青？	瘀青浮現的時間有固定嗎當下撞到就一定會出現嗎是不是有些瘀青不會當天撞擊立即浮現...	瘀青其實是挫傷造成的皮下出血，受傷後瘀青出現的時間，會隨受傷的程度不同，而有不同的時間出現...
大便是黑色的是什麼原因？	大便是黑色的然後心臟跳的很快血壓都低低的...	您應大醫院檢查先看心臟科如有吃黑色食物如仙草.豬血糕.鴨血...
常常胃痛要如何解決？	無端端的開始胃痛吃飯前後都有痛過所以不確定痛的原因...	您應再就醫看胃腸專科醫師，平日飲食應避免刺激性食物...
何謂腸躁症？	那我想請問腸躁症的定義是為時間長短及次數又是多久呢...	前 6 個月有反覆性腹痛或腹部不舒服，持續至少 3 個月且有下列 2 項以上...
火氣大會造成口乾	口乾舌燥唇乾吃東西漸漸沒	中醫治療，依脈開藥，很多

舌燥嗎？	味道眼乾痛發紅全身皮膚刺痛發癢難受關節痛肌肉無力很容易疲勞...	病，患者不舒服，不見得檢查得出來，檢查出來通常是無力已回天或只能用消炎藥或類固醇...
皮膚常常發癢是什麼原因？	醫師您好，最近我的皮膚開始很癢，只要一發癢就會紅，抓一下後皮膚就會腫腫的，這是混合型膠原纖維疾病的病徵之一嗎...	皮膚癢、紅疹，臨床上很常見，例如蕁麻疹，輕者數天，重者數十年，我寧可診斷為慢性蕁麻疹。但三總診斷為Collage Disease...
喉嚨常常破掉要如何解決？	喉嚨破掉一段時間了本是左邊喉嚨破掉但幾天就好了隨即馬上換右邊破掉這期間伴隨著 37 的體溫吃藥卻未能改善疼痛狀況右邊喉嚨破掉持續了兩周...	如果僅是這幾週才有這些情況，那只要生活起居保持正常，避免刺激的食物，再過一到兩週後便會改善，如果經常嘴破，又不是固定一個部位...
吃精神藥會變胖嗎？	我吃一年的精神藥為甚麼會變胖我從 46 公斤變成 65 公斤請問吃精神藥會變胖嗎...	這類的藥物的確會讓胃口稍微變好，但是如果配合飲食控制與適當的運動，應該不至於增加這麼多...
注意力不集中是什麼原因？	從小讀書就很容易分心注意力不容易集中而且記憶力也不好背個英文單字都被不起來...	如果診斷智力無明顯問題就要確認焦慮症狀是否達到疾病程度或是有其他成年注意力不集中的問題可與醫師再討論看看...
沒得過水痘可以懷孕嗎？	我是剛新婚不久的新人，目前正準備懷孕，但婚前並沒有做婚前健康檢查，小時候也沒得過水痘，若我想懷孕是否要去注射水痘疫苗或德國麻疹疫苗等...	您可以先自行檢驗是否有德國麻疹的抗體，若有則不需要擔心，沒有的話可以先打疫苗，不過在施打日算起的一至三個月內不建議懷孕...
如何控制糖尿病？	請問糖尿病要控制血糖，那以白飯跟稀飯來說哪一種適合我有聽說稀飯的含糖較高，是不是正確...	大部份的糖尿病患者的治療需要飲食控制、經常運動、及藥物治療三管齊下，才能使血糖穫得良好的控制...
腰痛會有什麼症狀？	請教一個問題因為我最近常常腰痛導致半夜常常睡不好連翻身都很難過去看過醫生	不知道您的腎臟發炎是如何診斷因為我們的腎絲球細胞不會再生若是尿道感染造成

	醫生大概只說腎臟發炎...	的腎盂腎炎就要好好治療...
頭痛會有什麼症狀？	我不舒服的地方在頭部，有時候會覺得頭痛不舒服，睡覺時閉上眼睛，會覺得眼球上方好像有東西壓迫著一樣...	偏頭痛要先考慮，但其他因素如眼壓過高等也有可能這種頭痛引起的劇烈刺痛總是發生在頭的一側，而且常集中在一隻眼睛上面...
頭部麻麻的很不舒服？	就最近有時候會突然半顆頭麻麻的，有時左有時右，可是大約 2 分鐘又不會了...	可能的原因相當多。疲倦，肌緊張性頭痛，青光眼（眼壓高），都可能這樣頭痛...
血小板過高會怎麼樣？	血小板過高會怎麼樣嗎母親日前去醫院無意間檢查出來，血小板過高，而且高出 3 倍左右...	血小板過高容易形成血栓，不過也要去查出血小板過高的病因，症狀部分要看該血栓塞在哪裡來做評估...
肩頸酸痛要如何解決？	個人頸部左後側有硬塊經常感覺肩頸酸痛影響睡眠經中醫門診（一般民間傷骨科）建議：做頸部針灸...	建議先檢查確認腫塊為良性，方可接受針灸治療，您可至一般外科或血液腫瘤科檢查...
為什麼會脖子僵硬？	脖子僵硬問題困擾我已久整各脖子的神經好像僵住每天都硬邦邦的，不時的轉動脖子這時脖子會稍微舒服些以至於不會那麼僵硬...	先去骨科或免疫風濕科看，若無脊椎（僵直性脊椎炎）問題，可能是自律神經失調或坐姿睡姿不良引起...
口臭的原因？	我好像有口臭，我嘴巴會酸酸的，我自己有感覺口腔內味道與正常時不一樣，我父母也說我有進食後會減緩一點...	有口臭的人是亦要考慮有「胃出門螺旋桿菌」感染的存在？在臨床上，口臭的原因很多，舉凡口腔、咽喉、氣管、肺臟、食道、胃以及十二指腸的病變...
膝蓋不舒服？	膝蓋部分常常覺得不舒服伸展時覺得很緊然後喀一聲覺得很舒服但大約 2 分鐘後又覺得不太舒服...	這有可能是膝關節內部半月板的問題，建議你還是就醫讓醫師好好檢查一下吧...
為何心跳很快？	我的心跳數據都是一百零幾下，最少 90 多，最多 115 這樣需要看醫生嗎...	正常心跳:60-100/分應先根據身高體重來判斷是否因為過胖造成心跳過速（因為您未附上身高體重），不過還是

		建議到心臟科門診諮詢一下...
容易拉肚子的原因？	之前一直拉肚子，本來以為是感冒引起的，不過看了感冒後，好像也沒好一點。胃一直覺得脹脹的...	只要稍微吃多一點或是冷熱交替，肚子就開始亂叫亂叫，也會再去拉肚子。這比較像是大腸激燥症不像是急性腸胃炎...
肝炎可以抽血檢驗嗎？	我想幫我爸爸弄健康檢查但我聽說肝炎的檢查光抽血也不見得會準，要做癌症檢查才會知道是真的嗎...	有部分病毒性肝炎及酒精性肝炎當然可以靠抽血確定，不過有慢性肝炎及有肝功能異常經常是兩回事...
胃食道逆流的原因？	我想請問一下什麼樣的情形下會造成胃食道逆流？那是什麼的一個疾病呢？什麼樣的情形會造成胃食道逆流呢...	食道逆流是胃酸會經由胃反向流向食道，造成食道發炎或受傷，引起潰瘍，除了因胃酸的分泌過多的原因之外...
為何眼睛會又癢又痛？	眼睛紅腫、發癢、發熱、眼睛四周腫大早上眼睛有膿狀的分泌物，或是睫毛上有眼垢...	根據您的描述，有可能是感染了急性結膜炎，建議您找眼科醫師檢查並接受治療（點眼藥水），記得常洗手...
禿頭的原因？	約1年前冬天感覺頭部冷冷的-才知道好像有禿頭傾向，請問造成禿頭原因為何...	落髮增加或髮量減少有許多的原因，雄性禿、藥物、節食、內分泌疾病、自體免疫疾病、拉扯感染等等許多的原因都可能造成落髮增加...
如何預防脂肪肝？	日前去照超音波，醫生說我肝比較油，應該是所謂的脂肪肝，但是，我沒有酗酒也沒有過重，只是膽固醇高了點...	會有脂肪肝，並應追蹤膽固醇，必要時應服藥，請回診追蹤檢查運動有幫助，多量血壓...