健康資訊觀念式檢索之查詢句生成

劉驊逵
慈濟大學醫學資訊學系
e-mail：102325106@gms.tcu.edu.tw

劉瑞瓏
慈濟大學醫學資訊學系
e-mail：rlliutcu@mail.tcu.edu.tw

摘要

隨著網路資訊的快速擴大，人們時常會藉由搜尋引擎來搜尋符合自己所需的醫療資訊。許多搜尋引擎是以關鍵字為搜尋基礎，但關鍵字之查詢句無法完整表達一般民眾之健康資訊需求，如能運用自然語言查詢應會更佳。本研究研發一個資訊科技，從自然語言描述的健康保健問題中，自動擷取主要事件並辨識其資訊需求之類別，再產生經擴充後之以關鍵字為主的查詢句，以幫助使用者獲得更多符合需求之健康保健資訊。我們實際以中文健康保健問題為實驗資料並與其它產生查詢句之方法同時送至搜尋引擎比較檢索回來網頁之準確度，實驗結果顯示此技術能顯著提高以健康保健資訊之準確度。

關鍵詞：查詢句生成、查詢句擴充、資訊檢索、觀念式檢索

壹、簡介

近年來網路發展十分迅速，網路變成人們獲取醫療資訊最主要的方法之一[12]。然而隨著網路資訊的快速擴大，人們時常會藉由搜尋引擎來搜尋符合自己醫療需求之資訊[1]。專業醫療網站提供許多健康保健資訊並聘請醫師在網站上來回答民眾之健康問題，提供一般民眾初步解決自身所遇到的健康保健問題(如:KingNet 國家網路醫院[1]、台灣 e 院[2])。民眾不知道有那些專業健康網站，因此當民眾有健康保健問題時，仍常透過搜尋引擎來搜尋健康保健問題，解決自身所遇到的健康保健問題[11]。

1.1 問題定義

本研究提出一個技術，讓使用者輸入自然語言描述的健康資訊需求，系統自動依其中的主要事件及資訊需求產生「以關鍵字為主」的查詢句，以幫助使用者獲得更多符合所需之健康保健資訊

1.2 研究動機及主要挑戰

自然語言是自然的人機互動媒介，是表現資訊需求最直接的方式之一，對於使用者而言。最自然的查詢方式是使用自然語言查詢所需要的資訊。但目前的搜尋引擎常是以關鍵字為搜尋基礎。使用者不易建立一個以關鍵字為基礎的查詢句[14]，故以關鍵字為基礎的查詢句不利於使用者完整表達自己所需要的資訊需求。因此若能運用自然語言進行檢索，較能簡單且清楚表達使用者的想法[3]。

運用自然語言進行檢索目前最大的困難是在自然語言處理中有各種的歧義(ambiguity)難以處理，包含了詞彙歧義、語法歧義、語義歧義，並且將自然語言查詢句轉換成關鍵字查詢句需先進行詞性之頗析並依詞性來斷詞，如何解決歧義及斷詞之精確性，一直是許多人討論之問題[24]。

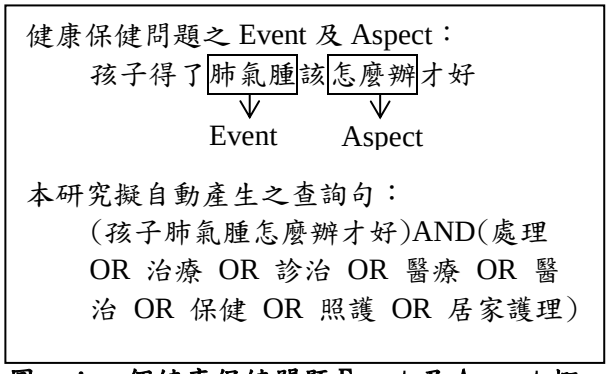
1.3 研究方法

本文所提出之資訊科技將使用者輸入之中文健康保健問題中的主要觀念及資訊需求類別擷取出來，結合關鍵字及代表資訊需求類別之字詞，產生一般搜尋引擎可接受之查詢句，期望獲得更多符合使用者需求的健康保健資訊。觀念式檢索是將中文健康保健問題分為主要事件(Event)及資訊需求類別(Aspect)兩個概念[16]。

Event 是在一個問題當中最重要的主題，

[1] KingNet 國家網路醫院：http://www.kingnet.com.tw/
[2] 台灣 e 院： http://sp1.hso.mohw.gov.tw/doctor/

通常為一疾病名或為一個事件。Aspect 為醫療資訊需求之類別。如，健康保健問題「孩子得了肺氣腫該怎麼辦才好」的主要事件是「肺氣腫」，而根據其中之"怎麼辦"，可知其醫療需求類別是「處理方式」(Process)，如圖一。



圖一：一個健康保健問題 Event 及 Aspect 概念及本研究擬自動產生之查詢句

本研究提出之方法其主要精神是擷取自然語言健康保健問題中的主要觀念並擴充其資訊需求類別字詞，以製成符合一般尋引擎適用的關鍵字查詢句，以期獲得更多符合使用者需求的健康保健資訊。此外，為實際驗證本研究所提出之方法，本研究以實際健康資訊網站之 FAQ 模擬使用者輸入之健康保健問題並與其它文獻提出之方法進行比較，評估檢索回來的相關網頁整體排名之效能，藉此具體衡量本研究所提技術之效能與貢獻。

本文結構如下：第二節探討目前健康保健常見問之檢索及查詢句擴充之方法，分別描述並比較其精神;第三節介紹本文之研究及方法與步驟，詳細說明查詢句產生方法之核心概念與方法流程;第四節為實驗評估，以實際健康保健問題進行評估及驗證;第五節為結論與未來展望。

貳、文獻探討

本研究之相關文獻包含兩類:「常見問題之檢索」及「查詢句擴充」。

2.1 常見問題之檢索

許多健康資訊網站會將民眾問題及回答加以整理，集合成一常見問題集(Frequently Asked Questions，FAQs)，讓民眾能快速獲得解答並減輕專業人士解答民眾問題之負擔。每一個 FAQ 皆由問題及答案兩個部分所組成，但在 FAQ 檢索中，問題部分仍屬最重要之核心。隨著資訊快速的增長，FAQs 也越來越龐大及複雜，如何從 FAQs 中根據使者給予的查詢句找出相似之 FAQ 逐漸變成一個重要的課題，目前較為常見 FAQ 之檢索大致可分為三類。

2.1.1 利用向量空間模型計算相似度

所謂的向量空間模型是指將查詢句與 FAQ 之字詞皆以多維度之向量表示。向量空間模型將每個 FAQ 及查詢句視為由特徵(feature)組成之向量，特徵可以是一個字、字詞或片語。利用查詢句與 FAQ 轉換而來的向量計算 Cosine 內積求得兩向量之夾角，以此來表達 FAQ 與查詢句的相似性。許多的研究也將字詞的 TF-IDF(字詞頻率 × 逆向文件頻率) 融入 VSM，給予向量空間中每個維度不同之權重[4]，TF 是表示一個字詞出現在文件中的次數，IDF 則是由總文件數除以包含該字詞之文件的數目。

2.1.2 根據語意相關度來計算 FAQ 與查詢句是否相似

語意相關度則是計算字詞在詞彙基礎知識庫(ontology)中之距離，來計算字詞在語意上之相似度，WordNet[3]就是一種詞彙基礎知識庫，該知識庫包括許多詞彙之間的關係，如：同義詞、階層關係等。此類方法會根據不同的關係來定義計分公式，以此來衡量查詢句與 FAQ 之間的相似度[4][23]。

2.1.3 根據查詢句與 FAQ 間字詞重疊之比例計算相似度

當 FAQ 及查詢句之間交集之字詞比例越高，表示 FAQ 與查詢句相似度就越高 [2]。這類方法只有考慮字詞重疊之比例，不考慮字詞之間語意的關係

上述所提到常見問題之檢索方法與本研究最大之差異是此類方法是計算查詢句跟 FAQ 之相似度並從 FAQs 中找出相似之 FAQ 提供給使用者。而本研究是藉由使用者所輸入的自然語言查詢句透過健康保健問題中的主要

[3] WordNet ：https://wordnet.princeton.edu/

觀念及資訊需求類別產生一個以關鍵字為基礎的查詢句，送至搜尋引擎找出相關之網頁。

2.2 查詢句擴充

使用者下達檢索字詞時，可能會因為使用不同的字詞，而檢索出不同的結果，有些符合查詢句概念的文件，可能會因為字詞的使用不同，而無法被檢索出來。為解決此問題，查詢句的擴充(query expansion)是較為常見的方法[9]。所謂的查詢句的擴充是根據使用者輸入的查詢句，添加合適的額外字詞，以求更完整的描述查詢句所隱含的概念及主題，以提高一些未被檢索到的文件被檢索出來的機率，使查詢結果更為準確。過去研究中查詢句的擴充方法大致可分為五類，以下分別說明這五類研究與本研究之差異。

2.2.1 從相關文件中挑選字詞

使用者所下達的查詢句進行第一次的檢索，從檢索回來的文件篩選出排名前面之文件進行分析，分析出重要性較高之字詞後將此字詞作為擴充字詞，最後再附加到原先的查詢句上。分析出重要字詞的方法有很多種，如透過潛在的語義分析(LSA)來挑選字詞[18]。此類方法的不足處是當第一次的檢索結果不佳時，排名前面的文件中包含相關文件的比率相對較低，如果從這些文件選取擴充字詞，很可能會選出與查詢主題不相關的字詞，並且本研究是在探討使用者查詢句還未送出之前的擴充，檢索更符合查詢句之相關文件。

2.2.2 依「需要被擴充之機率」來挑選擴充字詞

為了解決查詢句與文件不匹配的問題(查詢句中的字詞沒有出現在文件中)，計算哪一字詞不匹配機率最高，並擴充這個字詞[25]。字詞匹配機率(P(t|R))的正式公式是(r+1)/(|R|+2)，t 代表字詞，r 是指有包含 t 的相關文件數量，|R|是指有相關的文件數量。為了讓查詢句中的字詞與文件不匹配的情況減少，對於有問題的字詞進行擴充，讓它盡可能與相關文件產生匹配。

此方法會先進行第一次檢索，再從檢索回來的文件與查詢句中的每一字詞計算其匹配機率，並針對匹配機率最低之字詞加以擴充，因字詞匹配機率最低者亦是字詞不匹配機率最高者。此方法以關鍵字為基礎，使用者可能無法選擇適當的關鍵字來表達他們所需要的醫療資訊需求，而本研究將自然語言描述之健康保健問題，進行醫療需求之分類，並針對分完類的查詢句，擴充屬於這類別中較為常見的字詞。

2.2.3 從相關詞庫當中選擇擴充字詞

採用額外的資源來選擇擴充字詞，最常見的額外資源就是詞庫，從詞庫當中選擇同義詞當作擴充字詞，增加檢索回來的文件是相關文件之機會[10][19-20]。像是從 UMLS(Unified Medical Language System， [6])中依照語義來選擇相似字詞作為擴充字詞[5]。UMLS 稱為統一醫學語言系統，包含許多的受控詞表和術語。另一個常見之詞庫為 MeSH(Medical Subject Headings)，是美國國家醫學圖書館(NLM)所提出，包含了許多生醫領域中的醫學主題詞[15]。相對於這類字詞擴充方法，本研究的方法能與此方法相互結合提高檢索之效能。本研究是在探討擴充醫療需求的相關字詞是否能有效地增加檢索效能。

2.2.4 以查詢句中的醫療相關字詞來產生查詢句

選擇與醫療類別之相關字詞來產生 CNF 查詢句[17]。此方法在健康保健問題之應用上可能仍有不足。此乃因查詢句中的重要字詞仍可能與醫療類別有關，但是這些字詞不會被擷取出來當作關鍵字。例:「糖尿病能吃蘋果嗎」，此健康保健問題中，蘋果這個字詞不與醫療類別有關，不會被擷取出來，因蘋果沒有跟醫療類別有所關聯。但這個字詞在於此健康保健問題中是重要的，沒有將它當作關鍵字的話，搜尋引擎只會找到關於糖尿病之網頁，不會檢索糖尿病是否能吃蘋果之相關網頁。

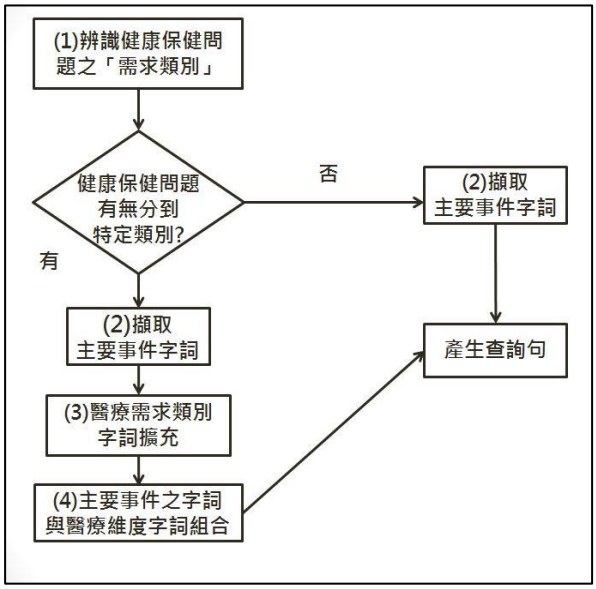
2.2.5 依字詞獨特性產生查詢句

透過計算在查詢句中每一字詞 TF-IDF，取獨特性最高的字詞當作關鍵字以此來產生查詢句[7]。TF-IDF 沒有考慮資訊需求描述中最主要的主題，故不會考慮查詢句的主要意義

來挑選字詞產生查詢句。本研究是透過資訊需求描述中的醫療需求辨識健康保健問題，並擴充其相關字詞。

參、研究方法

本研究提出一個方法自動將自然語言描述的中文健康保健問題擷取「主要事件」，並辨識醫療資訊需求之類別，藉此擴充醫療資訊需求之類別相關字詞，產生一般搜尋引擎可接受之查詢句，期望檢索到更多符合使用者健康資訊需求的健康保健資訊。本方法主要概念是將中文健康保健問題分為 Event 及 Aspect 兩個概念，依照其 Aspect 來擴充屬於同一醫療需求維度之字詞，連同 Event 組合成一般搜尋引擎可接受之查詢句，產生查詢句流程圖，如圖二所示。



圖二：查詢句產生之流程圖

3.1 分類健康保健問題

3.1.1 定義醫療需求類別

在醫療資訊需求類別(Aspect)的部分，使用者提出的健康保健問題的種類有很多，像是某疾病的治療、病因、預防措施、或是罹患疾病之生活型態…等[13][16]。因此我們將健康保健問題之需求分為十二個類別，分別是綜合描述 (description) 、預防 (prevention) 、處理 (process)、治療(medicine)、居家照護(homecare)、診斷(diagnosis)、危險因子(risk) 、症狀(sign)、檢驗(test)、 預後(prognosis)、致死率(mortality)、復發率(recurrence)，每一類別之詳細介紹請參考表一。

表一：資訊需求類別之階層架構及描述

醫療需求之類別	定義
1.1 Description	綜合描述疾病或身體不適
1.2 Prevention	某疾病或身體不適之預防及避免某疾病的發生
1.3 Process	某疾病或身體不適之處理、治療、保健、方法
1.3.1 Homecare	某疾病或身體不適之保健方法及居家照護
1.3.2 Medicine	某疾病或身體不適之醫療治療、用藥
1.4 Diagnosis	如何診斷某疾病或身體不適之原因，包括原因、症狀、醫學檢查
1.4.1 Risk	某疾病或身體不適之原因、風險因子
1.4.2 Sign	某疾病或身體不適之症狀
1.4.3 Test	檢驗是否患有某疾病
1.5 Prognosis	某疾病或身體不適的預後或是未來可能方展
1.5. 1 Mortality	某疾病之死亡率
1.5.2 Recurrence	某疾病之復發情況

3.1.2 醫療需求類別之辨識

本研究透過 SVM(Support vector machine)，對使用者輸入之健康保健問題進行需求類別分類。SVM 是一種以統計理論為基礎所發展出來的機器學習方法，主要用在分類(Classification)和回歸(Regression)上。其分類概念是給予一組事先標記好類別的資料，從中找出一個超平面(hyperplane)，使之將不同的集合分開，使得資料能被區分開來。當有新的未分類的測試資料時，利用計算該資料與超平面的關係，便可正確判定該資料所屬之類別。之前的研究也證明了使用 SVM 來進行文件之分類會比其它的演算法來的好，如：Nearest Neighbors (NN)、Naïve Bayes (NB)、Decision Tree (DT)和 Sparse Network of Winnows (SNoW) [26]。因此本研究採用 SVM 做為分類

的核心演算法，使用的工具是 SVM-light[4]。

對於訓練資料之處理分成 2 個步驟：(1)濾掉疾病名稱(2)訓練特徵之選取。

(1) 濾掉疾病名：優先過濾掉疾病名稱或者是症狀，以避免在訓練資料中某一疾病名稱或症狀固定出現在某一類別，而導致分類錯誤。其方法是從醫院之就診參考或是健康資訊網站，收集疾病名稱及症狀來建立詞庫，收集來源及方式如表二所示。目前包含 7642 筆的疾病名稱及症狀，已能濾掉訓練資料中 80%的疾病名稱及症狀，例:健康保健問題「糖尿病該如何治療?」移除掉疾病名之後為「該如何治療」，再進行特徵選取。

(2) 訓練特徵之選取：訓練之特徵是同時採用單字詞及雙字詞，單獨使用單字詞的話，會有某一字在某一類別出現太多次而造成分類錯誤，若只用雙字詞的話，字詞會有順序的問題存在。SVM 訓練資料之格式是給予每一字詞一個編號，並計算在此筆訓練資料中出現之次數。延續上個步驟，將「該如何治療」，用單字詞及雙字詞選取特徵，選出之特徵為"該"、"該如"、"如"、"如何"、"何"、"何治"、"治"、"治療"、"療"這幾個字詞，並將其轉換成 SVM 訓練之格式。

表二：疾病及症狀名稱之來源及選取規則

疾病及症狀名稱之來源	選取規則
衛生署民眾就醫指引手冊[1]	疾病及症狀全部選取
常見疾病 e 點通[2]	疾病及症狀全部選取
台灣癌症基金會[3]	選取台灣常見之癌症名稱
衛生福利部疾病管制署[4]	傳染疾病名稱全部選取
台灣 e 院[5]	症狀參考表全部選取
A＋線上醫學百科[6]	疾病名稱全部選取

1：http://www.mohw.gov.tw
2：http://www.tma.tw/
3：http://www.canceraway.org.tw
4：http://www.cdc.gov.tw
5：http://sp1.hso.mohw.gov.tw/doctor
6：http://cht.a-hospital.com

[4] SVM-light：http://svmlight.joachims.org/

3.2 擷取出主要事件字詞

3.2.1 查詢句斷詞

我們採用中研院的 CKIP 中文斷詞系統[5]將查詢句進行斷詞，並以詞性擷取出主要事件字詞(Event)。CKIP 中文斷詞系統是中研院詞庫小組開發之系統，執行斷詞並標記詞性。例如「孩子得了肺氣腫該怎麼辦才好?」經過斷詞後會產生「孩子(N) 得(Vt) 了(ASP) 肺氣腫(N) 該(ADV) 怎麼辦(Vi) 才好(N)」。CKIP 所提供之詞性，如表三。我們希望藉由這個方式來把查詢句切成一個一個字詞，以便進行後續的關鍵字之選擇。

表三:CKIP 提供之詞性
（資料來源:中研院斷詞系統）

精簡詞類	意義
A	非謂形容詞，不能充當謂語，不能用「不」和「很」修飾，如:男、女。
ADV	數量副詞、副詞、動詞前程度副詞…等
N	普通名詞、專有名詞…等，如:糖尿病、心臟病、「川崎」症。
DET	定詞， 如:乳癌「三」期
M	量詞，如:幾「劑」
Nv	名物化動詞，動詞當成名詞使用。
T	語助詞、感嘆詞。
Vi、Vt	動詞、形容詞，如:施打、篩檢、懷孕、疼。
FW	外文標記， 如:AIDS，「B」型肝炎

3.2.2 主要事件字詞之選擇

透過中研院斷詞後，我們依每個字詞之詞性，來選擇所需要的關鍵字。我們將採用負面列表的方式，剃除比較沒有可能是主要事件字詞的詞性字詞，因此我們不採用 ADV 及 T 這兩個詞性(視表三)。ADV 代表副詞，不選擇此詞性之原因是副詞通常用來修飾動詞，會是重要字詞的機會很小，如:剛剛、如何、總共、

[5] CKIP 中文斷詞系統：http://ckipsvr.iis.sinica.edu.tw/

一共…等字詞。T 代表語助詞或感嘆詞，語助詞或感嘆詞通常不是重要字詞因此不予選取。例：「孩子(N) 得(Vt) 了(ASP) 肺氣腫(N) 該(ADV) 怎麼辦(Vi) 才好(N)」，剔除非必要之詞性後會留下之字詞是「孩子 得 了 肺氣腫 怎麼辦 才好」。

3.2.3 剔除贅字及合併所有字詞

經過中研院斷詞完之查詢句，進行剔除贅字這個步驟。剃除贅字的方法是使用中央研究院漢語平衡語料庫[6] 中頻率出現最高之前 100 名當作贅字詞表，頻率高代表字詞十分常見，獨特性不高。例：選擇完主要事件之字詞後留下之字詞「孩子 得 了 肺氣腫 怎麼辦 才好」，剃除贅字後剩餘「孩子 肺氣腫 怎麼辦 才好」這些字詞。為了避免 CKIP 把重要的字詞斷錯，而導致搜尋引擎的檢索效果降低，我們會將剩下的字詞合併在一起，讓搜尋引擎自行處理。例：「孩子肺氣腫怎麼辦才好」。如果健康保健問題在分類時沒有被分類到任一類別，直接將擷取出來之字詞當作查詢句，不擴充任何字詞。

3.3 醫療需求類別字詞擴充

健康保健問題在進行需求類別分類後，系統即可進行擴充相似字詞。相似字詞是選取各個類別意思字詞，嚴格定義相似字詞，相似字詞之選取原則是從教育部所提供的線上字典選擇同義字，並剔除明顯不屬於醫療領域之字詞，像是 risk 類別的同義字有「來由」、「理由」、「情由」…等字詞，明顯不與醫療領域有關。此乃因若將不相關字詞一起送至搜尋引擎進行檢索時，搜尋引擎會考慮這些字詞，導致檢索回來一些不相關之網站。因此我們用各類別的意思字詞，嚴格定義相似字詞，讓擴充字詞更符合每個類別的意思，降低不相關網站被檢索回來之機會，各類別擴充字詞，如表四：

表四:醫療資訊需求類別之擴充字詞

醫療需求維度	擴充字詞
Description	何謂、什麼是、什麼叫做
Prevention	預防、避免、防範、提防
Process	處理、治療、診治、醫療、醫治、保健、照護、居家護理
Homecare	保健、照護、居家護理
Medicine	治療、診治、醫療、醫治
Diagnosis	診斷、原因、病因、起因、症狀、病狀、徵兆、病兆、檢驗、檢查
Risk	原因、病因、起因
Sign	症狀、病狀、徵兆、病兆
Test	檢驗、檢查
Prognosis	預後、死亡率、致死率、復發、再生
Mortality	死亡率、致死率
Recurrence	復發、再生

3.4 主要事件之字詞與資訊需求類別字詞之組合及擴充

切割出來的主要事件字詞及辨識出來的資訊需求類別字詞透過 Conjunctive Normal Form (CNF) 來進行組合、擴充。使用 AND 結合主要事件字詞及資訊需求類別字詞，資訊需求類別之相似字詞則是用 OR 組合。其主要原因是主要事件字詞及資訊需求類別字詞都必須出現在檢索回來的網站裡，相似字詞只需當中有一個字詞出現即可，其目的是為了盡可能搜尋符合使用者健康保健需求的網站。延續前例，當使用者之健康保健問題產生之查詢句為「孩子得了肺氣腫該怎麼辦才好?」時，系統產生之查詢句為：(孩子肺氣腫怎麼辦才好)AND(處理 OR 治療 OR 診治 OR 醫療 OR 醫治 OR 保健 OR 照護 OR 居家護理)。

[6]中央研究院漢語平衡語料庫：
http://www.aclclp.org.tw/use_wlawf_c.php

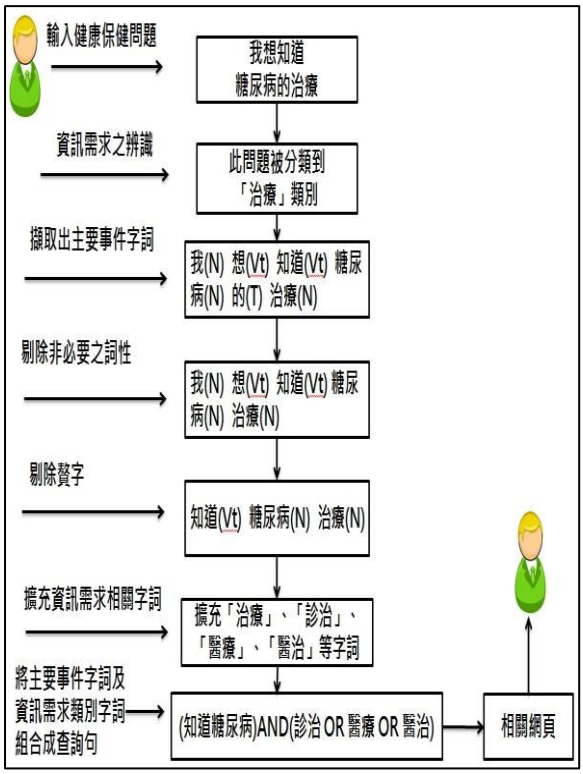
3.5 將查詢句送至搜尋引擎進行檢索

系統產生之查詢句會被送至搜尋引擎(如：Google)進行搜尋，使用者可因此而獲得較相關之健康資訊。延續前例，系統產生之查詢句可以幫助搜尋引擎更精確地著重於肺氣腫之診治照護之道。

最後我們舉個例子來說明我們系統執行的流程，有一健康保健問題為「我想知道糖尿病的治療?」，系統會先辨識此健康保健問題是屬於哪一資訊需求類別，我們利用 SVM 之分類技術來完成這一步驟，此問題會被分到「治療」這資訊需求類別。接下來系統會進行主要事件字詞之擷取。

我們透過中研院所研發的 CKIP 系統來處理主要事件字詞之擷取，CKIP 會依照詞性來進行斷詞，經過斷詞後會產生「我(N)　想(Vt)　知道(Vt)　糖尿病(N)　的(T)　治療(N)」，並剔除非必要之詞性。剔除非必要之字詞後會留下之字詞是「我　想　知道　糖尿病　治療」。並藉由中央研究院漢語平衡語料庫出現頻率最高之前 100 名當作贅字詞表，剔除贅字「我」、「想」這兩個字詞，剩下「知道　糖尿病　治療」這些字詞。當主要事件字詞擷取出來後會進行資訊需求類別字詞之擴充以增加檢索到相關網頁之機會。此問題屬於 Medicine 類別，因此擴充「診治」、「醫療」、「醫治」這些字詞。

最後將主要事件字詞及資訊需求類別字詞用 CNF 組合在一起，最後系統產生之查詢句為：(知道糖尿病治療)AND(診治　OR　醫療　OR　醫治)，並將此查詢句送至搜尋引擎進行檢索。如果健康保健問題在分類時沒有被分類到任一類別，直接將擷取出來之關鍵字當作查詢句，不擴充任何字詞，上述步驟，如圖三所示



圖三：系統流程之範例

總結而言，本研究提出一個透過中文健康保健問題中的主要事件及資訊需求類別結合關鍵字及代表醫療需求類別之字詞進行檢索，本研究主要之貢獻及特點有以下幾點：

(1) 自動辨識中文健康保健問題之資訊需求，並將健康保健問題的資訊需求分為十二個類別。

(2) 根據資訊需求類別來擴充資訊需求類別相關字詞，提升檢索回來之網頁是相關網頁之機會。

(3) 將自然語言描述的健康保健問題轉換成以「關鍵字」為基礎的查詢句，讓使用者能以自然語言的方式簡單且清楚地描述自身的健康保健問題，有效幫助使用者檢索到更多符合自身需求的健康保健資訊。

肆、實驗評估

為了驗證本研究所提出之方法，我們設計了一系列的實驗。在實驗中我們以其它文獻所提出的產生查詢句之方法當作對照組，並以本研究所提出之方法當作實驗組。實驗之過程主要分為兩部分：(1)資料收集：包含了分類器訓練資料時所需的 FAQ 與評估階段的測試查詢句；(2)實驗測試：分別針對實驗組及對照組以相同的查詢句進行測試。以下詳述實驗流程及

評估準則。

4.1 資料收集

資料收集分為以下幾個步驟：

(1) 從 85 個醫療專業網站及各醫院的網站收集 2150 筆健康保健問題，這些網站所列出的健康保健問題是實際民眾所問的問題。

(2) 對這2150筆健康保健問題進行手動分類，將其中的1720 筆作為訓練資料，430 筆作為為測試資料，用以測試分類器之效果。

(3) 另外從快速問醫生[7]有問必答健康網站收集 100 筆測試 FAQ，當作實驗測試 FAQ。此健康資訊網站包含 11 個科別，例如:外科、骨科、內科、兒科等，因此選擇從此網站收集 100 個 FAQ 做為測試的問句。為求公平選取資料來源，對於健康網站中前 20 個科別皆選取數量相同之 FAQ。

4.2 對照組設計

我們設計三種對照組來與實驗組進行效能比較:

(1)「原句」對照組：

原封不動的將測試的問句送至搜尋引擎進行檢索。

(2)「剔除非 Event 字詞」之對照組：

採用詞庫來剔除查詢句中的非 Event 字詞，詞庫是採用文獻[16]中所有非 Event 之字詞，其餘過程與實驗組相同。此對照組是為了證明中研院提供之斷詞方法能有效幫助我們擷取出關鍵字。

(3)「使用字典與 IDF」之對照組：

a) 針對中研院斷詞完的每一字詞去比對是否有在 MeSH 中有的話就擴充其同義字。

b) 假如所有字詞都沒有出現在字典裡，計算所有字詞之 IDF 取最大者，計算 IDF 的醫療文章之來源來自醫院、診所、醫師個人或是某基金會網站裡的文章。

c) 如所有字詞都沒有出現在醫療文章裡，隨機取一字詞當關鍵字。

此對照組主要證明採用醫療資訊需求之維度進行分類及擴充相關字詞是有效的。設計此對照組不僅有擴充相關字詞，也是從自然語言之描述產生關鍵字查詢句，與實驗方法最為符合，並且是文獻中檢索效果最好之對照組[17]。將實驗組及對照組產生之查詢句，同時送至 Google 搜尋引擎，選擇 Google 搜尋引擎的原因是 Google 搜尋引擎檢索回來之健康資訊品質，跟其它一般搜尋引擎相比會是較好的[8]。此外 Google 檢索回來之網站較少有網站開啟失敗的情況產生[22]。

4.3 評估準則

我們選用 MAP 作為評估準則，因 MAP 可以宏觀檢索回來的相關網頁整體排名之成效。 MAP(Mean Average Precision)計算方式如公式(1)所示，其中 AP(i)是指第 i 個查詢句平均準確度，N 則是查詢句之總數，以此實驗為 100 筆。MAP 是取多次查詢之 AP(Average Precision)的均值，AP 的計算公式如公式(2)，T i 是代表很相關網頁之總數。MAP 能代表一個查詢(檢索)系統的整體準確度。檢索系統將相關的網頁排在越前面，AP 值會越高，代表此檢索系統能檢索到相關網頁並將其排名在前面。反之當檢索系統將相關網頁排在越後面，AP 值會越低，表示此檢索系統無法將相關網頁排在前面或是檢索不到其相關網頁回來。

$$\text{MAP} = \frac{\sum_{i=1}^{N} AP(i)}{N} \qquad \text{公式(1)}$$

$$\text{AP}(i) = \frac{\sum_{j=1}^{T_i} \frac{j}{\text{第}j\text{篇目標文獻之排名}}}{T_i} \qquad \text{公式(2)}$$

以圖四舉例說明，第 i 個查詢句檢索回來的前 5 筆網站，很相關的網頁分別排在第一名、第二名、第五名，其 AP 會是(1/1 + 2/2 + 3/5)/5=0.52。我們使用統計之雙尾成對 T 檢定，以 95%信賴區間來判別實驗組與對照組之效能差異，是否有統計上之顯著意義。

排名	1	2	3	4	5	AP
	V $\frac{1}{1}$	V $\frac{2}{2}$			V $\frac{3}{5}$	0.56

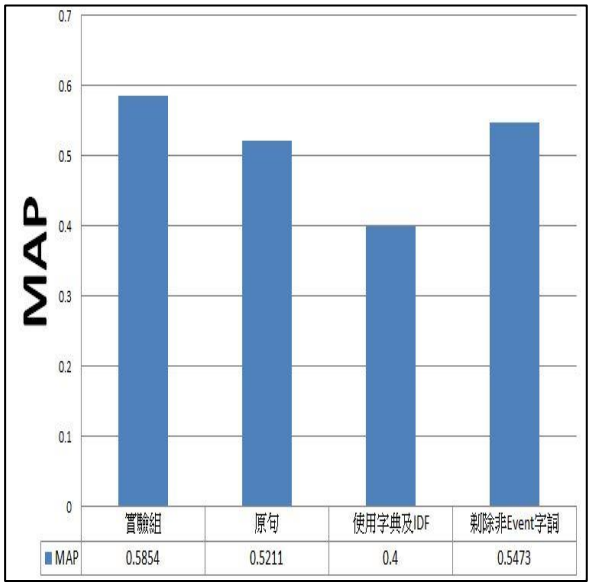
圖四：AP 範例

[7]快速問醫生有問必答健康網站：
http://www.120ask.com/

4.4 實驗結果

我們將 Google 搜尋引擎所檢索回來之前 10 筆網頁來計算其 MAP，以此來評估實驗組及對照組的檢索效能。只選擇前 10 筆網頁的原因是一般民眾透過搜尋引擎搜尋相關資訊時通常只會瀏覽第一頁的搜尋結果，而搜尋引擎檢索回來的網頁是以 10 筆網頁為一頁。

實驗結果如圖五所示，實驗組 MAP 有 0.58 以上之水準，並與其它對照組有顯著上之差異。「原句」對照組與「剃除非 Event 字詞」對照組差異不大，但都輸給實驗組。「字典及 IDF」之對照組是所有組別裡面 MAP 最低的。由此可證明實驗組所產生之查詢句比對照組產生之查詢句都來得好。舉例來說，針對「孩子得了肺氣腫該怎麼辦才好?」，實驗組產生查詢句「(孩子肺氣腫怎麼辦才好)AND(處理 OR 治療 OR 診治 OR 醫療 OR 醫治 OR 保健 OR 照護 OR 居家護理)」，而「原句」對照組送出之查詢句則是「孩子得了肺氣腫該怎麼辦才好?」



圖五：實驗組與對照組 MAP 之比較

標題:肺氣腫怎麼辦?中醫教你來辯證治療
出處:媽咪生活學堂
http://www.mmready.com/content/263335.html
內容(摘錄)
肺氣腫是一種呼吸道疾病，與支氣管炎等疾病有時會有些聯系，要看患者個人情況，不及時治療會給患者造成很多危害，那么患了肺氣腫該怎么辦呢?在中醫看來・・・肺氣腫怎么辦?中醫教你來辨證治療
(1)風寒外感
主證：咳喘并作，痰白而稀，惡寒頭痛，無汗，苔薄白，脈浮緊。
治法：散寒宣肺化痰平喘・・・
(2)痰濁壅肺
主證：氣喘咳嗽，痰多粘稠，咳而不爽，胸悶痛，苔白膩，脈滑。
治法：祛痰平喘・・・
(3)肺喘
主證：喘促短氣，語言無力，咳聲低弱，自寒畏風・・・
治法：益氣養陰定喘・・・
肺氣腫的危害是很大的，大家平時一定要少抽煙，而且不能盲目治療，一定要針對性治療。

圖六：很相關之網頁，但實驗組排在第二名，對照組排在十名外

實驗組檢索回來一筆很相關網站，「原句」對照組沒有檢索出一筆很相關網站，其主要原因是原句本身沒有提供太多使用者所需要的醫療資訊「需求」的字詞，但實驗組透過需求類別之分類後，擴充關於「處理」類別相關字詞，並將這些相關字詞一併送出。由圖六可知因有擴充「治療」字詞，增加了檢索回來的網站是在談論肺氣腫該如何進行治療之機會。而「原句」對照組前兩名檢索回來的網站，非相關字詞出現太多次(如："怎麼辦"、"得了")，故被排在了前兩名，如圖七。然而實驗組有擴充「處理」類別相關字詞，Google 搜尋引擎會一併考慮"OR"字詞，所以這實驗組會把這兩筆網站排除在十名之外。

標題: 人工蕁麻疹怎麼辦？用尼奧 si 普去治療效果
出處:愛問網
http://www.5lask.com/search/人工蕁麻疹怎麼辦？用尼奧 si 普去治療效果
內容(摘錄):
人工蕁麻疹怎麼辦？用尼奧 si 普去治療效果
男 30 歲來自湖南健康諮詢描述：人工蕁麻疹怎麼辦？
還要注意哪些問題呢？想得到怎樣的幫助：• • •
女 30 歲來自湖南健康諮詢描述：得了人工蕁麻疹怎麼辦？
還要注意哪些問題呢？想得到怎樣的幫助
• • •

圖七：不相關之網頁，但對照組排在第一名，實驗組排在十名外

我們亦分析實驗組比「原句」對照組差之例子「乾癬不吃什麼」。實驗組產生之查詢句「(乾癬不吃)AND(何謂 OR 什麼是 OR 什麼叫做）」，而「原句」對照組送出之查詢句則是「乾癬不吃什麼?」。

實驗組檢索回來之 AP 沒有比「原句」對照組來得好，其主要原因在於分類的錯誤。依照此健康保健問題應要分類到「homecare」類別，本應擴充屬於「homecare」類別之相關字詞，但是此健康保健問題卻被分類到「description」類別，擴充了屬於「description」之相關字詞，如圖八的「何謂」及「什麼」出現較為密集，導致實驗組將它排在十名內，而對照組沒有出現任何擴充字詞，但它卻準確談論乾癬之飲食，如圖九。

標題: 乾癬不是癬，談乾癬的十大誤解
出處: 臺大醫院—皮膚科／蔡呈芳
http://www.mdnkids.com/uho_health/detai1.asp?sn=202
內容(摘錄):
乾癬是一種慢性反覆發作的皮膚疾病，皮膚病灶的形態相當多樣化• •
一般民眾對乾癬的認識不足，也往往讓乾癬患者在承受身體的不適外，也要面對更多的社交、就學、就業上的壓力。本次作談會的目締，就在一一破解許多對乾癬錯誤的看法• • •
1.何謂乾癬？與一般所稱的頑癬有何不同？為什麼曾聽說有人用「癬藥」治好過乾癬？
2.乾癬像癌症，終身不治？
• • •
1. 何謂乾癬？
與一般所稱的頑癬有何不同？為什麼曾聽說有人用「癬藥」治好過乾癬？
台灣及日本所稱的乾癬，源自中國古籍所用的「干癬」，大陸則稱為「銀屑病」。
乾癬並非一般的「癬」，因為通常西醫把「癬」定義為• • •

圖八：不相關之網頁，但實驗組排在第三名，對照組排在十名外

標題: 中醫師說你不能吃你就不能吃系列-乾癬
出處:痞客邦
http://joesf0318.pixnet.net/blog/post/220431649-%5B 中醫師說你不能吃你就不能吃系列%5D2-乾癬
內容(摘錄):
中醫師規定你不能吃的東西還真多。
決定乾脆把他寫成一個系列文好了，每種疾病在慢慢補充上去！
今天之所以要選擇乾癬• • •
乾癬食物
1. 該避免的食物
菸、酒、檳榔（甚至連米酒都不要）
進補（當歸鴨、薑母鴨、羊肉爐、麻油、香油）
可能造成你過敏的食物（堅果、花生、瓜子、蝦蟹、竹筍、芋頭）
熱性食物（榴連、芒果、過多蔥薑蒜、辣、炸、麻辣鍋）
2. 可以吃的食物？
A：適量補充蔬果（胡蘿蔔、蕃茄）
• • •

圖九：很相關之網頁，但對照組排在第八名，實驗組排在十名外

伍、結論與未來展望

隨著網路資訊的快速擴大，人們時常會藉由搜尋引擎來搜尋符合自己所需的醫療資訊，自然語言是最自然的查詢方式。因此本研究研發一個資訊科技，從自然語言描述的健康保健問題中，自動擷取主要事件並辨識其醫療資訊需求之類別，再產生經擴充後之以關鍵字為主的查詢句。我們以實際的中文健康問題做實驗，結果證明以中文健康保健問題的主要觀念來切割自然語言之查詢句並擴充其相關資訊需求類別字詞能顯著提升以自然語言進行檢索之效能。

此方法還是有不足及可深入研究之處，包含：(1)針對辨識需求類別之準確度：當健康問題辨識錯誤時而導致擴充錯誤的類別相關字詞會讓檢索效果不佳，如能再擴充分類所需的資料，能提升辨識健康問題之準確度，避免擴充錯誤相關類別字詞。(2)檢索回來網頁之處理：檢索回來的網頁可以經過再一次的排序，透過較為知名的排序方法，將搜尋引擎所檢索回來的網頁重新經過一次排序讓檢索回來的相關網頁盡可能被排在前面，以此來提升檢索之效能 。

致謝: 本研究受科技部研究計畫補助，計畫編號 NSC 102-2221-E-320-007，謹此致謝。

參考文獻

[1] Broussard, R. and Zhang, Y.,"Seeking treatment options: Consumers search behaviors and cognitive activities", Proceedings of the American Society for Information Science and Technology, Vol. 50, Issue 1, pp. 1-10, 2013.

[2] Bernhard, D. and Gurevych, I.,"Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites", Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications pp. 44-52, 2008.

[3] Bader, J. L. and Theofanos, M. F., "Searching for cancer information on the internet: Analyzing natural language search queries", Journal of Medical Internet Research, 5(4), e31 , 2003.

[4] Burke, R. D., Hammond, K. J., Lytinen, S. L. and Tomuro, N.,"Question Answering from Frequently Asked Question Files Experiences with the FAQ FINDER System", AI Magazine Vol. 18 Number 2 (1997), 1997.

[5] Babashzadeh, A., Huang, J. X. and Daoud, M.,"Exploiting Semantics for Improving Clinical Information Retrieval", Proceedings of the 36th international ACM SIGIR , conference on Research and development in information retrieval , pp. 801-804, 2013.

[6] Bodenreider, O.,"The Unified Medical Language System (UMLS):integrating biomedical terminology", Nucleic Acids Research, 2004, Vol. 32, Database issue D267-D270, 2004.

[7] Balog, K. and Weerkamp, W.,"A few examples go a long way: Constructing query models from elaborate query formulations", Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 371-378, 2008.

[8] Chumber, S., Huber, J. and Gheezi, P.,"A Methodology to Analyze the Quality of Health Information on the Internet The Example of Diabetic Neuropathy", The Diabetes Educator February 2015 vol. 41 no. 1 pp. 95-105, 2015.

[9] Carpinetoni, C. and Romano, G.,"A Survey of Automatic Query Expansion in Information Retrieval", Journal of ACM Computing Surveys (CSUR), Vol. 44, Issue 1, January 2012 Article No. 1, 2012.

[10] Dramé, K., Mougin, F. and Mougin, G., "Query expansion using external resources for improving information retrieval in the biomedical domain", CEUR Workshop Proceedings-2014, 2014.

[11] Eysenbach, G. and Köhler, C.,"How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interview", British Medical Journal (BMJ), 324, 573–577 , 2002.

[12] Fox, S. and Duggan, M.," Health online 2013", Health,2013, 2013.

[13] Goeuriot, L. and Chapman, W.,"Building realistic potential patient queries for medical information retrieval evaluation", LREC workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014), 2014.

[14] Ivanitskaya, L., O'Boyle, I. and Casey, A. M., "Health information literacy and

competencies of information age students: results from the interactive online Research Readiness Self-Assessment (RRSA)", Journal of Medical Internet Research, 8(2), e6, 2006.
[15] Kabirzadeh, A. and Siamian, H.,"Survey of Keyword Adjustment of Published Articles Medical Subject Headings in Journal of Mazandaran University of Medical Sciences (2009-2010)", Acta Inform Med. 2013; 21(2): 98–102. Published online 2013 June. doi: 10.5455/aim.2013.21.98-102, 2013.
[16] Liu, R.-L. and Lin, S.-L., "A Conceptual Model for Retrieval of Chinese Frequently Asked Questions in Healthcare", Proceedings of the 8th Asia Information Retrieval Societies Conference (AIRS 2012), LNCS 7675, Springer-Verlag Berlin Heidelberg, pp. 366–375, Tianjin, Mainland China (17-19 December 2012), 2012.
[17] Liu, R.-L. and Huang, Y.-C.,"Medical query generation by term–category", Information Processing & Management, Vol. 47, Issue 1, pp. 68–79, 2011.
[18] Rahimi, M. and Zahedi, M., "Query expansion based on relevance feedback and latent semantic analysis", Journal of AI and Data Mining Vol. 2, 2014, No. 1, pp. 79-84, 2014.
[19] Rivas, A. R., Iglesias, E. L. and Borrajo, L., "Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval", The Scientific World Journal Vol. 2014, Article ID 132158, 10 pages , 2014.
[20] Thesprasith, O. and Jaruskulchai, C., "Query Expansion Using Medical Subject Headings Terms in the Biomedical Documents", Intelligent Information and Database Systems Vol. 8397 of the series Lecture Notes in Computer Science pp. 93-102, 2014.
[21] Somlo, G. L. and Howe, A. E.,"Using web helper agent profiles in query generation", Proceedings of the 2nd international joint conference on autonomous agents and multi agent systems, Melbourne, Australia , pp. 812–818, 2003.
[22] Wang, L. and Wang, J., "Using Internet Search Engines to Obtain Medical Information: A Comparative Study", Journal of Medical internet Research , Vol. 14, No. 3, 2012.
[23] Wu, C.-H., Yeh, J.-F. and Chen, M.-J.,"Domain-specific FAQ retrieval using independent aspects", ACM Transactions on Asian Language Information Processing Vol. 4, Issue 1, March 2005, pp. 1-17 ,2005.
[24] Wang, C.-H., Zhang, M. and Ma, S.-P.,"A Survey of Natural Language Processing in Information Retrivel", JOURNAL OF CHINESE INFORMATION PROCESSING, Vol. 21, No. 2, Mar, 1003-0077(2007)02-0035-11, 2007.
[25] Zhao, L. and Calla ,J.,"Automatic Term Mismatch Diagnosis for Selective Query Expansion", In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 515-524, 2012.
[26] Zhang, D. and Lee, W. S.,"Question classification using support vector machines", Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval pp. 26-32, 2003.