南　華　大　學

資訊管理學系

碩士論文

結合5W1H與本體論進行網路資料探勘技術之研究

A Research For Webmining By Combine 5W1H With Ontology

研 究 生：陳育銘

指導教授：王昌斌

中華民國九十七年六月

論文口試合格證明

南　華　大　學

資訊管理研究所

碩 士 學 位 論 文

利用 5W1H 結合本體論進行網路資料探勘

研究生：

經考試合格特此證明

口試委員：

指導教授：

系主任(所長)：

口試日期：中華民國　97 年　6 月　11 日

誌　　謝

這四年來，碩士班的階段到一段落了，也是另一段落的開始，此刻的心情洋溢著喜悅與不捨。回首這四年來，育銘要感謝的人很多，首先最要感謝的是恩師王昌斌教授，在論文研究期間不辭辛勞地指導與提攜，以及時常地關心與鼓勵支持育銘，雖隻字片語，意更甚於言表，老師，真的很謝謝您！此外，也要謝謝口考老師中華大學吳鴻輝教授及陳宗義老師，幫助多方面問題釐清與思考及提供寶貴的意見。

這段期間感謝南華同儕們彼此的鼓勵與照應，尤其是實驗室成員育弘、政宇在研究上的協助，與你們一同成長的相渡時光非常開心。

最後要感謝我最親愛與最支持我的老婆力凡及爸媽們感謝你們的勉勵，讓我能勇往前進，以完成人生的另一個階段，謝謝你們！

育銘 謹識

于 南華大學資管所

九十七年 六月

結合5W1H與本體論進行網路資料探勘技術之研究

研究生：陳育銘　　指導教授：王昌斌博士

南　華　大　學　資訊管理學系碩士班

摘　　要

隨著電腦與網際網路的蓬勃發展，利用網路搜尋資料已漸漸成為時代趨勢。在目前通用的搜尋方法中，都以單一或多個關鍵字做為搜尋重點，但往往都不能按照自己意圖的描述方式來做搜尋，所以會花費更多時間來找尋不符合意圖的資料。因此，要如何去達到符合自己意的的描述方式來找答案，是目前學界所追求的進一步研究目標。

本研究發展一解析系統，主要是來分析中文問句的意圖，此系統包含三個步驟，第一以中研院 CKIP 系統做斷詞及詞性標記，且以 5W1H 系統化的歸納方法來搭配有限自動機(Finite Automata)的比對手段來解析出意圖型態以及關鍵字，第二以關鍵字來搭配領域本體論(Ontology)做擴展成語意網，第三將解析後所對應的 5W1H 對應字與

語意網做結合。在實驗部分，我們會傳統單一關鍵字與本研究所提出的關聯對應字來做結合，將結合後的字串於網路搜尋引擎中做搜尋處理。其結果部分，我們可以改善其搜尋的準確率約 5%至 34%效能及可縮短搜尋時間。

關鍵詞：5W1H、意圖、有限狀態自動機、本體論

A Research For Webmining By Combine 5W1H With Ontology

Student：Yu-Ming Chen Advisors：Dr. Chin-Bin Wang

Department of Information Management
The M.B.A. Program
Nan-Hua University

ABSTRACT

With computer and the vigorous development of the Internet, use the Internet to search for information has been gradually become the trend of the times. In the current search methods, a description of the query search will often be the key, but often do not accord the user's intentions, then accuracy was challenged. Questions should be to understand how the user's intention is to pursue further research objectives.

In this paper, this research mainly makes analysis of the Chinese question sentence, it contains three steps, the first step is break sentence segmentation and Part-Of-Speech by Sinica CKIP (Chinese Knowledge Information processing) system, and find intention type and keyword by 5W1H induction and the Finite Automata method, the second step expands the semantic net by key words matching domain Ontology, the third step the 5W1H correspondence word and the semantic net do the association to the search engine does the search. This research result showed we proposed joins the 5W1H correspondence word method to be possible to improve its search the precision , can helps user to find intention answer by network.

Keyword: 5W1H, Intention, Finite Automata,Ontology,SemanticNet

目　錄

表　　目　　錄

圖　　目　　錄

第一章、緒論

第一節　研究背景

隨著科技的進步，利用網路搜尋資料已漸漸成為時代趨勢，大多的搜尋引擎以關鍵字為主，其搜尋結果較無法優先排序呈現給使用者瀏覽且資料也無法符合使用者意圖，因此我們需要花更多的精力及時間在做資料的篩選，以造成搜尋的不便。

現今的網路搜尋引擎，隨著時代的流行，大多還是偏向於以單一或多個關鍵字來做資料搜尋，而多半傾向使用多個關鍵字來提高準確率及縮短搜尋時間，但只是透過多個關鍵字來做搜尋資料仍是不足的且無法符合使用者的意圖，因此，透過近年來自然語言處理[8]研究的逐漸活絡，可能運用了此技術，來解決使用者在這方面的問題，也讓早在60年代就已經存在的研究方法，再度呈現新生命。而自然語言的技術主要是透過人們生活起居語言使用情形來建構出大量的語料庫[21] [12]及語言模型[35]，然而，現今電腦科技的突飛猛進，也讓快速的微處理器足以進行更複雜的計算及可做高容量語料的儲存訓練。因此，紛紛吸引了許多學者投入相關模型的建構與開發，更讓許多的難題有了新的理解及解決之道。而經過這些學者們的研究，讓此技術可應用在更多領域上，包括資訊檢索[19]、文件分類[31]、文件

摘要[5]、構詞語法[50]、詞義消歧[27]、字典編纂[48]、詞彙語意[11]、未知詞擷取[17]、Q&A等。

第二節　研究動機

隨著資訊科技的發展與網際網路的普及，人們習慣於網路上搜尋資料，大多以關鍵字或多個關鍵字為主，而往往檢索出來的文件中仍含有大量不是原來冀望獲得的答案，其主要原因在於關鍵詞沒有辦法傳達使用者的意圖[20]，而為了解決此一問題。本研究發展設計一系統，可讓使用者以自然語言方式輸入中文問句，經解析後以分析出意圖及關鍵字，再於網路上做相關的文件內容的搜尋與擷取，藉由擷取後的文件建立出知識間彼此的關聯，並將此知識內容建構成類似知識地圖，供使用者查詢使用。

本機制分為三大功能模組如圖 1-1，包含問題詢答機制、整合型網頁搜尋引擎機制及知識樹建構機制。本研究以問題詢答機制為主，來做更深入探討。

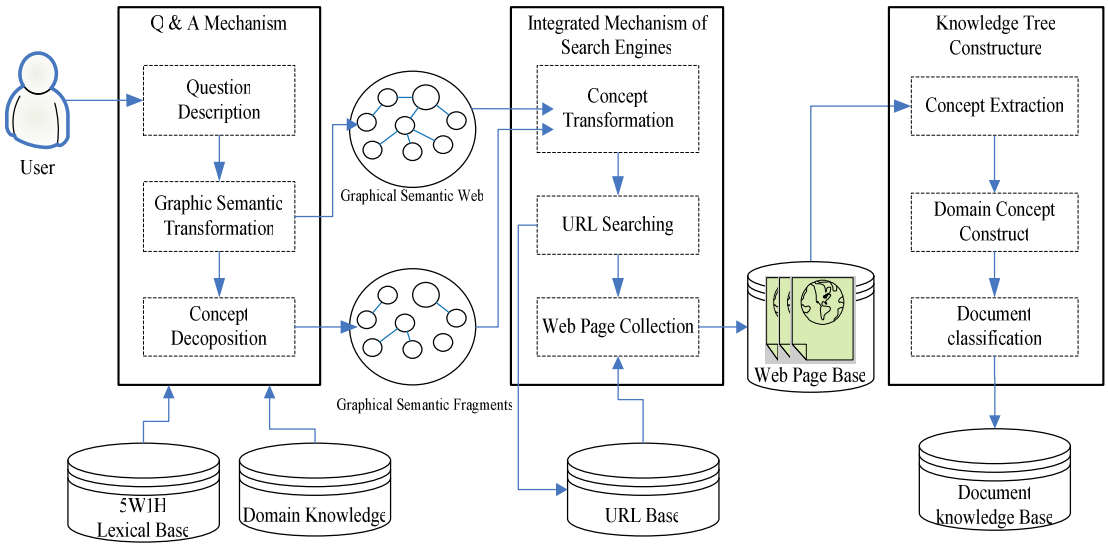


圖 1-1：整體機制架構

本研究為整體架構中的第一階段的問題詢答機制，主要針對使用者所下的自然語言問句做解析，以 5W1H 有系統化的歸納出問句型態模型[49]，快速解析所屬之意圖，且配合所對應之意圖資料庫及做領域本體論(Ontology) [23] [46]之語意網擴展，經意圖與關鍵字結合轉換後的字串於網頁整合型搜尋引擎機制中搜尋，期望讓使用者有效率依其意圖獲得所需答案，透過本機制來解決使用者之問題。

第三節　研究目的

在本篇論文，我們想解析使用者所輸入的自然語言中文問句，以建立具快速分析、可攜性佳、準確度高等特性之中文解析系統。本篇論文之研究目的有下列幾項：

一、分析使用者中文問句的結構、順序及組合方式，研究中文資訊擷取技術。

二、系統化的歸納5W1H問句型態及意圖對應字。

三、分析及擷取中文問句中關鍵字所對應之對應字。

四、建立高準確度、快速分析擷取之中文擷取系統。

五、研究跨領域且高可攜性之中文資訊擷取系統。

六、分析此中文解析系統之實驗結果，並做分析與討論。

第四節 研究流程

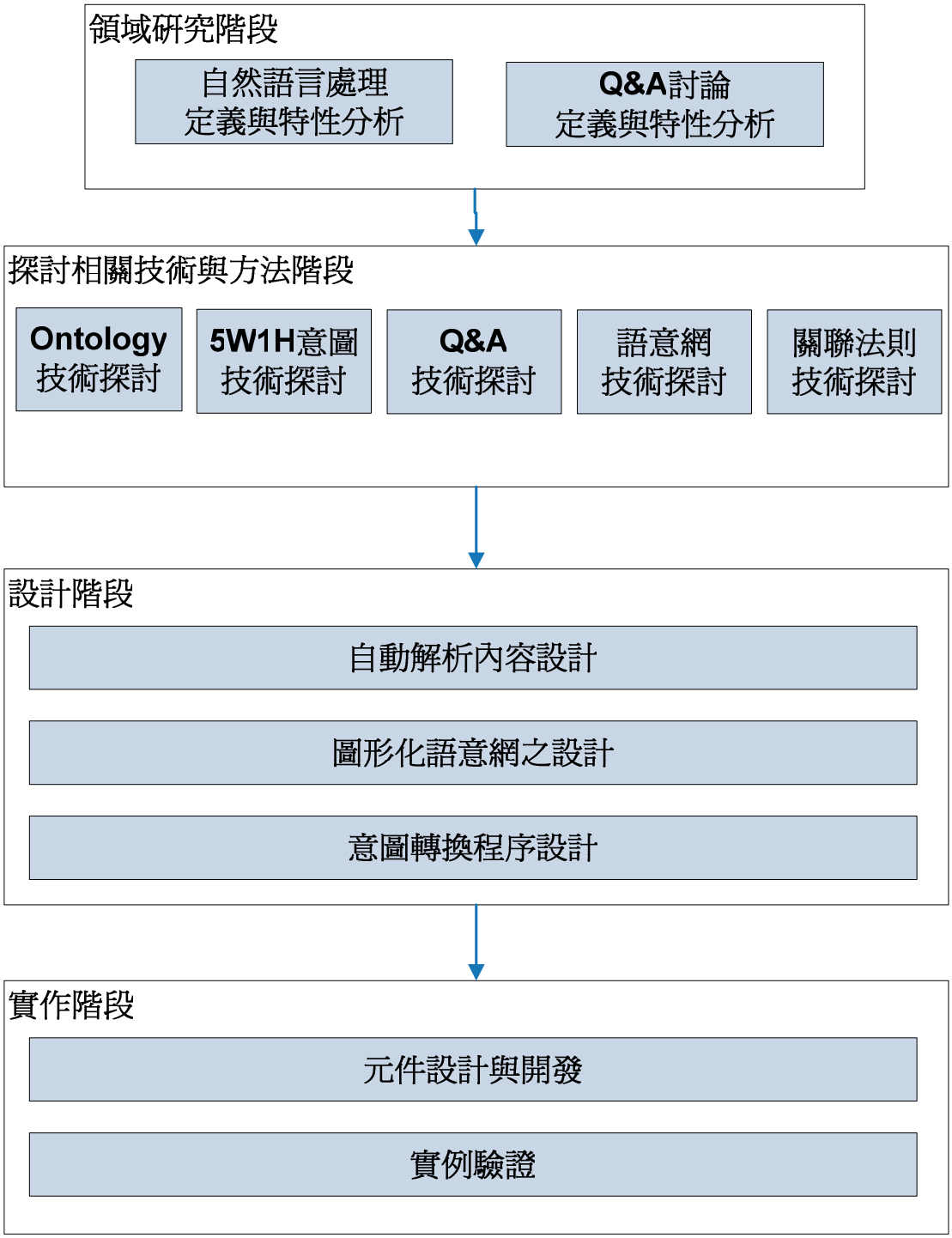


圖 1-2：本研究之研究流程

第五節　論文架構

本論文的章節架構如下：第一章研究背景、動機、目的、流程與論文架構，並概括描述研究的整體架構。第二章將介紹相關的研究及本篇研究論文中所運用到的技術，包含語意網、自然語言處理相關技術、5W1H 意圖、本體論、Q&A 系統及關聯規則等相關技術探討。第三章會詳細描述本篇研究論文的作法及整個系統的流程；第四章介紹系統的使用及說明並且介紹實驗測試資料、方法以及實驗後的結果；最後在第五章中會為本篇研究論文做一個整體性的結論、貢獻及未來發展的描述。

第二章　文獻探討

第一節 問句

在大部分的情況下關鍵詞有助於檢索出我們想要的答案，但是在符合關鍵詞符合的情況下，往往檢索出來的文件中仍含有大量不是原來冀望獲得的答案，而其主要原因在於關鍵詞沒有辦法傳達使用者的意圖。

所以我們認為一個詢問句中可以作為協助檢索的分為兩部分，一個是「關鍵詞部分（keyword segment）」，另外則是「意圖部分（intention segment）」，關鍵詞部分代表構成一個詢問句的關鍵詞，這關鍵詞部分有可能為一個詞或是數個詞所構成的集合；而意圖部分則代表該詢問句的意圖，在這邊我們認為意圖可以由原詢問句中部分片語或是一些片語的組合而成。詳細的定義將在本章介紹。

壹、問句分類

許多語言學學者都提出其對於問句研究之看法，根據張鐘尹[14]的分析，就語法形式而言，疑問句可分成句子和非句子兩大類，再歸成「疑問詞問句」、「選擇問句」、「句尾語助詞問句」、「獨立語助詞問句」、「是非問句」、「附加問句」及「直述問句」。就溝通

功能而言，疑問句可分為外在訊息問句、言談問句、關係問句及表意問句四大類。從說話者的肯定度來看疑問句有從說話者不確定性高的至不確定性低的；從訊息的角度來看，則有從說話者在尋求訊息的至傳遞訊息的疑問句。同時，疑問句亦顯現出從尋求較客觀、指示性的訊息，至傳遞較主觀、以說話者為出發點的態度和看法的分佈。因此這說明即使在句構層次意義的主觀化或說話者介入程度的表達，那種機制的運作亦明顯可見。

其研究結果顯示，疑問句的語法形式與溝通功能雖是多對多的關係，其中卻仍存有某種特定的對應關係。說話者傾向於使用疑問詞問句、是非問句及句尾語助詞問句為「嗎」的問句來尋求自己不瞭解答案的外在訊息，也就是說一般的詢問句(query sentence)多以上述三種類型的問句存在。因此，本論文針對以「疑問詞問句」類型的問句來作分析。

貳、意圖的定義

對於一個自然語言問句我們認為除了關鍵詞部分之外，仍有其他可以作為分辨問句間差異性的部分，舉例而言:「怎麼治療頭痛？」、「為什麼要治療頭痛？」、「治療頭痛的方法有哪些？」。如果只觀察關鍵字，可能會得到「治療」、「感冒」都為以上三句的關鍵字；基於如此相同的結果，我們就無法從關鍵字來分辨第一和第三句應該

最為接近，因為此二句皆旨在詢問治療感冒的方法，而第二句則是在詢問之所以要治療感冒的原因。

因此，一個自然語言問句中的「意圖部分」我們對其定義為：「問句中所傳達最直接想獲得的答案，不需包含前提；意圖可以是原詢問句之子句或片語，甚至結合其他特定片語而成。」透過對問句的分析，使得含有相同意義卻以不同問句句型表現的問句，所萃取出來的意圖部分仍然能夠保持相同。

對於上述例句，當我們從「關鍵字部分」以及「意圖部分」來觀察，如表所示，便可以輕易的分辨的異同。

表2-1 詢問句對應之關鍵字及意圖例子對應表

詢問句	關鍵字部分	意圖部分
怎麼治療頭痛?	治療、頭痛	治療感冒的方法
為什麼要治療頭痛	治療、頭痛	治療感冒的原因
治療頭痛的方法有哪些?	治療、頭痛	治療感冒的方法

參、意圖的萃取

由上一節的說明得知，如果能從問句中正確的萃取出意圖，對於問句意圖的便希有很大的幫助。從語言學的角度來看，問句的語意實與問句的語法形式息息相關，尤其在知道詢問句主要由疑問詞問句、

是非問句及句末語助詞問句構成時，我們針對「疑問詞問句」句型去討探，研究在各種句法結構下每個疑問句的意圖。

肆、疑問詞問句

疑問詞問句相對於英文的WH 問句有相當接近的地位，疑問詞通常出現在與不帶疑問訊息詞相同文法功能的位置上[32]。而在中文有許多疑問詞，例如：「什麼」、「誰」、「怎麼」、「怎麼樣」、「為什麼」、「多少」、「哪裡」、「幹嘛」、「為何」。

通常疑問詞可以協助判斷疑問意圖的核心，像是問句中如果問到「為什麼」，可以想見的該句就是在問某件事物的原因；但是，有些疑問詞會隨著在句子的相對語法位置不同，其意義也不相同。如表2-2所示，「怎麼」這個疑問詞，若出現在副詞前面可作為詢問某件事物的原因，而在動詞前卻作為詢問做某件事的方法[49]。

表2-2 疑問句之意圖對應表

詢問句	意圖部分
如何了解心臟病?	方法
B 型肝炎患者的碗筷應如何處理?	處理方法
為什麼會有頭痛?	頭痛的原因

所以經由語言學上的一些研究結果，以及從收集到的問句中整理歸納，我們定義一套結合語法規則及語意的語意文法，當問句符合語意文法中某一則時，其相對應的意圖之萃取方式也清楚的被規範著。表2-3列出部分語意文法及其意圖部分萃取方式，並舉例說明之。

表 2-3 語意文法例子整理表

類型	問句	語意文法	意圖
What	什麼是運算?	什麼(Nep)是(SHI)運算(VC)	運算之意函
Who	誰會教數學除法?	誰(Nh)會(D)教(VC)數學(Na)除法(Na)	教學習障礙的人
Why	你為何會發生車禍?	你(Nh)為何(D)會(D)發生(VJ)車禍(Na)	發生車禍的原因
When	小孩何時開始學走路?	小孩(Na)何時(Nd)開始(VL)學(VC)走路(VA)	小孩開始學走路的時間
How	要怎麼學數學?	要(D)怎麼(D)學(VC)數學(Na)	學數學的方法
Where	阿里山在哪邊?	阿里山(Nc)在(P)哪(Nep)邊(Ncd)	阿里山的地點

伍、關鍵字的萃取

相對於意圖的萃取，關鍵詞的萃取也是一個不可忽略的部分，藉由關鍵詞萃取我們將對詢問句找出其關鍵字部分。對中文而言，斷詞以及詞性標記的問題一直阻礙國內計算語言學的發展。本研究以AutoTag 作為斷詞及詞性標記的工具。此軟體為中研院資訊科學所CKIP 小組所研發的，經由AutoTag 的協助，可以將一個句子依照分析的結果轉換成一個帶有詞性的詞序列。

在判斷此一詞序列中哪一個詞為關鍵詞時，首先在一般做關鍵詞查詢時，多半會使用的是「名詞」或是「動詞」，所以我們從斷詞後的句子中找出名詞及動詞的部分作為關鍵詞。但是AutoTag 所標記的詞性分類相當細，即使是名詞類仍有許多細分，而部分類別雖屬於名詞卻不作為關鍵詞，例如表2-4所示：

表2-4 中研院平衡語料庫詞類標記集整理表[7]

簡化標記	對應的CKIP詞類標記	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/

Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	*Daa*	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/

Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/

VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE	/*的, 之, 得, 地*/	
SHI	/*是*/	
FW	/*外文標記*/	

上述表2-4中有些為非關鍵詞之名詞詞類，如定詞Ne、量詞Nf、方位詞Ng及代名詞Nh等，而有些詞雖然符合以上規則，但是出現頻率相當高；相對而言，其重要性便降低，不足以視為關鍵詞。所以我們會經由統計收集到的語料庫得到一些高頻率的詞，再經過人工篩選做去除處理。如下表2-5我們將會篩選出直接關係的動詞及名詞部分，其餘則做過濾。

表2-5 詢問句之關鍵字對應表

詢問句	關鍵字部分
鼻子手術後會留疤痕嗎？	鼻子(Na)、手術(Na)、留(VC)、疤痕(Na)
中醫如何治療糖尿病？	中醫(Na)、治療(VC)、糖尿病(Na)
為什麼嬰兒呼吸有雜音？	嬰兒(Na)、呼吸(VC)、雜音(Na)

陸、本節結論

本研究將會以「疑問詞問句」為主，而要如何去萃取「意圖部分」及「關鍵字部分」，所研究會依前5節所述，我們會透過語言學上的一些研究結果，以此概念將眾多問句中做整理歸納，定義了一套結合語法規則及語意的語意文法，來提供本研究的使用。

第二節 語意網

語意網是一種以網路結構為基礎之知識表示法，其主要利用節點(Node)與弧(Arc)進行知識架構組成；當中，節點可以表示物件、概念或特定領域中之情境，而弧則表示節點間之關聯。因此，Ruan 等人(2000) [42]利用語意網之概念，描述醫學專有名詞與其相關資訊來源間之關係，以建構線上醫學資料辭典。

此方法論乃將任意之醫學專有名詞視為語意網結構中的一個節點，連接兩節點之路徑代表兩節點之關係。透過此種表示方式，系統可追蹤任意兩節點間之所有可能路徑，以找出代表兩專有名詞間之所有可能關係。此醫學辭典系統主要應用於臨床醫學領域，其可提供處方籤與藥物資訊給予使用者作為參考。

Smolentsev[45]乃提出一稱為動態語意網(Dynamic Semantic Network，DSN)之知識表示法。其以語意網形式描述資料之概念、物

件和關係的階層屬性，具有(1)整合宣告式知識與程序式知識、(2)可平行查詢語意網之各元件及(3)根據累積之知識發展網絡關係等特性。此外，該表示法的各個知識節點包含運算程序(即解釋所收集資訊之資訊處理演算法，用以發展問題答案、知識節點間聯繫方式與某特定時間之行為)，因此具有自我學習之特性。

此外，Mitri[34]提出一多屬性效能(Multi-Attribute Utility，MAU)之語意網模式，並應用於表示決策評估之知識。此方法將語意網知識表示法融合於決策理論之多屬性效能模式中，利用語意網建構決策判斷所牽涉之觀念、觀念類型、觀念間關係、判斷紀錄和推論策略等知識之語意關係。此些觀念的結合提供一強力之推論能力，即使缺乏直接可用之證據，亦能透過此模式估計一可行之決策方案。該模式乃應用於國際行銷資料庫上，以判斷有潛力之海外市場或外國投資事業。

第三節 自然語言相關技術

Kao 提出一個自然語言介面系統的基本架構[30]如圖 2-1，其系統架構可分為五個主要的單元：自然語言剖析器、語意解譯器、查詢產生器、查詢解譯器及資料庫。整個中文自然語言處理大致可分為：斷詞、語法分析、語意理解、產生查詢代碼及資料查詢。

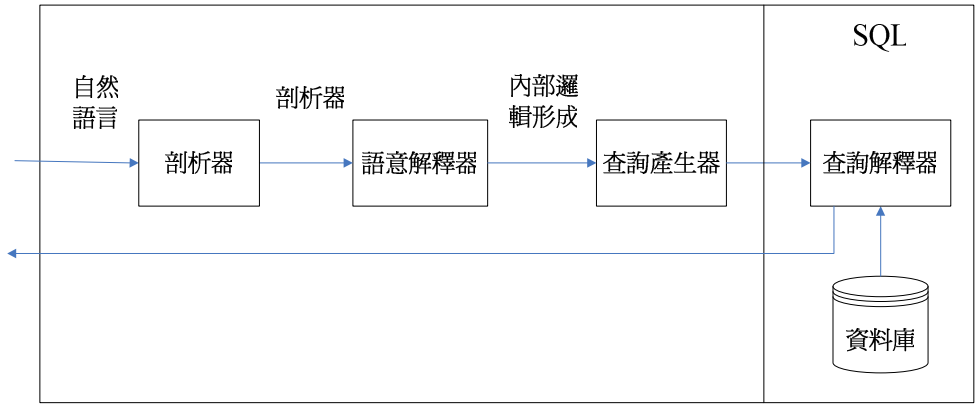


圖2-1 自然語言介面系統架構

壹、斷詞規則

中文字的基本單位為字，而英文句子中最小的單位則是詞(Word)。因此自然語言研究首先面對的問題就是斷詞。另外中文文句中又具有相當多且複雜的歧義問題在。在陳永德[16]的研究列出了中文的歧義類型：句子結構歧義、詞彙歧義、詞類歧義、及詞間歧義。為了解決歧義問題利用斷詞規則來降低歧義性所造成斷詞錯誤，而一般常用的斷詞規則有三大類：長詞優先、前詞優先及詞間頻率對比等規則。

1.長詞優先

中文詞彙的組成字數不一，除了由字和字組成的詞外，尚可以由字和詞、甚至是由詞與詞來組成另一個詞。如「南華大學」，我們可以斷成「南華」和「大學」兩個詞，或者我們也可以只斷為南華大學，在長詞優先的原則下我們會採斷為「南華大學」一詞， 南華大學一詞所包含的意義要比南華或者大學意涵來得多且精確。

2.前詞優先

當兩個相鄰詞組成相鄰部分的字相同且重疊，假設為三個字的字組，前兩個字可以斷成一個詞(前詞)，而後兩字又可以斷成另一個詞(後詞)，在前詞優先下我們會採取將此一字組斷為前詞。如有一組字為「印表機車」，則我們可以斷為「印表機車」或者「機車」，在此一規則下我們斷為「印表機」。

3.詞間頻率對比

此斷詞規則即由詞彙的出現頻率來決定，如前例「南華大學」，在決定要取「南華」、「大學」或者「南華大學」為斷詞結果時，是按詞彙出現的頻率來決定要取那個詞，那個出現頻率較高，則以該詞為優先，而詞頻可由事先讀入語料訓練取得。

貳、斷詞法分類

斷詞的方法可分為詞庫式斷詞法、規則式斷詞法、統計式斷詞法及混合式斷詞法等四種，而其中的混合式斷詞法是由規則式及統計式斷詞法所合併而成[2,10,16]。

1.詞庫式斷詞法

利用人工的方式收集所需的詞彙，預先將其存入詞庫之中，而再利用這個預先建立的詞庫，使用「長詞優先」及「前詞優先」的規則，對需要斷詞的中文文句進行比對，完成斷詞的工作。其優點是詞彙可

以事先經由人工的過濾，挑選出適合的詞彙，可以避免無意義的詞彙。

2.規則式斷詞法

規則式斷詞的學者認為斷詞只是自然語言的一部分，所以強調的是語言的現象。因此事先建立出一套語法規則(Syntax Rule)作為判斷的依據，再將其與詞庫的資料作比對，找出真正的詞彙組合[2]。例如利用一些區域特性，定出某些特殊情形下詞使用特性規則，如定出「以...為例」的規則句型，來表示「以」字出現，其後通常伴隨著「為例」一詞的出現，借由簡單的規則來作中文自然語言處理[9]。

3.統計式斷詞法

統計式斷詞法又有數種不同的斷詞方式，不過以蔡文祥與范文康先生所提出的鬆弛式斷詞法及 Sproat-Shih 的統計斷詞法較為常見[16]，而這兩個統計式斷詞法都必需經由事先用大量的語料來訓練才能使用。

統計式斷詞法的長處在於其不受限於語言的種類及所訂定之語法規則，單純由大量的訓練語料中來取得其所需之相關資訊。但是要得到一個完整的機率統計模式，相對地也需要相當大的語料庫，一旦語料不足或所收集之語料有所偏差，則易造成所統計之機率值可信度降低而導致斷詞錯誤，此外，若統計模型過於龐大，則系統在斷詞所需花費的時間相對的也會變的較長。

4.混合式斷詞法

Yeh 與 Lee 提出以聯併(Unification)為處理基礎之中文斷詞法。首先利用詞庫搜尋出所有可能的斷詞組合，接著利用構詞規則簡化斷詞組合，再以一階馬可夫機率模型列出所有可能的結果，依照機率值排列所有的可能組合，最後使用 HPSG 剖析器(Head-driven Phrase Structure Grammar Parser)，逐一過濾這些斷詞組合，確認該斷詞是否合於文法[16]。

參、詞庫分類

斷詞的歷程中，除了上述的斷詞規則及方法之外，另一個重要的要件就是詞庫，各種斷詞法必需配合詞庫才能順利地作好斷詞工作，而詞庫可以說是斷詞系統的知識庫。詞庫主要可分為以下三種類型，而此三類型的詞庫並非相互排斥，而是可以並存交互應用及配合的。

1.詞庫法

事先由人工收集該應用領域有關的詞彙，建立一個完整的詞庫。

2.規則法

除了收錄詞庫法詞庫的資料外，還存放有關的文法或者句型特性。

3.統計法

詞庫的詞彙來自經由訓練的語意資料，經過統計相鄰字間所出現

的次數再轉為機率模式，找出字與字及詞與詞之間關聯性，將字、詞的相關資訊預先存於詞庫之中。

肆、語意分析方式

一般常用的語意分析模型為馬可夫語言模型(Markov Language Model) [30]，其作法利用事先大量語料統計，將所有可能的詞序找出來，形成一個語意網路架構，再利用詞庫中的資訊找出每個字詞出現的機率或加入其前後詞關係的機率，再計算各不同路徑的機率，擁有最大機率值的路徑，即為最有可能的詞序組合。在李坤霖[4]的研究中，透過語意文法及停用字的篩選，將詢問問句分成意圖部分及關鍵詞部分，意圖部分是找出使用者的意圖目的。關鍵詞部分利用事先建好的關鍵詞樹作比對，而比對的過程會分別求算出詞間的相似度及節點間的機率，進而找出最佳路徑，而這類利用關鍵詞比對處理語意分析方式為目前國內自然語言學者[29]較常使用。

第四節 Q&A 系統相關研究

有許多的學者對Q&A 系統已作出了一些具體的研究成果，第八屆與第九屆的Text REtrieval Conference(TREC-8, TREC-9) 介紹了一系列與question answering 相關的研究，其中主要可分為幾種不同典型的作法，本文針對其作法上的特色作出下列的整理：

壹、LASSO

Sanda M. Harabagiu 等人對Q&A 系統作了一系列的研究並得到了一些具體的成果，特別是針對將knowledge-based技術應用到Q&A系統的設計的方法。

在LASSO 系統[36]中作者們提出一個名為LASSO的Q&A系統架構，其主要的方法是藉由使用創新的自然語言處理方式從大量的文件中找到答案。首先，藉由結合由簡要語法分析結果所得到的句形資訊及能代表問題特性的語義資訊來完成對問題的處理（例如，問題的類型及問題的焦點）。接著，答案的搜尋是架構在名為paragraph index及有關的新檢索方法上。最後為了取得答案及評估其正確性，其使用了一系列的abductive 技術，有些是基於經驗法則，有些則是基於詞彙語義的資訊。

LASSO 系統包含三個主要的模組，分別是Question Processing module、Paragraph Indexing module 及Answer Processing module，分述如下：

- Question Processing module;其主要的工作分為下列四點：

(1) 決定問題的類型

藉由分類出問句的型式，如what、why、who、how、where及when等不同的問題，來決定問題的類型。

(2) 決定期望的答案類型

藉由問題的類型的決定找出其對應的答案類型，如who 明顯的對應到person的答案。

(3) 建立出問題的焦點

焦點是定義問題的一個字或是一串字的排列，而且是可以明確的指出什麼是問題所要尋找的，或什麼是與問題相關的。焦點也是協助找出查詢語句中關鍵字的重要依據。

(4) 將問題轉換為給搜尋引擎利用的查詢語句

取出問句關鍵字的程序是基於一組有ordered heuristics，每一個heuristic 會傳回一組關鍵字，是用來累增問題的關鍵字。

- Paragraph Indexing module;主要有三個部份：

(1) Search engine

主要是藉由NIST的Zprise IR search engine 來完成，Zprise IR search engine 是利用一個cosine vector space model 來建立的，這個模型不允許取出存在文件中的所有關鍵字，但是可以藉由計算查詢問句向量及文件向量間cosine 角的大小來測量其間的相似度[24]。LASSO並利用SGML的擴充以協助建立index，及利用Boolean index 來提高查詢recall 及precision。

(2) Paragraph filtering

透過取出在文件中的paragraph 來作為search engine 查詢的問

句，並以一個限定關鍵字數目的PARAGRAPH n.來限定paragraph 字組的大小。

(3) Paragraph ordering

藉由一個包含有三個不同評分的 radix sort 來完成paragraphs 的順位，這三個測定的標準分別是：最大的Same-word-sequence-score 、最大的Distance-score 及最小的Missing-keyword-score。

- Answer Processing module;主要有二個部份：

(1) Parser

給合來自各式各樣的廣泛的詞典的語義資訊用來確認已經確定的entities，並利用Brill's 詞性標記系統及Wordnet 來協助parser 處理heuristics，以順利的取得如人物、地點、組織、日期、貨幣及產品等資訊。這樣的能力可有效地提供從候選的paragraphs 中找出可能的答案。

(2) Answer Extraction

在parser 從paragraphs 中選出候選的答案後，Answer Processing module 利用一個名為answer-window 的評分機制依序排列出答案。

LASSO 的優點是利用了解單的自然語言處理方式， 改善了一般IR(information retrieval)及IE(information extraction)在無法從從大量文件中準確的取得答案的問題。

Sanda M. Harabagiu 等人在[25]一文中則描述問題分類與

abductive 推論的知識對Q&A 準確性的影響，作者相信Q&A 系統的效能是依賴在其所利用的知識來源之上。並設計了一個名為SOMBRERO 的知識處理模組，主要是為協助LASSO 系統在取得語意的知識及擴展推論結果而設計。SOMBRERO 將可能的候選答案轉換為與問題一樣的語義呈現方式，由一組abductive rules 來控制在問題與答案間呈現方式的一致性。作者發展一個衡量權重的abduction來產生答案正確性的評分以擷取出最有可能性的答案序列。

SOMBRERO 的語意呈現方式是一個anonymous relations 的個案結構，是能夠讓答案與問題透過個案關係來達到一致性的方法。有兩個案例在這種呈現方式中有特別的含義：(1)答案的型態(the answer type)，在問題的架構中可以被呈現出來而且可以用來確認出答案的類型及一致性，(2)焦點(the focus)，在語法解析器中是被定義來調整及比較答案類型的功用，所有的關鍵字都是透過問題焦點的語法所調整而取出的。SOMBRERO 提供了一些abductive rules 用來從已知的知識中找出答案的正確性，其中有兩個重要的因素影響到LASSO 的準確性提昇：首先，問題的分類提供了校正問題焦點正確性的能力，SOMBRERO 以此排除了不含有問題焦點的所有候選答案。第二，透過協調問題與答案間語義的呈現方式可以將問題的分類用在答案正確性的評估上，事實上SOMBRERO 是藉由一個衡量權重的abduction

來計算介於問題與答案之間的語義距離(semantic distance)以分析出答案的適用性。

SOMBRERO 引用了如Wordnet 相關的知識庫及詞彙系統來作為常識性問題的知識分類及詞義（性）標注之用，這種方式普遍的運用在自然語言處理及資訊檢索的相關研究之中，特別是針對open-domain 之語義及語句分析的工作之上。

Sanda M. Harabagiu 等人在[26]則描述了一個整合許多knowledge-based 自然語言處理技術及surface-based 的方法而成的Question-Answering 的系統，來完成從大量文件中取得答案的工作，並且得到不錯的成果。在作法上，不同於一般的IE 系統，為了滿足open-domain 的限制，問題處理程序是透過判斷出問題的型態（例如，what, why, who, how, where 等）及期望的答案類型來取代IE 系統中的linguistic pattern 比對。接著，由經驗法則建立出的heuristic order 可以用抽取出與問題焦點相關的關鍵字組，由這些字組所取得的paragraph 可以提供給IR系統快速的取得可能的答案集合，並由一組評分的機制依序排列出答案。作者針對實現一個open-domain Q&A 系統規納出一些觀點：

- 針對open-domain Q&A 系統的流程而言，運用簡單的knowledge-based 自然語言處理技術及surface-based 的方法即可

得到令人驚訝的成果。問題分類的語義呈現(例如，what, why, who, how, where 等)可以確認出問題的類別(question class)進而從中取得可能的答案類型(answer type)、問題的焦點(question focus) 及問題關鍵字的語義類別(semantic class of question keywords)。經由語義的類別可以取得複合的關鍵字組，以提供給IR 系統取得答案集合及作為評估答案正確性的依據。

- 透過一個語彙的知識階層架構（如Wordnet 的語義階層架構）可提供完成語義轉換及協助問題的分類(question taxonomy)。運用已知詞彙所對應的知識分類的位置，明確的提供了問題中關鍵字詞的相對關係及釐出能代表此問句意念的重要關鍵字。
- 一個Q&A 系統的效能是決定在其所利用的知識來源，運用knowledge-based 的方法能有效地增進答案的準確性。

此研究所實現的系統透過運用現有的知識系統（如Wordnet）作為協助在問題分類、答案尋找及結果修正的重要參考依據，並由實驗證明了knowledge-based的方法能直接的提昇Q&A 系統的效能。

貳、GuruQA

John Prager 等[37]提出了一個名為predictive annotation 的問題解答技術，其方法是針對下列五點觀察：

- 問題可以藉由它們所要尋求的答案被分類

- 答案通是存在詞組的型態中
- 答案的詞組同樣可以使用分類問題的方法來分類
- 答案可以用淺顯的的語法解析技術從文件中被取出
- 用來作為問題解答的詞組的內容通常是文章中的某一小段

基於上述的五點觀察，作者使用了下列幾個主要的方法設計出一名為GuruQA 的QA 系統：

(1)問題分類的方法

定義了一個稱為QA-Token 的問題類型分類表作為問題及答案的分類依據，如同LOSSA 系統利用wh-word 的方式。

(2)字詞的檢索的方法

使用一個稱為Textract 的文件處理系統來作為字詞的檢索的主要工具，Textract 提供了建立文件主題及註解的能力。文件一開始以字串的方式輸入，Textract 再逐字的建立其主題的格式（如how long），潛在的詞性及順位並在偵測到感興趣的字詞時為其標示上對應的註解，例如將Bill Clinton 標示為PERSON，January 標示為TIME。

(3)文件索引的方法

利用一個稱為Resporator 的模組作為確認出文件中可能的解答詞組並對應QA-Token 為其註解。索引的程序會將Textract 取得的字串逐一建立索引並加入由QA-Token 所取得的分類標註。

(4)查詢問句的分析

問句會依照系統所建立的問題樣板被轉換為符合系統search engine 所需的格式，並針對問句的主題（wh-word 的樣式）配合QA-Token找出對應的查詢字組。

(5)Matching and Ranking

在Matching 的作法上，GuruQA 是以段落為基本的單位而非整篇的文章，系統透過一個最佳的matching 參數來控制段落的大小。這種方式的好處是讓現有的search engine 能很容易地被用在QA 系統的設計，但是潛在的缺點則是同一篇文章中可能會出現多個答案的段落。在Ranking 的方式上，不同於傳統的search engine 使用tf*idf 的作法，其捨棄了idf 而改用很普遍的權重等級，主要的作法是計算連續段落中Matching 且不重複的查詢關鍵字組。

(6)答案的選擇

GuruQA 的答案選擇主要是依據AnSel 及WerLect 這兩個演算法，其主要的作法是評算search engine 所取回的前 10 個段落中已由QA-Token 所確認的項目。評算的標準是藉由對已定義出之七個特徵值作加權計算的線性分類法則，詳細的演算法及計算的特徵值在[38,39]中有詳細的介紹。

整體而言，predictive annotation 最大的優勢是其問題解答技術使用了QA-Token 的方法來取代繁瑣的語意分析工作，並以簡易的加權

計算比對段落中的查詢關鍵字組達到答案比對的工作。GuruQA 方法沒有使用到深奧的NLP 或是任何的IE 樣板比對，也沒有使用詞性標記或任何的同義詞典及如WordNet 之類的Ontology。相對的，predictive annotation 的方式使用在fact-oriented 問題解答有明顯的效用，但弱勢則是無法作到了解文件意涵、同義字取代及更深層知識分析等的工作。

參、XR^3

Michael Laszlo等[33]使用淺顯的文件處理技術設計出一個名為 XR^3 的答案擷取系統。這些技術的出發點是每一個問題的目標型態都明顯地描繪出了候選問題的屬性。在實際的作法上可歸納出下列幾點：

- 將文件視為一連串的text windows

一個window 是在一文章中可被找到之任何一個連續250 字的字串，並且是儘可能在其中心最少包含有一個關鍵詞。在去除前後兩端及排除重覆的windows 之後，可以得到一列的windows。最後，藉由比對各windows 之間的相似度並除去相似度小於個maximum windows margin 系統參數的windows，得到最後代表此一文章的windows list。

- 在問句中所出現的關鍵字同樣的會出現在答案中

在早期的抽詞過程中同義字並非必然地有效用，因為動詞及副詞會因表達方式的差異而有所不同，所以過早的建立同義字檢索對取得正確答案並無太大的助益。然而，名詞及專有名詞扮演著更重要的角色，因此在萃取答案時再建立同義字的查詢會更有效用。

- 透過問題表達方式可所確認及擷取出可能的對應答案原始的文件可藉由規則性的表達式子來取出可能的配對目標（例如：who會指向人名），配合以人工建立的樣本及起動的字詞（例如：貨幣及時間的單位名稱列表）可應用在整個流程之中，特別是取出對答案的關鍵性專有名詞。
- 藉由分析問題的焦點(focus)及聯想關係(link)可提昇答案擷取的正確性問題的焦點是指在一個window 中絕對會出現的部份，如果是，則可認定此window 是包含著有效的答案。問題的聯想關係則有點像是語句的結構，XR3系統定義出六個可能的link type，分別是：existence、attribute、time、location、reason 及means，如“Who is the Queen of Holland?”這個問句的link type 是existence，”Where is Inoco based?”的link type 是location。焦點及聯想關係都是由pattern matching 所決定，問題可視為由stopwords（例如：介系詞、行列式、代名詞及一般的動詞像是“is”及“does”）所鏈結的一連串段落，這些stopwords 會暴露出文件中的語意關

係及可靠的patterns。

作者從相關的工作中推測出，透過建立一個有明顯區隔的專有名詞概念目錄及definition-based 的問題對提昇QA 系統在問題解答有非常實質的效用。

Robert Gaizauskas 等在[22,44]中描述了一個結合了IR 系統及NLP 系統的QA-LaSIE Q&A 系統。主要是利用IR 系統從大量的文件中預先取出相關的的文件或是段落集合，接著NLP 系統再以語法分析的方式從集合中選出候選的答案。QA-LaSIE 以一個含有18 條法則的question grammar 來處理輸入的查詢語句，並將為一個問句建立成為一個”quasi-logical form”(QLF)作為語法解析器在分析語句時所用。在找尋答案時為找出符合事件型態的類別，系統則利用了Wordnet所建立的ontology 來作為一般廣泛性常識分類的參考架構。

在系統的作法上其將每一個問題視為一個query 並傳遞給IR 系統，IR 系統再從文件集合中取出排位最高的幾個文件或是段落，接著再將之傳送給IE 系統作答案的篩選工作。在這裡IR 系統被視為是IE 系統的過濾器，因為IR 系統原本的設計就是為了處理大量的資料，而IE 系統雖可比較精確的處理文句的分析，但是速度相對的就非常的緩慢而且無法針對每一個查詢一一的處理。所以在作業的順序上，一開始可以利用IR 系統處理資料的廣度及速度，再加上IE 系統

解決問題的深度來提昇整體系統的效能。

肆、本節結論

從上述針對解決Q&A 問題的相關研究中可整理出下列幾個重要的結論：

- 以question type 的shallow language understanding 取代繁重的NLP 工作。

(1) 藉由問題的型態及分類可找出可能的答案配對。

(2) 排除了繁雜的自然語言處理工作。

(3) 藉由建立更完善的問題型態與答案型態的配對，可更有效提昇答題的正確性。

- 以windows or passage 的方式來取代整篇文章的比對更能找出正確的答案。

(1) 答案的關鍵字詞通常只出現在文章中的某一段，即在全文檢索中的IDF 值在Q&A 系統中通常被忽略，因此如果只針對文件中的某一段落來找尋答案可排除因在全文檢索所造成的失焦現象。

(2) 當某文件的許多段落中皆出現相同的關鍵字，則可認定此文章非常有可能是答案。

- 利用問題的焦點(Focus) 直接的指出了使用者所關心的議題。

(1) 從問題的型態可看出問題的焦點

(2) 焦點代表了問題的核心

- knowledge-based 的方法可作為問題概念分類的依據。

(1) 一般常識性的問題可藉由如Wordnet 知識庫所建立之ontology 來完成知識分類的工作。

(2) 特殊領域的問題可藉由domain-specific ontology 來完成專業知識的分類工作。

(3) domain-specific ontology 的建立及應用在知識管理相關工作上愈來愈受重視。

(4) 配合domain-specific ontology 及Wordnet 知識庫可達成系統 domain-independent 的目標。

利用專有名詞詞庫及分類目錄可有效的提昇解答的準確率。

(1) 問題的焦點通常是某一個特定的人、事、時、地、物。

(2) 配合專有名詞的檢索可分辨出文件是屬於某一個特定的領域。

- 結合IR 系統及IE 系統的優點可平衡解答的回應率及準確率。

(1) IR 技術主要是為了處理大量文件而設計，因此可視為是一個強大的文件過濾器。

(2) IE 技術是為了準確的取得某一關鍵事件而設計，因此作為核心事件特徵的辨識器。

(3) 結合前端的IR 系統及後端的IE 系統，可在初步取得大量的相關

文件（高回應率），並在後期得到準確的目標文件(高準確率)。

第五節 本體論(Ontology)

本體論一詞來自於哲學領域， 為表述哲學理論的術語，根據Smith & Welty的研究[30]，兩位哲學家認為哲學上的本體論是所有真實領域中實體(Etities in all spheres of being)的分類，而且此一分類是決定性的(Definitive)、徹底的(Exhaustive)。本體論被廣泛使用於電腦科學的各個領域之中，包括人工智慧、自然語言處理、資訊檢索、知識管理等等。各領域對於本體論有著不同的論述與用途。

在電腦科學領域中，本體論被發展為用來提供機器可處理的資訊來源， 並供不同的代理人(Agent)之間溝通。在過去數十年中， 有許多關於本體論的定義被提出，其中最常被引用的是Gruber[23]於1993年所提出的定義：「本體論是對於群體共享的概念化之正式的、明確的表示形式」。根據Gruber的定義，「概念化(Conceptualization)」是指對現存的某個現象或領域的確定現象之相關概念抽象模型；「共享(Shared)」是指本體論是一個共享的部分，屬於群體而非個人；「正式的(Formal)」是指本體論是機器可以讀的、可以理解的；「明確的(Explicit)」是指本體論的概念形態及限制以明確的方式表示出來。

第六節 關聯規則(Assocation Rule)

關聯規則（Association Rule）探勘方式首先由Agrawal[41]等人於

1993 年首先提出，其目的在於對於大量的交易資料中找尋出隱藏的有用資訊以利企業決策運用。由於這種資訊將提供客戶消費的趨勢、傾向與喜好，所以，若能夠充分掌握顧客行為便能為企業帶來更多商機。關聯規則的定義描述為：假定Ｉ＝｛Ｉ1, I 2，…，Ｉn｝為項目之集合，由Ｉ中的子集合所組成的集合則稱之為交易，這些交易紀錄了有哪些項目被購買，並且會給予每個交易一份唯一的編號，稱為tid。一群項目的組合稱為項目組(Itemset)，若項目組中的個數為k，則稱該項目組為k-項目組(k-itemset)。下表2-7為一交易資料庫片段，每筆交易都有一份唯一的Tid編號，其中T10是一個由4 個項目所組成的4-itemset。

表 2-6 關聯法則交易資料庫資料表

Tid	Itemset
T10	ACDE
T20	CD
T30	BCE

在關聯規則問題一般都分為兩個部份探討，第一部分為找出所有支持度大於使用者設定的最小支持度閥值項目組，這些項目組都稱為高頻項目組（Frequent Itemset ,Large Itemset）。而支持度的計算則如下：

Support(i) = 項目I 出現次數頻率/資料庫中交易筆數。

在產生高頻項目時，若出現的頻率過低，這樣的項目組對使用者來說是沒有意義的，我們有興趣的會是較常出現的項目集。第二階段為產生關聯規則，我們將會計算Confidence，公式如下：

Confidence(A->B) = A,B 同時在資料庫中出現的次數/項目A在資料庫中出現次數。

例如：(牛奶->麵包，支持度(support)=0.2，信心水準(confidence)=0.8)，此關聯規則的意思是「大多數的消費者購買牛奶同時也會購買麵包」，這個購買的組合出現的比例為20%，而購買牛奶的消費者將有80%的機率會去選擇購買麵包。利用這些資訊可以在大賣場重新安排商品位置，將顧客較可能同時購買的物品放在鄰近的位置，以達到刺激消費者消費以及增加企業利潤。

第七節 中文詞知識庫小組及中文斷詞系統

壹、中文詞知識庫小組

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[7]，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WWW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，本計劃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-建構本體、語言及常識知識庫

研究如何自動化擷取語言知識及一般常識，我們期望由計畫中發展的語言處理技術配合擷取的知識能自動的分析WWW中的大量文本，從中抽取知識。

知識建構是一件耗時費事的大工程，我們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。我們將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。我們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，我們著眼於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，我們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。我們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」。未來，我們會繼續朝知識邏輯與推理、知識結構整合，及自動推理與定理證明方面努力。

三、知識應用-知識為本的中文語言處理技術

我們將注重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。以上步驟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、中文斷詞系統

詞是最小有意義且可以自由使用的語言單位。任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理，例如機器 翻譯、語言分析、語言了解、資訊抽取。因此中文自動分詞的工作成了語言處理不可或缺的技術。基本上自動分詞多利用詞典中收錄的詞和文本做比對，找出可能包 含的詞，由於存在歧義的切分結果，因此多數的中文分詞程式多討論如何解決分詞歧義的問題，而較少討論如何處理詞典中未收錄的詞出現的問題。

由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以

盡列所有的中文詞。當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成分詞系統 因為參考詞彙的不足而產生錯誤的切分。為了解決這個問題，最有效的方法是補充領域詞典加強詞彙的搜集。因此新的詞彙或關鍵詞的自動抽取成為分詞的先期準備 步驟。領域關鍵詞彙多出現在該領域的文件中而少出現在其它領域，因此抽取關鍵詞時多利用此特性。高頻的關鍵詞比較容易抽取，少數低頻的新詞不容事先搜集， 必須線上辨識。構詞律、詞素、詞彙及詞彙共現訊息，為線上新詞辨識依據。本系統提供了一個解決方案，可以自動抽取新詞建立領域用詞或線上即時分詞功能。為 一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統。此一系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞 依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。除了基本詞彙庫外，使用者可依需要附加領域專屬詞庫。詞類標記為選擇 性功能，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類。分詞系統採用之詞典俱可擴充性，使用者可依據不同領域文件，補充以領域詞典做為分詞之用。

中文斷詞作法依序可分為以下幾個步驟：

1.初步斷詞

2.未知詞偵測

3.中國人名擷取

4.歐美譯名擷取

5.複合詞擷取

6.bottom-up merging algorithm

7.重新斷詞

經過初步斷詞(使用連續三詞的長詞優先演算法)後，絕大多數的未知詞會被斷成較小的單位，即此未知詞的詞素，我們希望在接下來的步驟中， 將這些詞素重新組合成未知詞。經觀察99%的未知詞其詞構當中至少會有一個單字的詞素。因此希望利用未知詞偵測這個步驟去判定初步斷詞後的單字哪些是詞素，哪些是獨用詞彙，之後的擷取步驟再根據這樣的資訊，把處理的焦點放在這些判定為詞素的單字身上，看看是否能和其相鄰的token來合併成未知詞。我們針對一些特定類型的未知詞，如：中國人名，歐美譯名，複合詞等作了詞構分析及簡單的語言模型，剩餘的其它未知詞則交給bottom-up merging algorithm做最後的擷取動作。目前為止，所擷取出來的pattern稱為"未知詞候選者"，參考這些"未知詞候 選者"搭配原始的辭典再做一次斷詞可得到最後的結果。重新斷詞的目的是在於：有時某一個未知詞可能在文章某處並沒有擷取出來，但是在文章的 另一處被擷取出來，所以重新斷詞可以使兩處都能有正確的斷詞結果。另外，未知詞

候選者彼此間有時會有covering或overlapping這種衝突的現象，重新斷詞能夠幫助我們在這些有衝突的未知詞候選者之中挑選出一個正確答案作為最後確定的未知詞。

中文斷詞系統對文字內容進行標記處理後之詞類標記對照表於本章節中表2-4中表示，其主要將詞彙藉由中央研究院現代漢語平衡語料庫進行詞累標示。

第三章　圖形化語意網轉換機制

本研究架構如圖 3-1 所示，主要包括語意分析機制(Semantic Analysis Mechanism)、語意網轉換機制(Semantic Net Transformation Mechanism)及意圖轉換機制(Intention Transformation Mechanism)，三機制主要幫助我們產出問題意圖(Intention)、關鍵字及對應字結合。

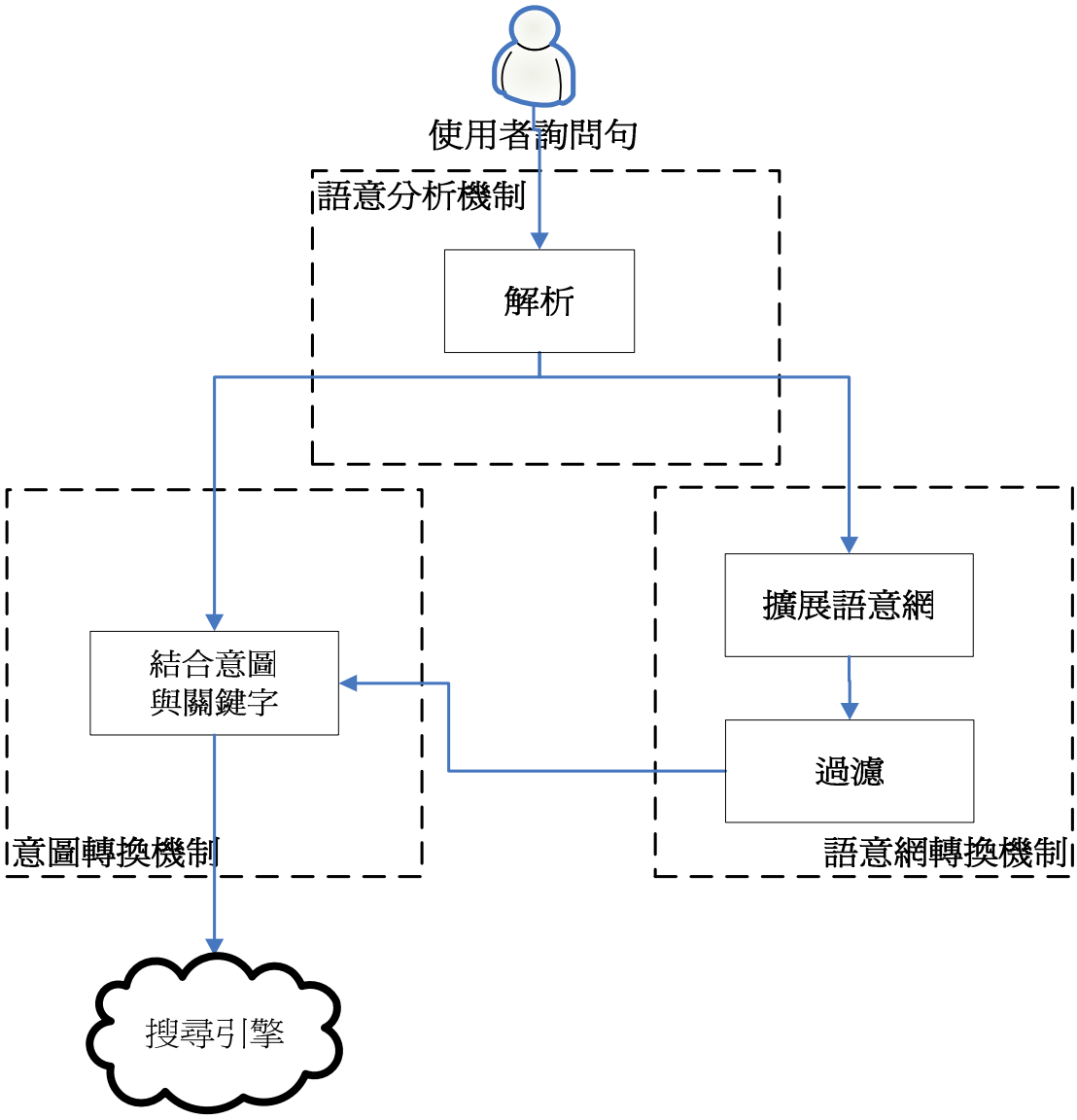


圖 3-1 圖形化語意網轉換機制架構

本研究提出一以 5W1H 結合領域本體論做資料探勘，以圖 3-1 來表示，透過意圖及語意網的結合，提高搜尋的準確性，以解決使用者之問題。本機制除了提供一個使用者介面及供使用者輸入問題及查閱搜尋結果的回應外，它是以語意分析機制(Semantic Analysis Mechanism)、語意網轉換機制(Semantic Net Transformation Mechanism)及意圖轉換機制(Intention Transformation Mechanism)三部分為核心。語意分析機制主要將使用者描述的問題，利用自然語言處理與斷詞(Natural Language Processing and Segmented)及詞性標記技術處理，透過前處理(Pre-Processing)解析後將 5W1H 意圖及關鍵字萃取出來。語意網轉換機制主要是將拆解後關鍵字與意圖利用本體論來進行擴展成 5W1H 語意網。意圖轉換機制主要將關鍵字與 5W1H 意圖資料庫中找出所對應的資料結合，經轉換後於網路搜尋引擎上做查詢。

第一節　語意分析機制

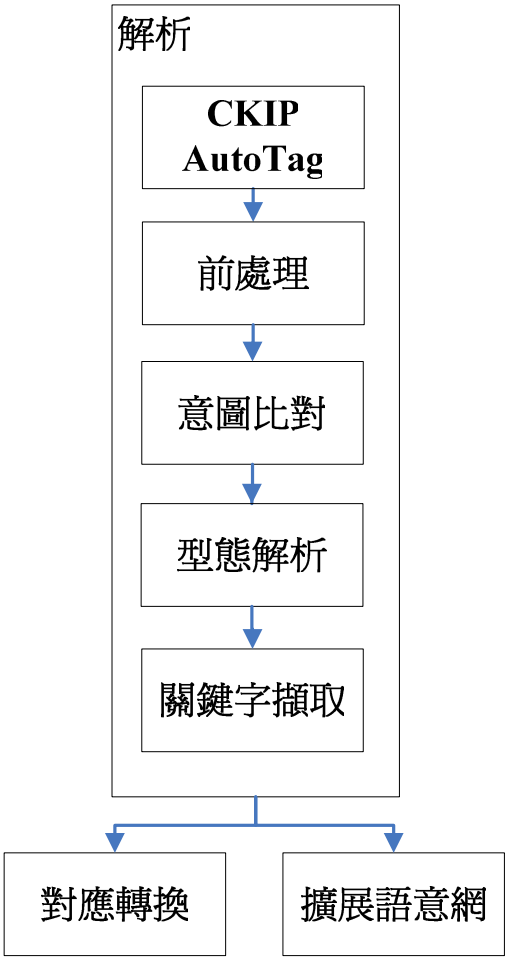


圖3-2 語意分析機制

本機制如圖 3-2 所示，主要包含五步驟，將使用者的問句利用中研院 CKIP[6]做自動斷詞及標記詞性、意圖類型比對及關鍵字擷取，其步驟主要功能如下：

1. CKIP(Chinese Knowledge Information processing)自動標記

主要透過中研院 CKIP 系統做斷詞及詞性標記，我們將以一問句

例子「如何學習除法」來說明，下圖 3-3 主要說明我們將「如何學習除法」直接輸入於系統介面上，我們會得到回傳的訊息為圖 3-4 所示可做點選多個項目，我們所需的回傳訊息為「包含未知詞的斷詞標記結果」可按下直接看到回傳的訊息，為三個部分「如何(D) 學習(VC) 除法(Na)」，如圖 3-5 所示。

圖 3-3 中央研究院中文斷詞系統介面步驟圖[5]

圖 3-4 中央研究院中文斷詞系統解析步驟圖

圖 3-5 中央研究院中文斷詞系統解析後步驟圖

2. 前處理(Pre-Processing)

本研究透過 CKIP 的斷詞及詞性標記處理後，會產生一未斷詞完成的狀況，我們將用一例子「什麼是四則運算」來做說明，以圖 3-6 表示，會斷成「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」五部

分，這樣就無法達到萃取到關鍵字的目的，那要如何萃取到我們所需的關鍵字，本研究將透過資料庫 SQL 語法規則及語意文法的設計來達到萃取關鍵字之目的，如圖 3-7 資料庫所示，我們會將設計好的規則定好後，會將「四(Neu) 則(Nf) 運算(VC)」三部分，在畫面呈現部分會直接結合成「四則運算」，可讓本研究達到目的，在資料庫部分仍是不足，期望做更多的增加，供未來做更深入的研究。

圖 3-6 中央研究院中文斷詞系統解析未完全步驟圖

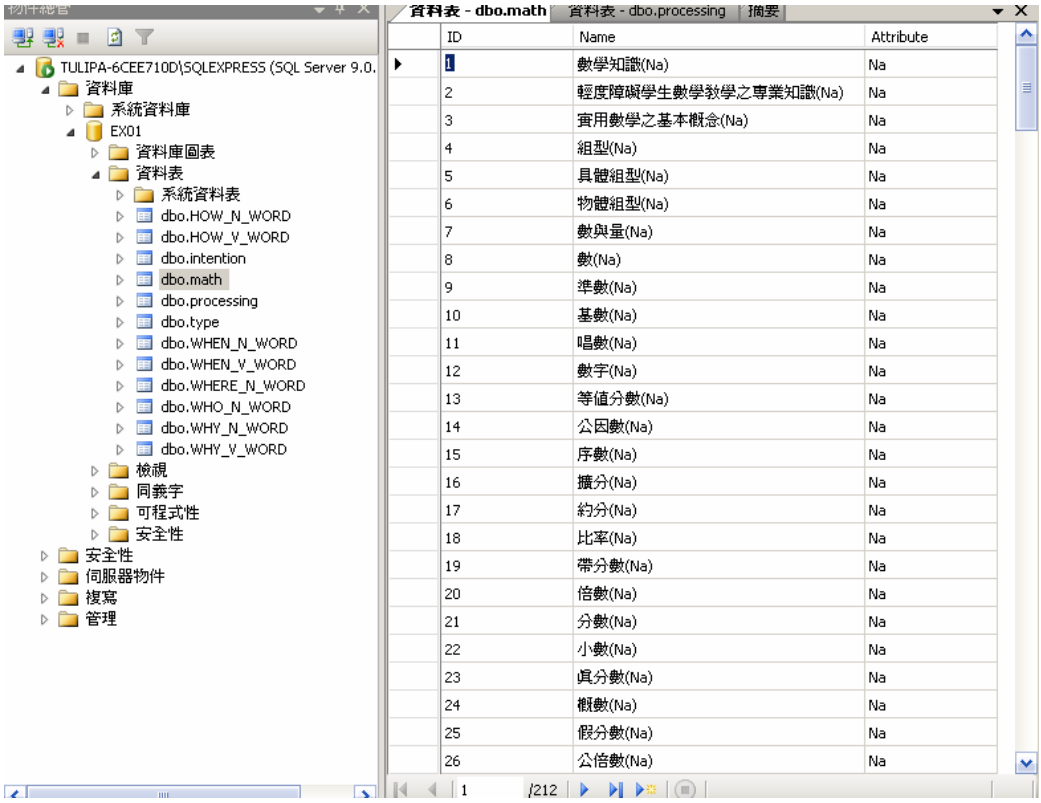


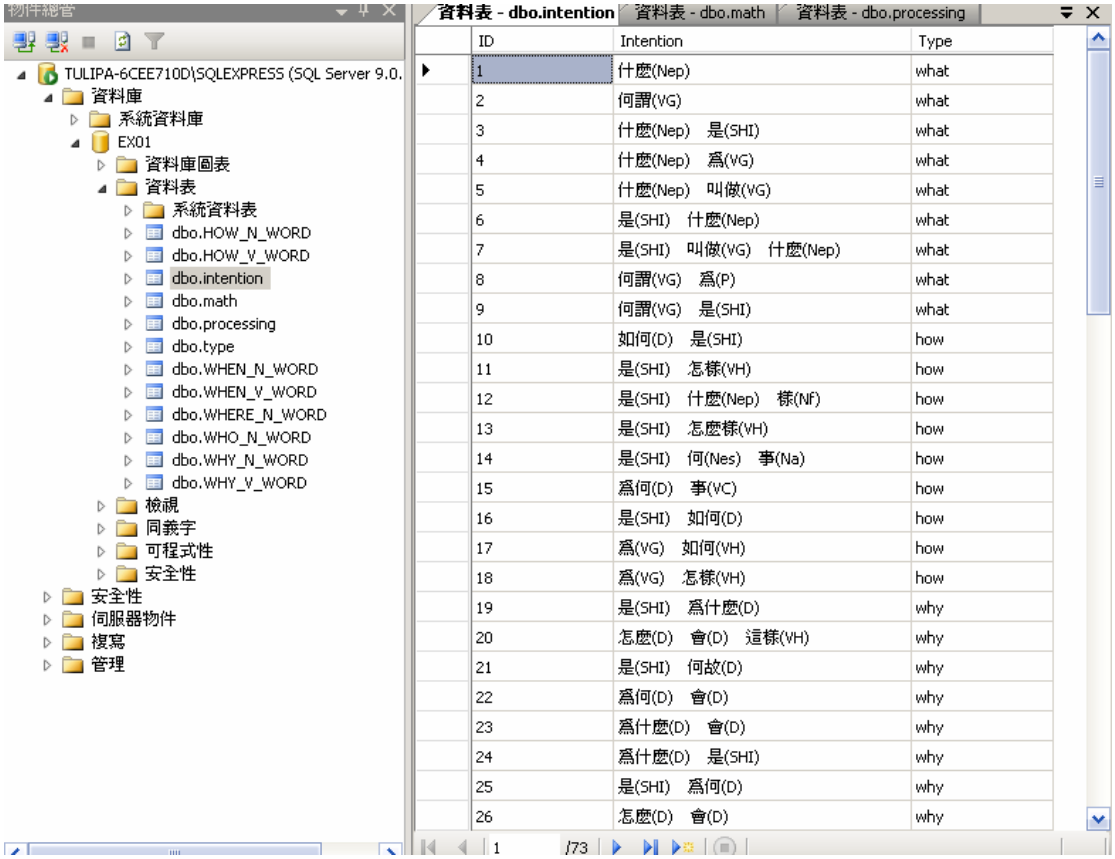
圖3-7 SQL測試資料庫之領域詞庫圖

3. 意圖比對(Intention Matching)

本研究經前處理後，再將透過設計建構後的資料庫來做比對，我們將以上一例子「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」做說明，我們會將建構後的資料庫如下表 3-1 與 CKIP 處理過後的句子做意圖比對，將例子中的「什麼(Nep)」與資料庫如圖 3-8 所示，我們將會比對出其意圖屬「What」，其詞庫由楊宸彥及牛維娟[3] [14]之研究論文改良後歸納出下表 3-1 及圖 3-8 所示。

表 3-1 5W1H 同義詞資料表-例子

what	how	why	Who	when	where
什麼	如何是	是為什麼	誰	是何時	在哪
何謂	是怎樣	怎麼會這樣	何人	在什麼時候	在何地
什麼是	是什麼樣	是何故	何等人	在什麼時間	在哪裡
什麼為	是怎麼樣	為何會	什麼人	於幾時	什麼地方
什麼叫做	是何事	為什麼會	是誰	在什麼階段	什麼地點
是什麼	為何事	為什麼是	為誰	在何時	在何處
是叫做什麼	是如何	是為何	誰是	在幾時	在哪邊
何謂為	為如何	怎麼會	是何人	於何時	在何方
何謂是	為怎樣	為什麼要	何人是	何時是	哪裡是
	哪些		是什麼人	何時為	是哪裡
			誰會	幾點是	是何方
			有誰	是幾點	是何處
			誰要	是什麼時候	是什麼地方
			是有誰	是什麼時間	是什麼地點
				是幾時	是哪裡
				是何時	是在哪



ID	Intention	Type
1	什麼(Nep)	what
2	何謂(VG)	what
3	什麼(Nep) 是(SHI)	what
4	什麼(Nep) 爲(VG)	what
5	什麼(Nep) 叫做(VG)	what
6	是(SHI) 什麼(Nep)	what
7	是(SHI) 叫做(VG) 什麼(Nep)	what
8	何謂(VG) 爲(P)	what
9	何謂(VG) 是(SHI)	what
10	如何(D) 是(SHI)	how
11	是(SHI) 怎樣(VH)	how
12	是(SHI) 什麼(Nep) 樣(Nf)	how
13	是(SHI) 怎麼樣(VH)	how
14	是(SHI) 何(Nes) 事(Na)	how
15	爲何(D) 事(VC)	how
16	是(SHI) 如何(D)	how
17	爲(VG) 如何(VH)	how
18	爲(VG) 怎樣(VH)	how
19	是(SHI) 爲什麼(D)	why
20	怎麼(D) 會(D) 這樣(VH)	why
21	是(SHI) 何故(D)	why
22	爲何(D) 會(D)	why
23	爲什麼(D) 會(D)	why
24	爲什麼(D) 是(SHI)	why
25	是(SHI) 爲何(D)	why
26	怎麼(D) 會(D)	why

圖3-8 SQL測試資料庫之5W1H同義詞庫圖

4. 型態解析(Type Parsing)

本研究歸納出5W1H之六種型態模型，如下圖3-10至3-20所示，再以Knuth-Morris-Pratt(KMP) [18]演算法做型態解析，以建立有限自動機(Finite Automata)與透過設計建構後的語意型態庫做快速掃瞄，我們以處理後的上述例子「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」做說明，來判斷出型態類型及做驗證其意圖。

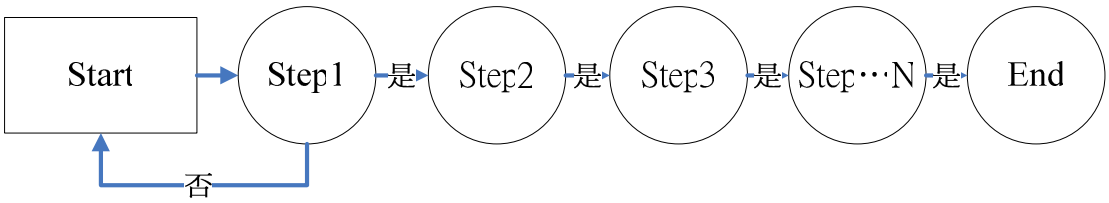


圖3-9 有限自動機流程圖

我們將以「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」例子與圖3-9來做明說流程，主要處理流程是比對詞性為主，而步驟一我們會先去判斷(Nep)是否與資料庫相符合，如相符合的話，將再到步驟二做判斷(SHI)是否與資料庫相符合直到比對步驟結束，此目的是為了可加快比對的速度及系統效能。

以下有本研究所提出的5W1H之六種型態簡易模型[32]，來提供比對驗證意圖之用，包含How、What、Who、Where、What、Why，主要透過網路上使用者的問句，所歸納出來，如下圖所示：

第一種型態How

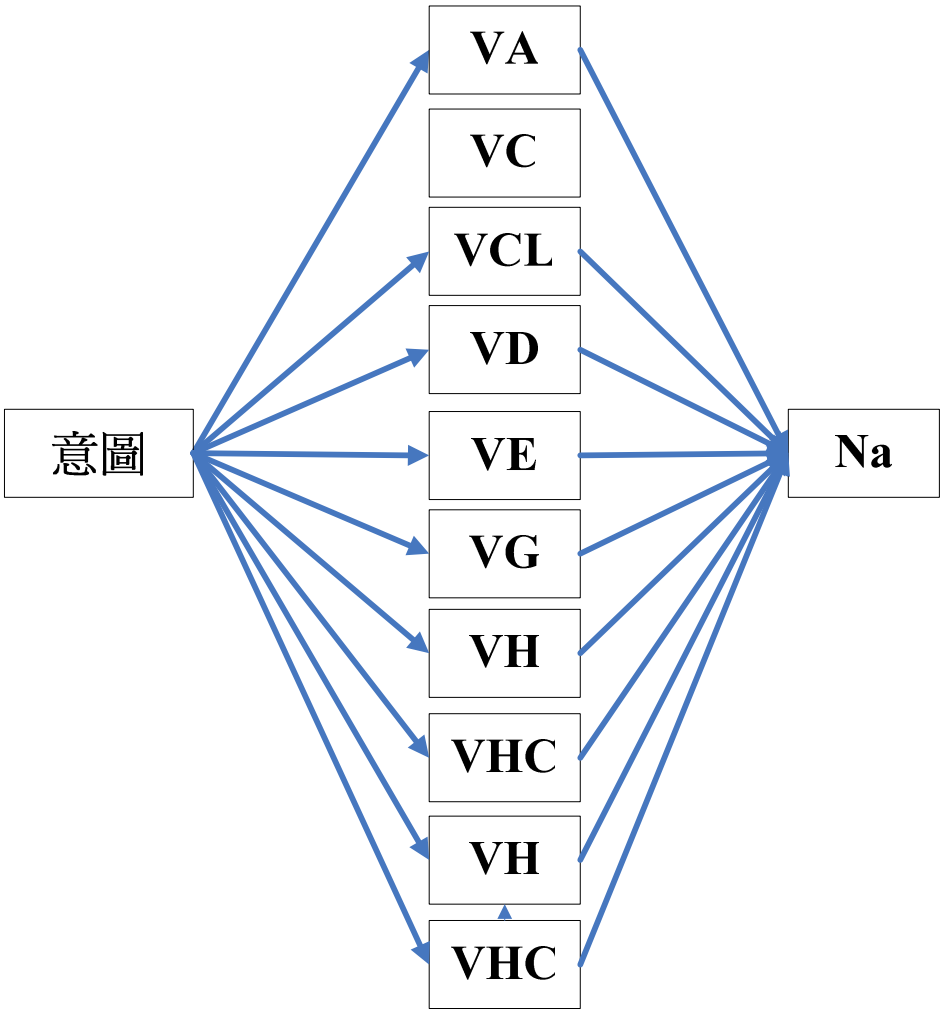


圖3-10 類型How之第一型

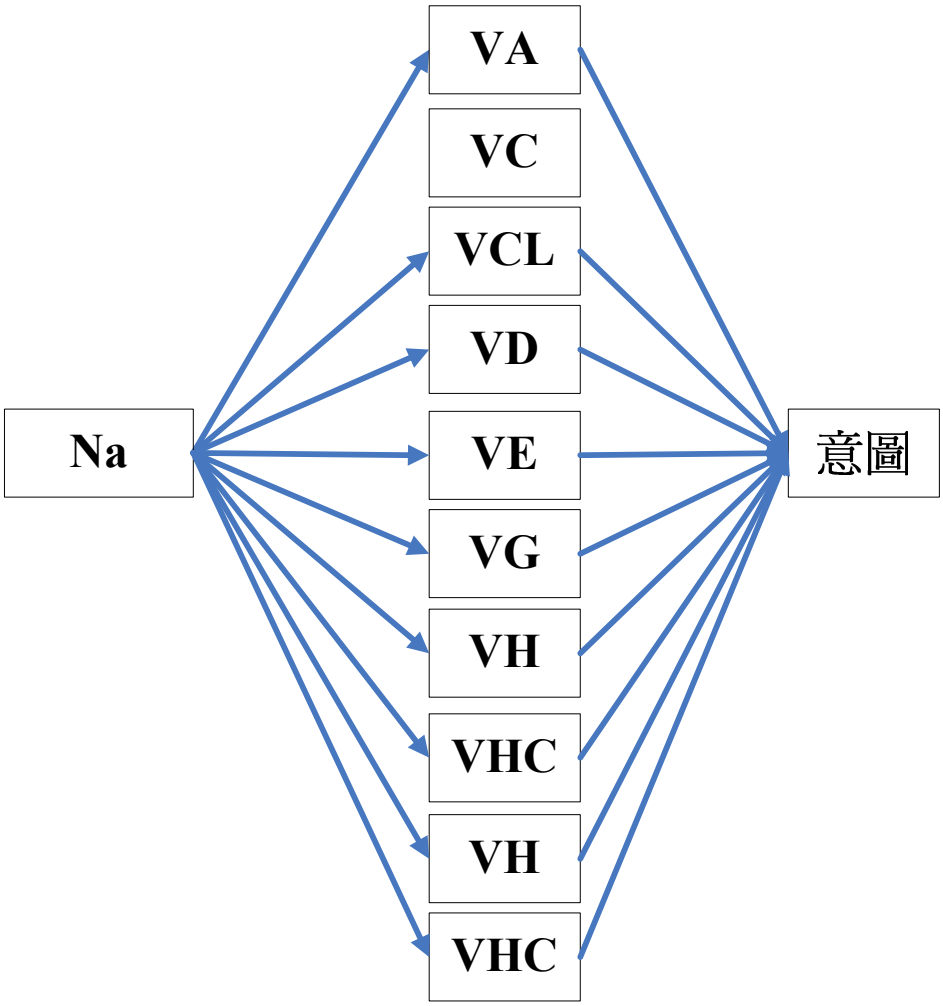


圖3-11 類型How之第二型

第二種型態What

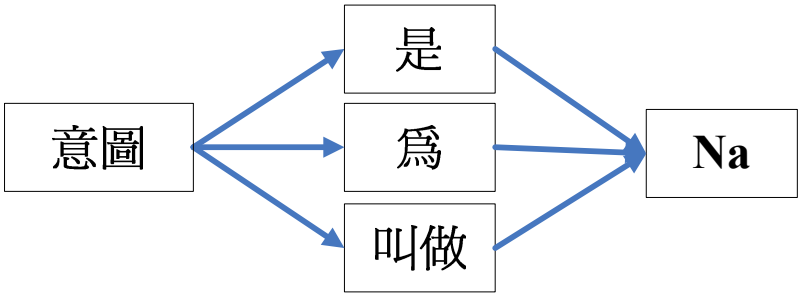


圖3-12 類型What之第一型

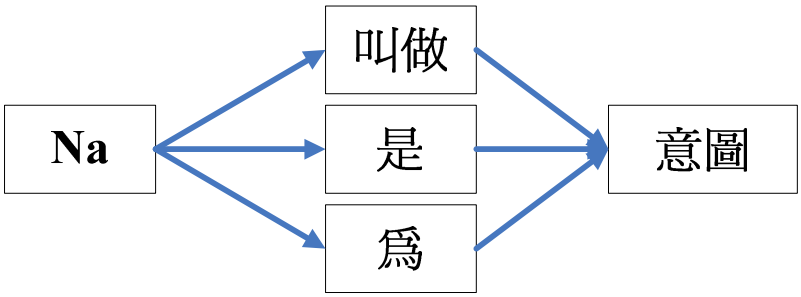


圖3-13 類型What之第二型

第三種型態Who

圖3-14 類型Who之第一型

第四種型態Why

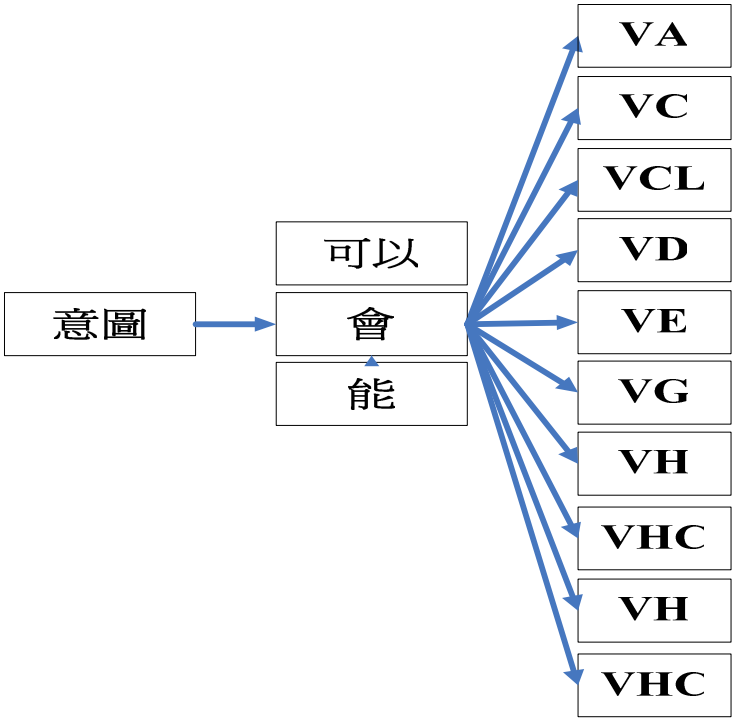


圖3-15 類型Why之第一型

第五種型態When

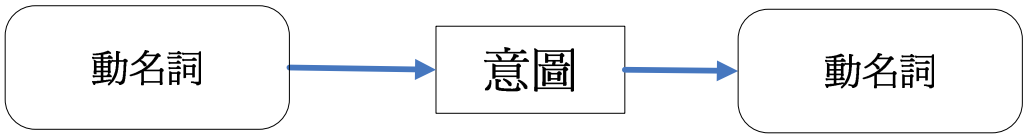


圖3-16 類型When之第一型

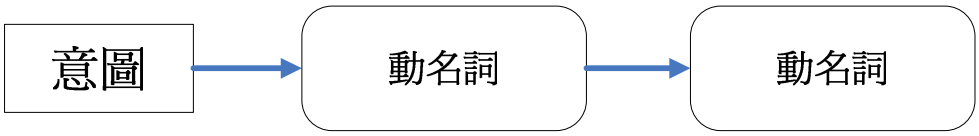


圖3-17 類型When之第二型

圖3-18 類型When之第三型

第六種型態Where

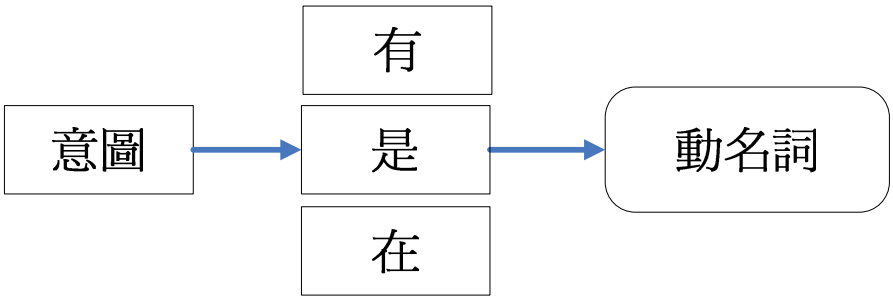


圖3-19 類型Where之第一型

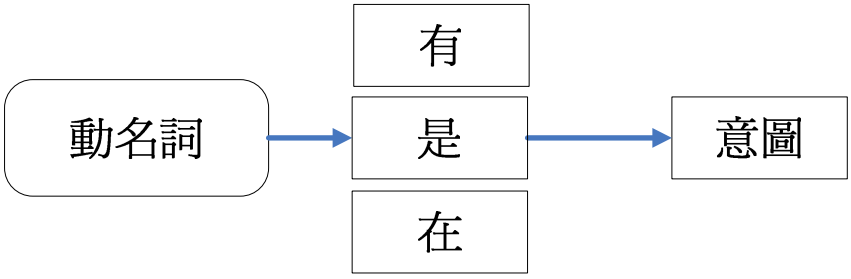


圖3-20 類型Where之第二型

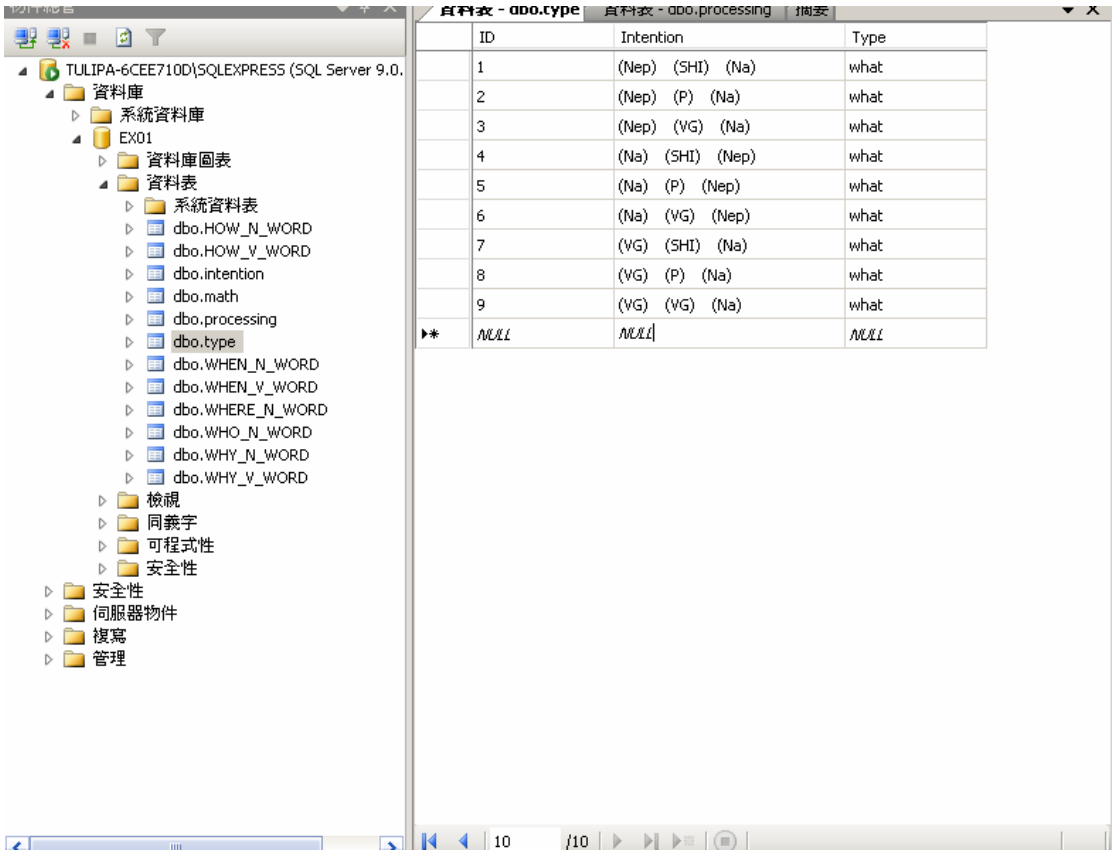


圖3-21 SQL測試資料之語意型態庫圖

上圖如上述所示，我們將以「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」為例子做說明之，以比對詞性「 (Nep) (SHI) (Neu) (Nf) (VC)」為主，來判斷出類型及驗證其意圖為「What」，可供使用者在系統頁面上的判斷，是否為我們所問之意圖。

Kunth-Morris-Pratt演算法：

輸入：P及T，字樣及本文字串；f 陣列(失敗鏈陣列)如果T的長度事前不可預知，在演算法使用它的部分可以用一個end-of-string測試來取代。P的長度應該在設定f 陣列時已求出。

輸出：T字串中P的一份拷貝之開始註標。如果找不到可與P 匹配者，註標值將為T_L+1。

```
function KMPmatch ( P, T : String; f : IndexArray ) : index;
var
    j , k : index;
   { j indexes text characters;
    k indexes the pattern and f array. }
begin
    j := 1; k := 1;
    while j ≤ T_L and k ≤ P_L do
          if k = 0 or j k t = p then
              j := j +1; k := k +1
          else { follow fail arrow }
              k := f [k]
          end { if }
       end { while };
       if L k ≤ P then L KMPmatch := j − P { match found }
       else KMPmatch := j { := +1 L j T , no match }
       end { if }
end { KMPmatch }
```

公式 3-1 KMP 演算法

5. 關鍵字擷取(Keyword Extract)

在這步驟也是以一「什麼(Nep) 是(SHI) 四(Neu) 則(Nf) 運算(VC)」為例子，經處理過後的斷詞與設計建構後的資料庫做比對，將

關鍵字四(Neu) 則(Nf) 運算(VC)比對出來成為「四則運算」，以提供下一步驟做擴展語意網。

第二節 語意網轉換機制

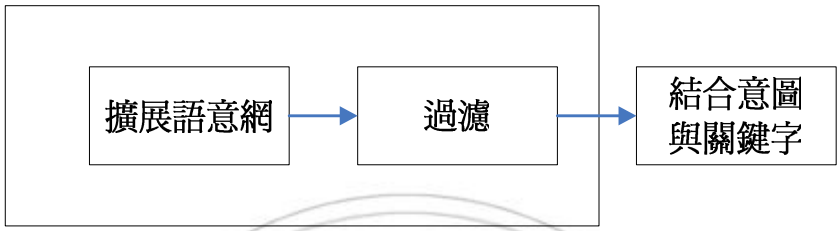


圖3-22 語意網轉換機制

本機制包括兩步驟如圖 3-22 所示，將包含名詞「國小數學」等的關鍵字，透過設計建構後的本體論 Ontology「幾何」、「數與量」、「公式應用」等關聯來直接比對出更多關聯字出來，以擴展成語意網。其說明如下：

1. 擴展語意網(Extended Semantic Net)

我們將以一關鍵字「國小數學」例子做說明，被延伸後的關鍵字如「幾何」、「數與量」、「公式應用」等，利用本體論關聯做延伸相關概念出來，形成語意網，最多可延伸至第七層，下圖例子只延伸至第四層。以下圖 3-23 為例。

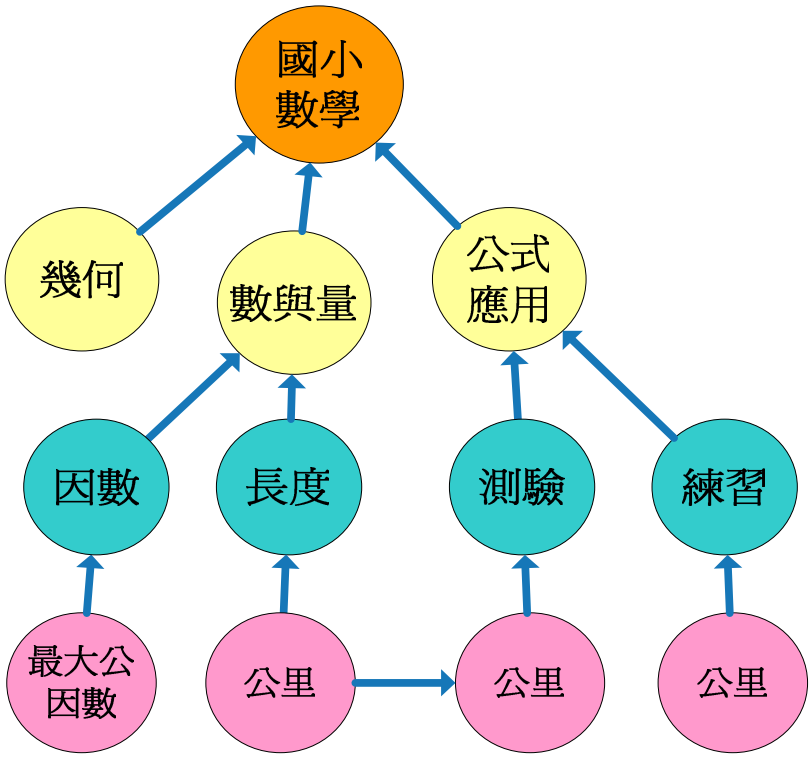


圖3-23 國小數學領域之本體論測試架構圖

2. 過濾(Filtering)

由於透過本體論關聯來做延伸，如果延伸太多層的話，會造成資訊過多的情形及關聯性較無相關，且造成系統效能負擔，所以本研究在系統上將設立一參數門檻值來做過濾，如下圖 3-23 及圖 3-24 來做對照過濾一層資訊。

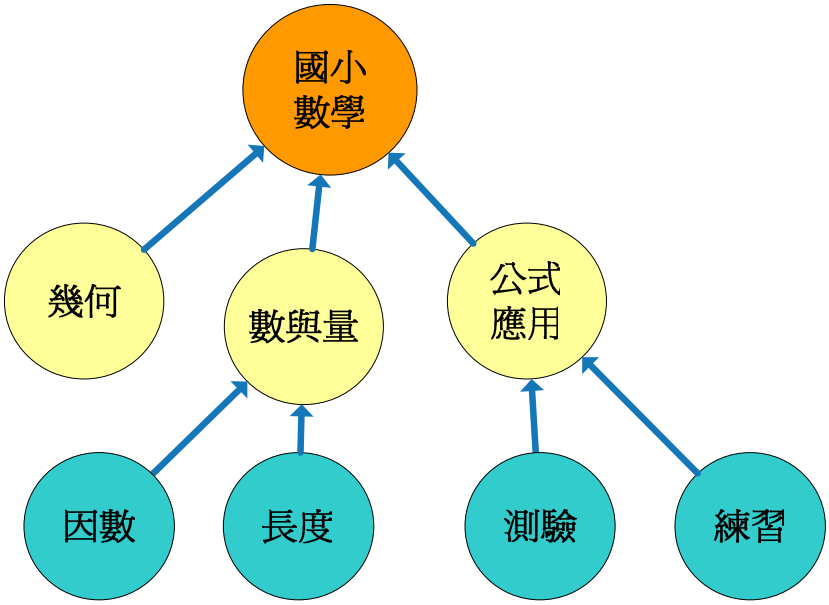


圖3-24國小數學領域之本體論過濾測試架構圖

第三節 意圖轉換機制

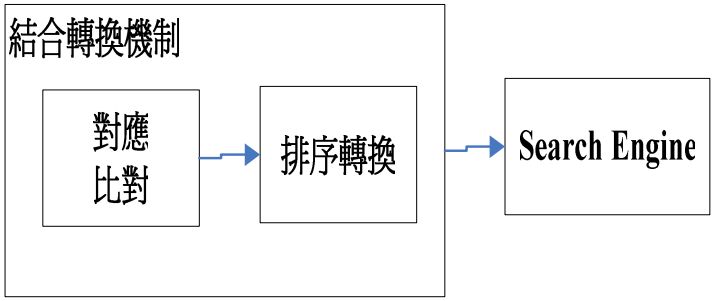


圖3-25 意圖轉換機制

本機制有兩步驟，目的是將意圖與關鍵字結合後，以 5W1H 對應詞庫中的「對應詞」對應出來，再轉換成一般句型及過濾詞性後於搜尋引擎做資料搜尋。

1. 對應比對(Corresponsively)

透過5W1H對應詞庫例子如表3-3的比對，我們根據J. Han and K. Kambe[28]所提之關聯法則探勘-Apriori演算法，如何將對應字找出

來，我們以下面表3-2例子做說明。

表3-2 意圖Why之支持度與信心水準計算表

關鍵字	對應字	支持度	信心水準
智能障礙	因子	25/30=0.833	(16/25)/(25/30)=0.768

表3-2以意圖Why為例，將關鍵字「智能障礙」及對應字「因子」兩個關鍵字，同時於搜尋引擎做網頁搜尋，我們會搜尋30篇樣本，來做計算出其支持度及信心水準。而支持度就是30網頁中同時出現「智能障礙」及「因子」的次數有25篇。而信心水準部分則是在這25篇的網頁中對應字「因子」所出現的次數有16篇，所以我們算的信心水準為0.768。

表 3-3 5W1H 對應資料-各意圖之例子

意圖類型	Keyword 例子	對應詞
why	智能、障礙	來源、源由、理由、原因、起源、介紹、因素、因子、成因、發生、發現、過程
	肚子	
	頭痛	
	心臟病	
	腸病毒	
when	小孩、爬、開始	日期、起源、時間、幾月、

	煮菜、放油	多大、時期、時機、時間點、發生、發現、發源、發跡
	種、台灣、稻米	
	澆、花	
	學習、注音	
what	長期、記憶	介紹、定義、認識、情況、狀況、解釋
	語言、發展	
	血壓	
	肝癌	
	聽覺障礙	
how	計算、等腰三角形面積	解法、答案、公式、算法、教法、方法、步驟、題庫、題材、類型、題目、方式、用途、公式、流程、辦法、情況、狀況、問題、認識、學習、教、算、輔導、輔助、運算、計算、練習、求得、求、解、教導、解釋
	製作、網頁	
	做、蛋糕	
	製造、機器人	
	去、九份	
where	去、北投	發源地、原點、原處、發生

	南華大學	地、原生地、出處、方向、距離、位置、地點、方位、地址
	台北捷運站	
	離、台北火車站	
	台北、101、走	
who	12 屆、中華民國、總統	發明人、原著、作者、發明者、創作人、創造人、名字、姓名、姓氏、介紹、解釋
	著作、三國志	
	發明、飛機	
	發明、電燈	
	微軟、創辦人	

（1）對應字之關聯法則探勘

根據J. Han and K. Kambe所提之關聯法則探勘-Apriori演算法，其目的在於對於大量的交易資料中找尋出隱藏的有用資訊，以此概念做為以關鍵字來搜尋出隱藏的5W1H對應字。例如：顧客購買印表機時，他們是否也會購買報表紙?或購買IBM PC主機時是否會傾向於搭配ViewSoinc螢幕。

本研究透過透過此概念，於網路搜尋引擎做文件搜尋，先將預設方式將關鍵字及各意圖之對應詞，於網路搜尋 30 篇相關的網頁資

料，以作為本研究之測試樣本，利用此樣本進行 Apriori 演算法計算，將關鍵字及隱藏之 5W1H 對應字的支持度及信心水準計算彙整出來。

以上表3-3信心水準來判斷，以關鍵字智能障礙與各對應字都有滿高的信心水準，也符合其關聯法則之條件。

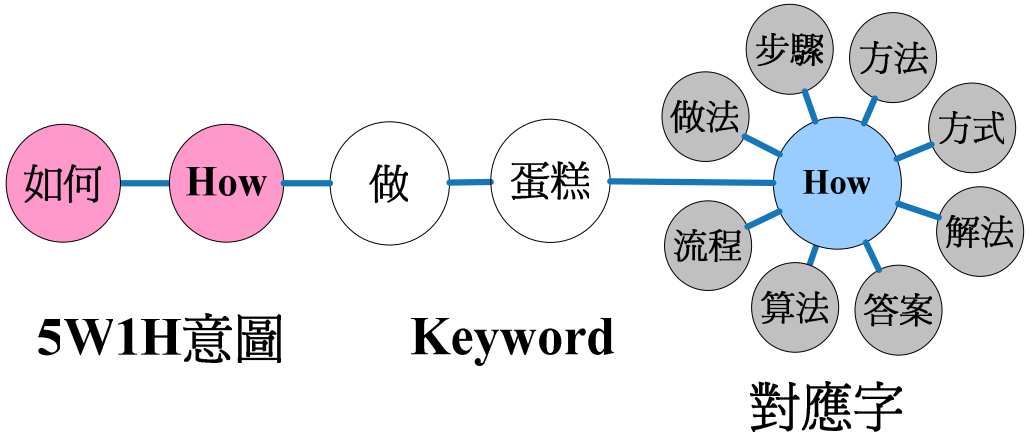


圖3-26 意圖How之對應例子示意圖

上圖所示，我們會將會以一處理過後的例子「如何做蛋糕」做說明，將比對後的意圖How及關鍵字「做」與「蛋糕」，於意圖對應資料庫中比對出屬於How的對應字，其包含有「做法」、「步驟」、「方法」、「流程」等等，將與關鍵字做結合後成兩關鍵字、三個關鍵字及四個關鍵字。

```
L1=find_frequent_1-itemsets(D);
for (k=2; Lk-1≠φ ;k++) {
    Ck=apriori-gen (Lk-1,min_sup);
    for each transaction t∈D{ //scan D for counts
        Ct=subset(Ck,t); //get the subsets of t that are candidates
        for each candidate∈ Ct
            c.count ++;
    }
    Lk={c∈ Ck|c.count >= min_sup}
}
return L=∪k Lk;

Apriori_gen (L k-1:frequent (k-1)-itemsets; min_sup:minimum support
threshold)
    for each itemset l1∈ Lk-1
        for each itemset l2∈ Lk-1
            if (l1[1]=l2[1])^ (l1[2]=l2[2]) …….^ (l1[k-2]=l2[k-2]) ^
              (l1[k-1]<l2[k-1])then {
              c= l1∞l2;//join step: generate candidates
              if has_infrequent_subset(c, Lk-1) then
                  delete c; //prune step:remove unfruitful candidate
              else add c to Ck;
             }
      reutn Ck;

procedure has_infrequent_subset (c: candidate k-itemset; L k-1:frequent
(k-1)-itemsets);
    //use prior knowledge
    for each (k-1)-subset s of c
        if s  L k-1 then
            return TRUE;
return FALSE;
```

圖3-27 Apriori 演算法[28]

演算法一開始先找出1-頻繁項目組，進入Apriori_gen演算法，透過Apriori_gen來產生後選項目組，由Lk-1產生Ck。對產生的候選項

目，我們會利用has_infrequent_subset演算法來做修剪的動作，以節省搜尋時間。

2. 排序轉換(Ranking Transformation)

經上一步驟後，從資料表中對應出來的「對應詞」配合關鍵字來做排序轉換，本研究經過文件訓練後如「學習」+「四則運算」＋「算法」+「方法」+「解法」＝「名詞」＋「對應字」，透過王亭雅[1]所提之方式可做直接結合轉換來表示。

以下圖3-28例子表示：

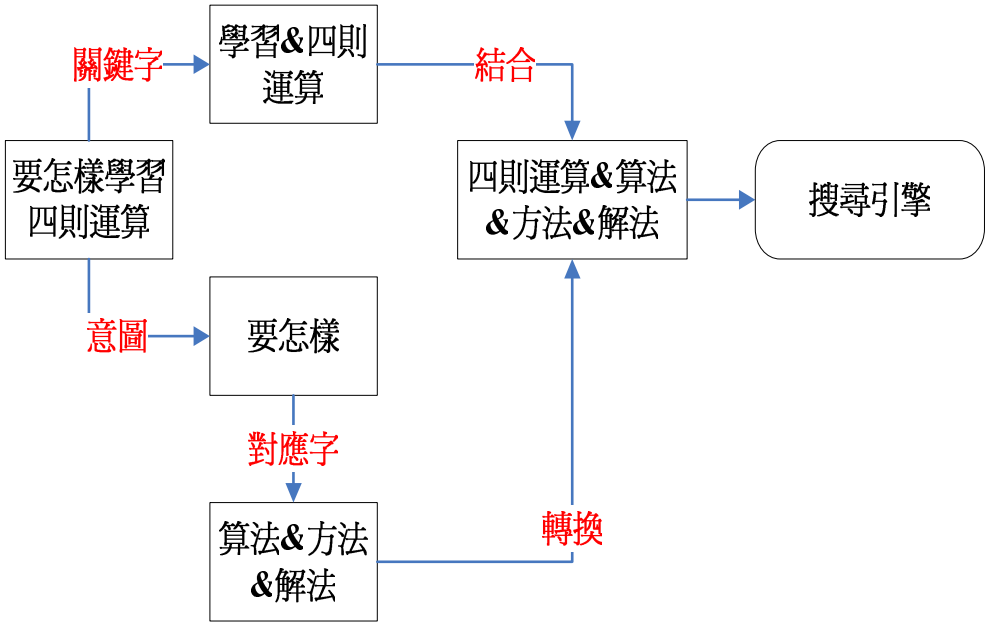


圖3-28 意圖轉換示意圖

第四章、系統開發與實作

依據第三章所設計的架構，本研究設計發展對應比對與轉換機制，本機制主要處理關鍵字與意圖擷取、對應字比對，讓使用者更快速且準確的獲取網路上最貼近原意的網頁知識。

第一節 實驗環境介紹

本實驗與測試是利用下列環境所建造，主要可分為軟體與硬體兩部份。

壹、 軟體

一、 作業系統：Microsoft Windows XP Professional Edition Service Pack 2。

二、 資料庫系統：Microsoft SQL Server 2005 Express Edition。

三、 程式語言：Java。

四、 程式開發平台：NetBeans 6.0、JDK 1.6.0、SQL JDBC 1.1.1501.101。

貳、 硬體

一、 機器：ASUS　L5000C Notebook。

二、 CPU：Intel Pentium(R)4 CPU 2.67GHz。

三、 記憶體：512MB。

實驗目的主要是找出意圖後以比對出對應字來與關鍵字一同於搜尋引擎做搜尋，為了是能更快速及準確的搜尋到所需的網路資訊，以達到透過本機制讓其效能有所提升。

第二節資料來源與限制

本研究之樣本收集，主要透過網路搜尋引擎 Google 做資料搜尋，而 Google 本身的搜排序參數是無法做改變的，所以我們就假設以各領域的 5W1H 問句來做搜尋，將關鍵字分別配合六種意圖型態『How』、『Who』、『Where』、『What』、『When』及『Why』所對應之『對應字』做網頁搜尋，先以「傳統單一關鍵字」、加上本研究所提之「第一個對應關鍵字」、「第二個對應關鍵字」及「第三個對應關鍵字」做網頁搜尋。

至於資料庫方面，則建立數學領域知識庫(Math Domain Knowledge Base)包含 305 條國小數學名詞、動詞與專有名詞；5W1H 同義詞資料庫(5W1H Synonym Rule)有 72 條規則，分為單字與多字相關詞；而在問句型態庫相關規則(Query Type Rule)有 107 條規則及本研究最核心的資料庫 5W1H 對應字庫有六種意圖共 100 條對應字。

在搜尋樣本方面，我們將依排列的前三十筆有效樣本作為本研究之測試樣本，再以黃耀民[15]論文中所提之一般常用準確率(Precision

ratio)公式來作評估。

精準率：所有搜尋樣本中，可檢索出的有相關樣本，其公式如下：

$$precision = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{retrieved\}|} = \frac{A}{A+B} \quad \text{.....公式(1)}$$

第三節 實驗結果

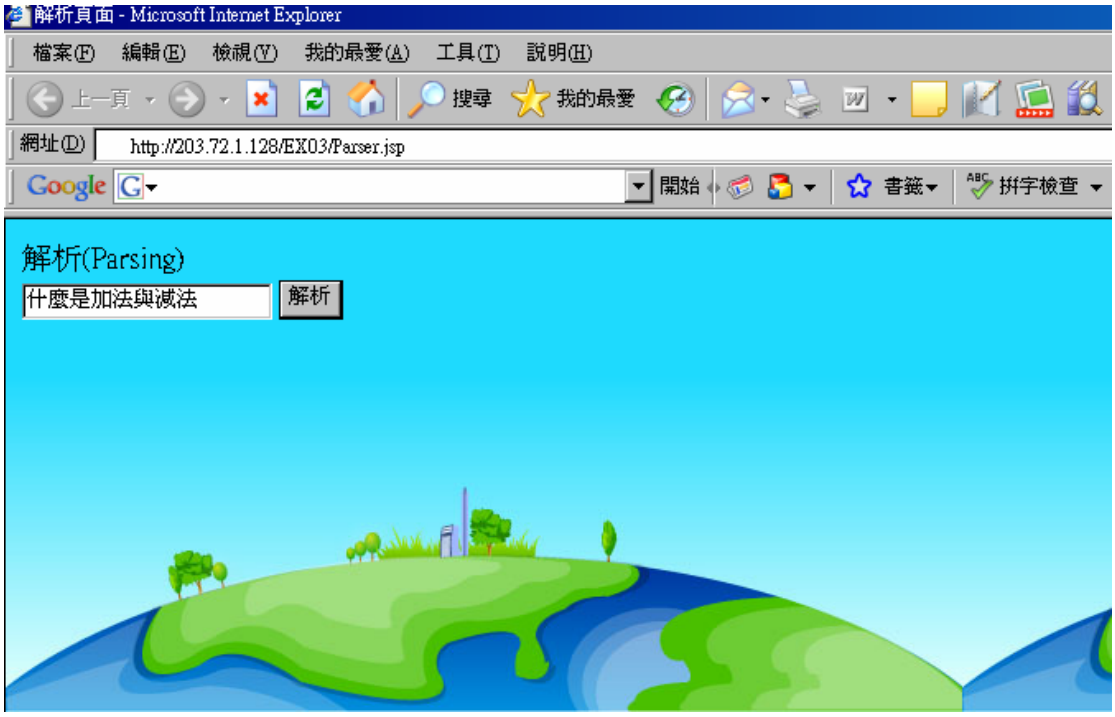


圖4-1 測試解析頁面圖

圖 4-1 表示在此系統頁面中，我們可依自然語言方式輸入一問句，以一例子「什麼是加法與減法」來做說明，再將輸入後的問句按下解析來做下一步驟的系統解析判斷。

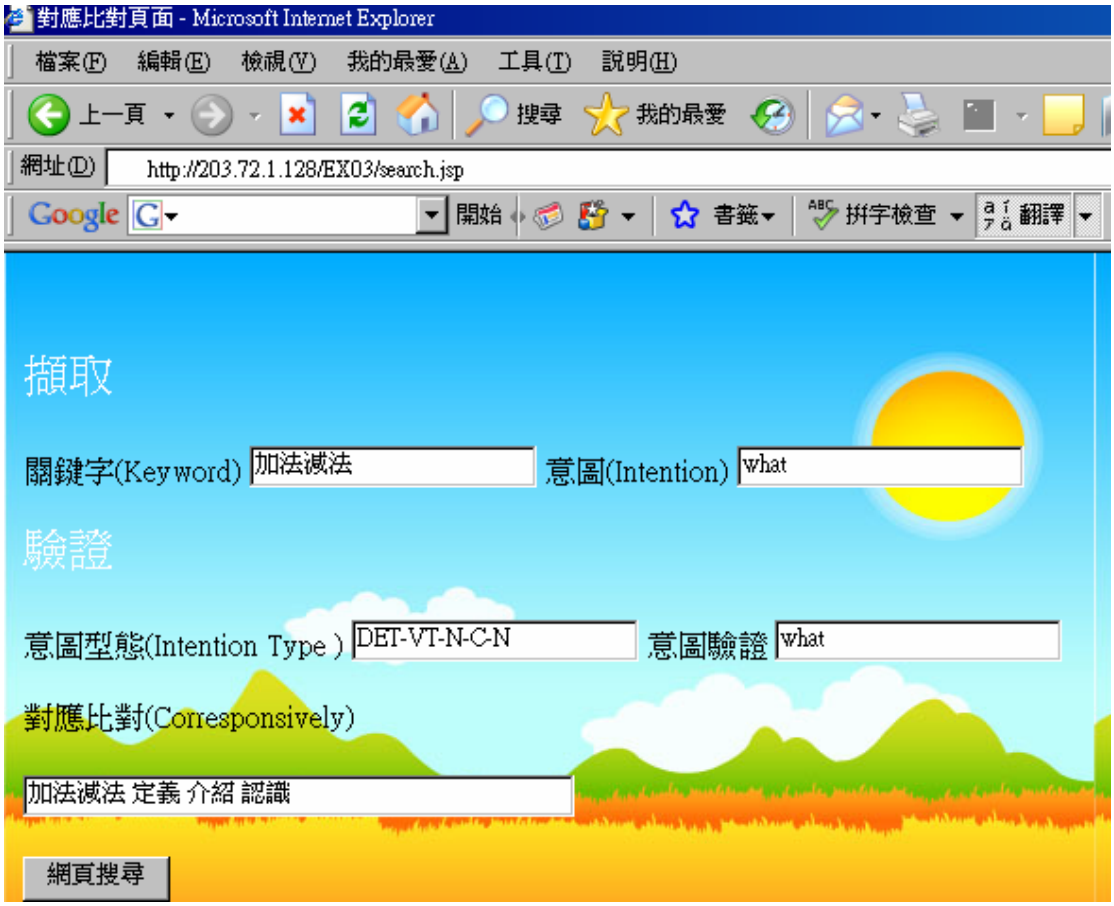


圖4-2 測試驗證對應頁面圖

圖 4-2 表示，經上一步驟的處理後，我們會將關鍵字『加法』、『減法』及意圖『what』比對出來，再透過一驗證機制將使用者問句意圖型態『DET-VT-N-C-N』比對出來後，做驗證出為「What」的意圖，再由對應比對處理後，將對應字「定義」、「介紹」及「認識」一同與關鍵字於網路搜尋引擎做網頁搜尋，期望以更快速及準確的搜尋到所需的網路資訊，以達到透過本機制讓其效能有所提升。

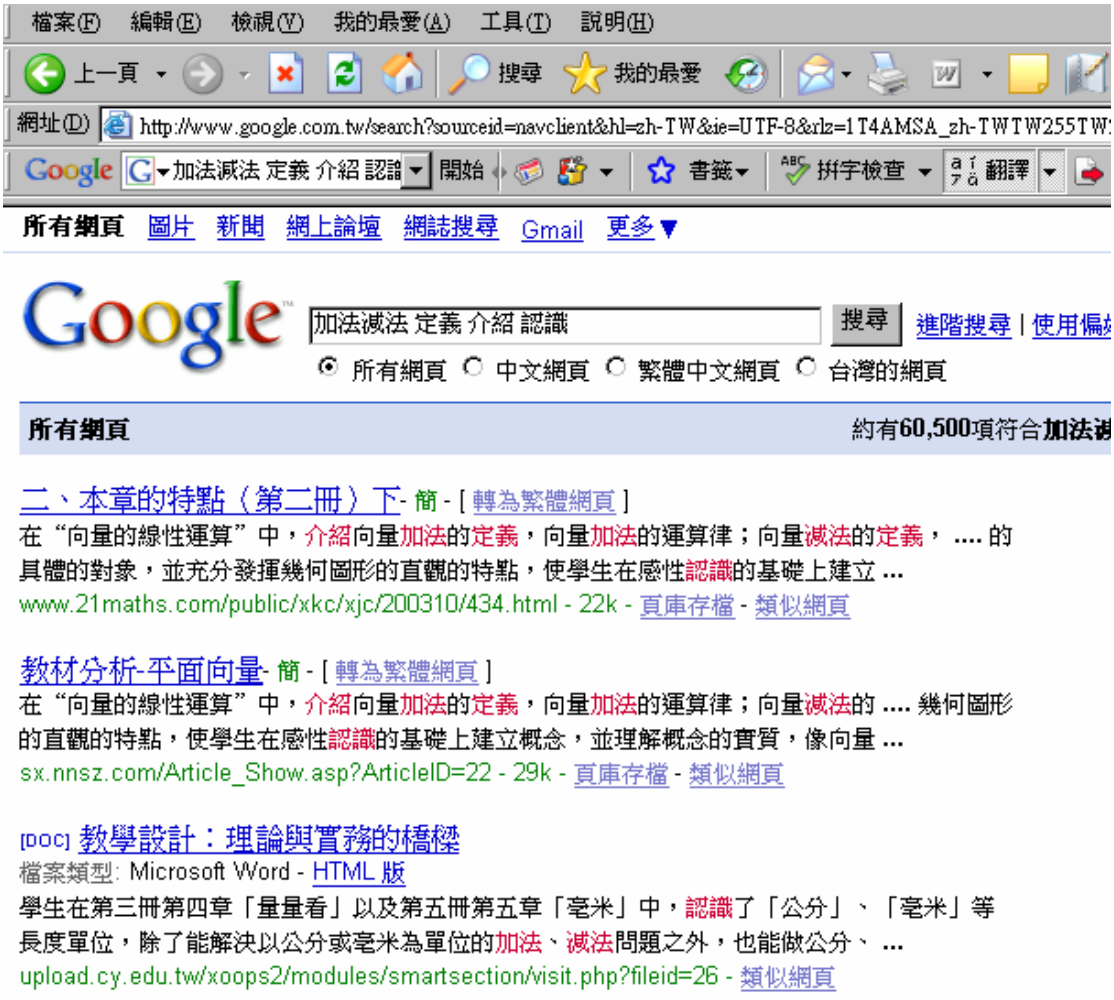


圖4-3 搜尋引擎搜尋頁面圖

圖 4-3 主要是將關鍵字「加法、減法」與對應字「定義、介紹、認識」來做結合到 Google 搜尋引擎上做資料搜尋，透過本研究所提出的對應字，與傳統單一關鍵字一同於搜尋引擎做搜尋，能縮短搜尋的時間且提高搜尋資料的準確率。

表 4-1 5W1H 意圖測試整理資料表

例子	意圖	傳統單一 關鍵字	本研究之 擴展對應詞
如何計算等腰三角形面積?	How	計算、等腰三角形面積	解法、算法、答案
要怎樣製作網頁?		製作、網頁	方法、步驟、流程
要如何做蛋糕?		做、蛋糕	做法、方法、流程
怎樣才能製造出簡易機器人?		製造、簡易、 機器人	方法、流程、步驟
要如何去九份?		去、九份	方法、方式、辦法

例子	意圖	傳統單一 關鍵字	本研究之 擴展對應詞
為什麼會有智能障礙?	Why	智能、障礙	原因、因素、成因
為何會常拉肚子?		常拉、肚子	原因、因素、成因
我為什麼會發		發生、頭痛	原因、因素、成因

生頭痛?			
他為什麼會產生心臟病?		產生、心臟病	起因、原因、因素
小孩為什麼會得到腸病毒?		小孩、得到、腸病毒	原因、介紹、成因

例子	意圖	傳統單一關鍵字	本研究之擴展對應詞
誰是 12 屆中華民國總統?	Who	第 12 屆、中華民國、總統	名字、姓名、介紹
著作三國志的人是誰?		著作、三國志	作者、原著、發明人
誰發明了飛機?		發明、飛機	姓名、介紹、發明人
電燈是由誰發明的?		電燈、發明	發明人、創造人、發明者
誰是微軟的創辦人?		微軟、創辦人	姓名、名字、介紹

例子	意圖	傳統單一關鍵字	本研究之擴展對應詞

小孩什麼時候開始學爬?	When	小孩、開始、學爬	時間、多大、幾月
煮菜要何時開始放油?		煮菜、開始、放油	時間點、時機、時間
台灣稻米何時開始播種?		台灣、稻米、開始、播種	時間、時期、日期
百合什麼時候可以澆水?		百合、澆水	時間點、時機、時間
國小何時開始學注音?		國小、開始、學、注音	時間、時期、時間點

例子	意圖	傳統單一關鍵字	本研究之擴展對應詞
走哪邊可以去基隆?	Where	去、基隆	地點、位置、方向
南華大學在什麼地方?		南華大學	地址、地點、地方
那邊可以去台北捷運站		去、台北捷運站	地點、位置、方向

離台中火車站有多遠?		離、台中火車站	距離、地點、方向
台北 101 在何方?		台北、101	方向、位置、方位

例子	意圖	傳統單一關鍵字	本研究之擴展對應詞
什麼叫做聽覺障礙?	What	聽覺、障礙	定義、認識、介紹
何謂長期記憶?		長期記憶	定義、認識、介紹
什麼是語言發展?		語言、發展	介紹、認識、定義
高血壓是什麼?		高血壓	定義、認識、介紹
何謂為肝癌?		肝癌	認識、介紹、定義

上表 4-1 主要以六種意圖內包含了五個個別不同的測試例子來做說明，透過處理過後的例子來得到傳統單一關鍵字及本研究之擴展對應字，分別以傳統單一關鍵字與本研究所擴展的對應字分別做加入於搜尋引擎做網頁搜尋。

表 4-2 意圖 How 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
How	0.433	0.533	0.567	0.567
	0.167	0.5	0.667	0.7
	0.1	0.533	0.733	0.867
	0.167	0.333	0.367	0.4
	0.5	0.533	0.5	0.5
平均值	0.2734	0.4864	0.5668	0.6068

以上表 4-2 一與表 4-1 以「How」的例子「如何計算等腰三角形面積?」做說明，經處理後會產出兩部分傳統單一關鍵字「計算、等腰三角形面積」及對應字「解法」、「算法」、「答案」，而傳統單一關鍵字就是直接於搜尋引擎做搜尋後做準確率及平均值計算，在兩個關鍵字部分則是「計算、等腰三角形面積」加上「解法」一同於搜尋引擎做搜尋，而三個關鍵字部分則是「計算、等腰三角形面積」加上「解法」、「算法」，在四個關鍵字部分則是「計算、等腰三角形面積」加上「解法」、「算法」、「答案」一起做網頁資料搜尋。下表 4-3 至 4-7 以此類推來做表示。

表 4-3 意圖 What 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
What	0.333	0.6	0.733	0.833
	0.267	0.5	0.567	0.633
	0.367	0.467	0.6	0.7
	0.3	0.467	0.567	0.567
	0.6	0.667	0.733	0.833

平均值	0.3734	0.5402	0.64	0.7132

呈上述的說明，表 4-3 的準確率平均值乃是四個關鍵字 0.7132 為最高值，比傳統單一關鍵字做網頁搜尋的準確率平均值 0.3734 來的高，平均差異值可達 0.340 左右。

表 4-4 意圖 When 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
When	0.333	0.367	0.5	0.6
	0.167	0.2	0.233	0.267
	0.1	0.133	0.267	0.233
	0.133	0.267	0.333	0.367
	0.1	0.267	0.233	0.367
平均值	0.1666	0.2468	0.3132	0.3668

而表 4-4 的準確率平均值 0.3668 則是以四個關鍵字為最高值，傳統單一關鍵字的準確率平均值 0.1666，平均差異值可達 0.2 左右。

表 4-5 意圖 Where 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
Where	0.333	0.5	0.6	0.533
	0.467	0.6	0.633	0.667
	0.6	0.633	0.733	0.533
	0.3	0.4	0.433	0.5
	0.333	0.433	0.5	0.367
平均值	0.4066	0.5132	0.5798	0.52

在表 4-5 的準確率最高平均值 0.5798 在意圖 where 中是以三個關

鍵字為最高，比起傳統單一關鍵字的準確率平均值是 0.4066，而平均差異值可達 0.179。

表 4-6 意圖 Who 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
Who	0.333	0.533	0.5	0.433
	0.067	0.1	0.167	0.167
	0.267	0.333	0.4	0.3
	0.4	0.433	0.433	0.467
	0.5	0.533	0.567	0.467
平均值	0.3134	0.3864	0.4134	0.3668

在表 4-6 中以三個關鍵字的平均準確率值最高為 0.5798，在意圖 who 裡，比起傳統單一關鍵字的準確率平均值 0.3134，在平均值方面較無明顯差異只差 0.1 左右的值。

表 4-7 意圖 Why 準確率平均表

意圖	傳統單一關鍵字	兩個關鍵字	三個關鍵字	四個關鍵字
Why	0.4	0.5	0.533	0.6
	0.067	0.3	0.333	0.4
	0.4	0.633	0.733	0.767
	0.333	0.6	0.667	0.733
	0.267	0.5	0.5	0.567
平均值	0.2934	0.5066	0.5532	0.6134

表 4-7 說明在意圖 why 中，四個關鍵字的準確率平均值為 0.6134 與傳統單一關鍵字的準確率平均值 0.2934 來高的許多，且平均差異值可達 0.320。

下圖 4-4 及 4-5 表示，傳統單一關鍵字與本研究所提出的擴充 5W1H 對應字之平均準確率成長圖。

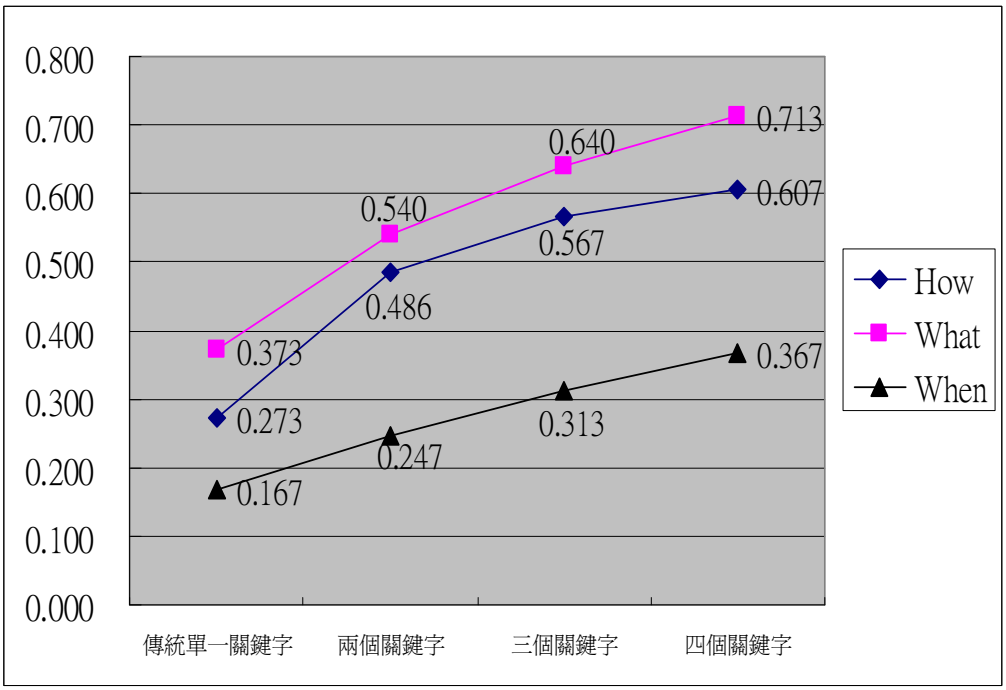


圖 4-4 How、What 及 When 平均準確率成長圖

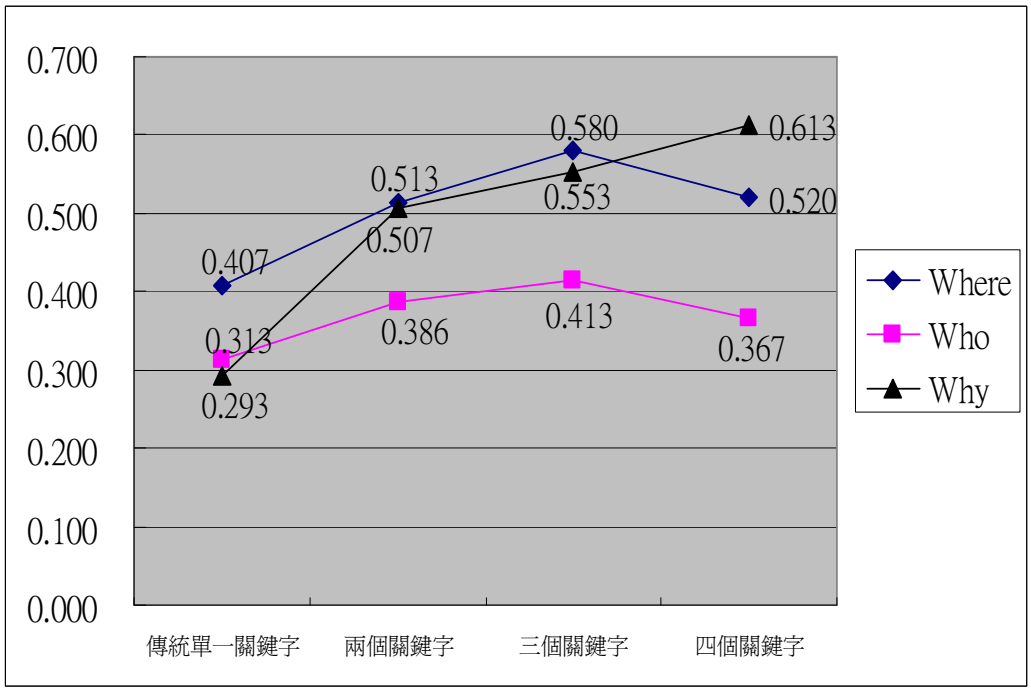


圖 4-5 Where、Who 及 Why 準確率成長圖

表 4-8 5W1H 意圖比較差異平均表

意圖	兩個關鍵字	三個關鍵字	四個關鍵字
How	0.213	0.293	0.333
What	0.167	0.267	**0.340 最高**
When	0.080	0.147	0.200
Where	0.107	0.173	0.113
Who	0.073	0.100	**0.053 最低**
Why	0.213	0.260	0.320

表 4-8 主要依本研究所提出的對應字加入後，將搜尋資料計算後，與傳統的單一關鍵字做準確率之差異，根據表 4-8 表示，每一意圖會依所搜尋的領域不同，會提升 5.3%至 34%的準確率。以圖 4-6 來表示各意圖差異。

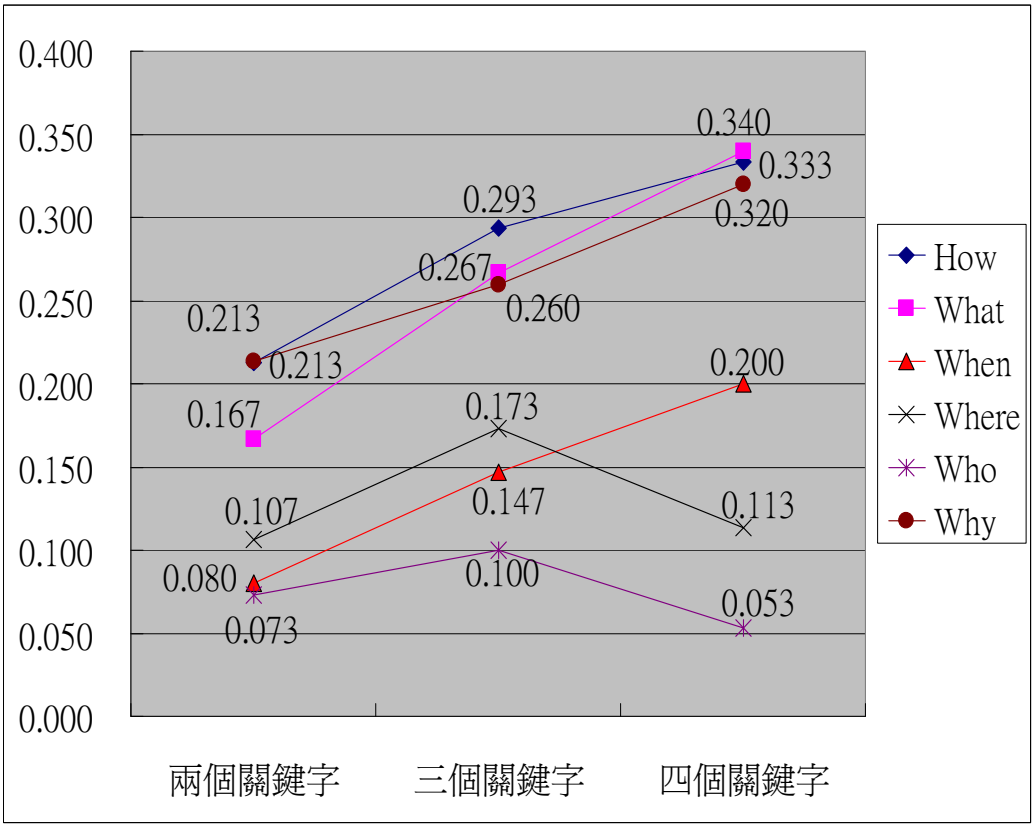


圖 4-6 5W1H 意圖平均成長差異圖

圖 4-6 主要表示將表 2 至表 7 的(「四個關鍵字」、「三個關鍵字」、「兩個關鍵字」各平均值)－(傳統單一關鍵字平均值)＝各平均值成長差異製成圖表方式，在 What 的部分最高平均值為 0.340，而在最低的平均值為 0.053 則是屬 Who 部分，可看出此平均差異性。

第五章、結論與未來展望

第一節　結論

本研究主要是在探討：自然語言的處理技術、意圖與關鍵字的擷取、擴展語意網、型態解析、問題轉換及配合多種資料庫做分析。本研究主要針對幾項問題來進行分類，從使用者問句中去擷取出意圖類型及關鍵字來進行語意擴展及結合轉換，最後透過搜尋引擎來找到更好的資料回覆給使用者，達到本研究之目的。

本研究之具體貢獻如下：

第一點：本研究歸納出 5W1H 之一般化模型，能更快速判斷出問句類型及意圖。

第二點：本研究歸納出5W1H對應字Ontology與關鍵字做結合搜尋可提高準確率。

第三點：有限自動機(Finite Automata)的建立，可改善比對的效能。

第四點：本研究所提之方法與傳統單一搜尋方法比較後，可提高5.3%~34%的準確性。

第五點：本研究所提之六種意圖的5W1H對應字，可幫助縮短搜尋的時間且可提高搜尋的準確率，為本研究最有價值的地方。

第二節　未來展望

針對於本研究所提出之架構，有以下四點可供未來做更深入研究與探討：

第一點：5W1H 的類型屬一般雛型，待未來增加更多領域類型。

第二點：在資料庫方面，未來將以訓練方式來形成一回饋機制來設定權重。

第三點：結合後的成效高低不一，待未來有更好的演算方法以作改善。

第四點：期望未來加入更多演算方法來改善，如模糊理論、基因演算法、類神經網路等。

第五點：系統將透過長時間的資料庫訓練與累積，於未來會以自動化方式做學習訓練。

參　考　文　獻

一、中文部份

[1] 王亭雅、吳昇(2003)，網頁搜尋引擎問答，國立中正大學資訊工程系碩士論文。

[2] 王聖中(1994)，語法式中文斷詞之研究，私立淡江大學資訊工程系研究所碩士論文。

[3] 牛維娟、李錫捷(2003)，應用於USENET 之Q&A 系統之研究與設計，私立元智大學資訊管理系碩士論文。

[4] 李坤霖(2000)，網際網路FAQ檢索中意圖萃取及語意比對之研究，國立成功大學資訊工程系碩士論文。

[5] 李祥賓、柯淑津(2001)，「新聞文件摘要之研究」，第十四屆計算語言學研討會論文集，pp. 23-42。

[6] 中央研究院中文斷詞系統，URL：http://ckipsvr.iis.sinica.edu.tw/

[7] 中央研究院詞庫小組，URL：http://godel.iis.sinica.edu.tw/CKIP/index.htm

[8] 吳志虹(2001)，應用關鍵頁搜尋及知識分類技術於Q&A系統之研究與設計，私立元智大學資訊管理系碩士論文。

[9] 林啟文(1993)，具可移動性中文自然語言查詢介面之研究，國立中興大學應用數學系碩士論文。

[10]周思誠，簡世杰，謝偉強，陳昭宏(2000)，電通人分機語音查詢系統，電腦與通訊，第86 期，pp.24-33。

[11]柯淑津、宋啟聖(2002)，「中文複合詞的語法結構分析及詞彙語意表達」，第三屆中文詞彙語意學研討會論文集，pp. 57-68。

[12]馬偉雲、謝佑明、楊昌樺、陳克健(2001)，「中文語料庫構建及管理系統設計」，第十四屆計算語言學研討會論文集，pp. 175-191。

[13]陳永德(1997)，中文斷詞中長詞優先、詞頻對比及前詞優先規則之使用，國立臺灣大學心理學研究所博士論文。

[14]楊宸彥(2002)，運用剖析概念圖進行中文詢答之研究，國立台灣大學資訊工程學碩士論文。

[15]黃耀民(2005)，「以子句擷取為基礎並應用於文件分類之自動摘要之研究」，台灣師範大學資訊工程研究所碩士論文。

[16]黃韻璆 (1999)，自然語言應用於銀行電話服務系統之研究，國立中興大學應用數學系碩士論文。

[17]曾慧馨、劉昭麟、高照明、陳克健(2001)，「中文動詞自動分類研究」，第十四屆計算語言學研討會論文集，pp. 253-272。

二、西文部份

[18]Baase, S., and Gelder, A. V., “Computer Algorithms ,Introduction to Design and Analysis”, Addison Wesley Longman, 2000.
[19]Baeza-Yates, R. and B. Ribeiro-Neto, Modern Information Retrieval,The ACM Press,Addison Wesley Longman Limited, Essex, England,1999.
[20]Chung-Yin Chang, “A Discourse Analysis of Questions in Mandarin Conversion,” M.A. Thesis, National Taiwan University GraduateInstitute of Linguistics, June 1997.
[21]Chen, K.H. and Chen, H.H., “Extracting Noun Phrases for Large-Scale Text Corpora: A Hybrid Approach and Its Automatic Evaluation,” Proceedings of the 34th Annual Meeting of Association for Computational Linguistics (ACL-94), pp. 234-241, 1994.
[22]Gaizauskas, R., Humphreys, K., : “A Combined IR/NLP Approach to Question Answering Against Large Text Collections”, To appear in Proceedings of RIAO 2000:Content-Based Multimedia Information Access, Paris, April, 2000.
[23]Gruber, T. R., “A translation approach to portable ontology specifications”, Knowledge Acquisition, 5(2), pp.199-220, 1993.
[24]Harabagiu, S., Maiorano, S.,: “Finding Answers in Large Collections of Texts:Paragraph Indexing + Abductive Inference”, in AAAI Fall Symposium on Question Answering Systems, pp. 63-71, 1999.
[25]Harabagiu, S., Pasca, M.: “Mining Textual Answers with Knowledge-BasedIndicators”, in Proceedings of FLAIRS-2000, Orlando FL,2000.
[26]Harabagiu, S., Pasca, M., Maiorano,“Experiments with Open-Domain Textual Question Answering”, in Proceedings of COLING-2000,Saarbruken Germany, 2000.
[27]Ide, N. and J. Véronis, “Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art”, Computational Linguistics, 24, 1, pp. 1-40,The MIT Press, London, England, 1998.
[28]J.Han and K.Kamber,Simon Fraser University,Data Mining :Concepts and Techniques, 5nd ed., San Francisco: Morgan Kaufmann, pp. 230-235,2001.
[29]Jordan, P.W., “Using Terminological Knowledge Representation Languages to Manage Linguistic Resources,” cmp-lg/9605024 (URL：http://xxx.lanl.gov/), 1996.
[30]Kao,M.Cercone, N. and Luk, W.. Providing Quality Responses with

Natural Language Interface: the Null Value Problem. IEEETransactions on Software Engineering, 14(7), 959-984, 1988.
[31] Ker, S-J. and J-N. Chen, “A Text Categorization Based on Summarization Techniques”, In Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, pp. 79-83, Hong Kong,2000.
[32] Li, Charles and Sandra A. Thompson. “Mandarin Chinese: A functional reference grammar”. Berkeley and LosAngeles University of California Press. 1981.
[33] Laszlo, M., Kosseim, L., Lapalme, G.: “Goal-Driven Answer Extraction”, The Ninth Text REtrieval Conference (TREC 9) ,2000.
[34] Mitri, M., ”Combing semantic networks with multi-attribute utility models: An evaluative database indexing method,” Proceedings, Ninth Conference on Artificial Intelligence for Applications,pp. 462,1993.
[35] Manning, C. D. and H. Schütze, Foundations of Statistical Natural LanguageProcessing, The MIT Press, London, England, 1999.
[36] Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju , R.,Rus, V “LASSO: A Tool for Surfing the Answer Net”, The Eighth TextREtrieval Conference (TREC 8),pp. 175,1999.
[37] Prager, J., Brown, E., Coden, A., Radev, D.,: “Question-answering by predictiveannotation”, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval , July 24 - 28, 2000, Athens Greece,pp.184-191, 2000.
[38] Prager, J., radev, D., Brown, E., Coden, A., Samn, V., “The Use of Predictive Annotation for Question Answering in TREC8”, The Eighth Text REtrieval Conference (TREC 8), pp. 399, 1999.
[39] Prager, J., Brown, E., Radev, D.R., Czuba, K., “One Search Engine or Two for Question-Answering”, The Ninth Text REtrieval Conference (TREC 9).2000.
[40] QA Track Specifications,URL :http://www.research.att.com/~singhal /qa-track-sepc.txt
[41] R. Agawal, T. Imielinski and A. Swami. “Mining Association Rules between Sets of Items in Large Databases,” Proc. Of ACM SIGMOD, pp. 207-216 ,1993.
[42] Ruan, W., Buerkle, T. and Dudeck, J., “Object-oriented design for automated navigation of semantic networks inside a medical data dictionary,” Artificial Intelligence in Medicine, Vol. 18, No. 1, pp. 83-103,2000.
[43] Rohini Srihari and Wei Li. A Question Answering System Supported by Information Extraction. Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00),

166-172.
[44] Scott, S., Gaizauskas, R.“University of Sheffield TREC-9 Q&A System”, TheNinth Text REtrieval Conference (TREC 9) , 2000.
[45] Smolentsev, S. V, “The identification and self-organization problems in dynamic semantic networks,” IEEE International Conference on Artificial Intelligence Systems,pp. 35-39, 2002.
[46] Smith, B. and Welty, C., “Ontology: Towards a New Synthesis”, In Proceedings of the International Conference on Formal Ontology in Information Systems, Ogunquit, Maine, USA, pp. 3-9, 2001.
[47] Salton, G., and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc, 1983.
[48] Wilks, Y. A., B. M. Slator and L. M. Guthrie, Electric Words, The MIT Press, London,England, 1996.
[49] Wei-Tien Dylan Tsai, “The Hows of Why and the Whys of How”, 台灣語言學的創造力學術研討會, 2000.
[50] Yarowsky, D. and R. Wicentowski, “Minimally Supervised orphological Analysis by Multimodal Alignment”, In Proceedings of the ACL-2000, pp. 207-216, HongKong, 2000.

附　　錄　　一

Why 意圖與對應字之支持度及信心水準計算表

關鍵字	對應字	支持度	信心水準
智能障礙	原因	30/30=1.0	(26/30)/(30/30)=0.867

關鍵字	對應字	支持度	信心水準
智能障礙	因素	30/30=1.0	(25/30)/(30/30)=0.833

關鍵字	對應字	支持度	信心水準
智能障礙	成因	30/30=1.0	(27/30)/(30/30)=0.900

關鍵字	對應字	支持度	信心水準
智能障礙	因子	25/30=0.833	(16/25)/(25/30)=0.768

關鍵字	對應字	支持度	信心水準
智能障礙	理由	24/30=0.8	(13/24)/(24/30)=0.678

關鍵字	對應字	支持度	信心水準
智能障礙	發生	28/30=0.933	(15/28)/(28/30)=0.574

關鍵字	對應字	支持度	信心水準
智能障礙	發現	26/30=0.867	(16/26)/(26/30)=0.709

關鍵字	對應字	支持度	信心水準
智能障礙	過程	29/30=0.967	(17/29)/(29/30)=0.677

關鍵字	對應字	支持度	信心水準
智能障礙	源由	27/30=0.9	(17/27)/(27/30)=0.700

關鍵字	對應字	支持度	信心水準
智能障礙	起源	28/30=0.933	(18/28)/(28/30)=0.668

關鍵字	對應字	支持度	信心水準
智能障礙	來源	25/30=0.833	(10/25)/(25/30)=0.480

關鍵字	對應字	支持度	信心水準
智能障礙	介紹	30/30=1.0	(15/30)/(30/30)=0.500

When 意圖與對應字之支持度及信心水準計算

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	時間	29/30=0.967	(26/29)/(29/30)=0.928

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	多大	28/30=0.933	(24/28)/(28/30)=0.918

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	幾月	27/30=0.9	(22/27)/(27/30)=0.905

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	日期	25/30=0.833	(18/25)/(25/30)=0.864

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	起源	28/30=0.933	(13/28)/(28/30)=0.498

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	發生	29/30=0.967	(14/29)/(29/30)=0.517

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	發現	28/30=0.933	(12/28)/(28/30)=0.459

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	時期	29/30=0.967	(21/29)/(29/30)=0.749

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	時機	26/30=0.867	(15/26)/(26/30)=0.665

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	時間點	25/30=0.833	(16/25)/(25/30)=0.768

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	發源	25/30=0.833	(5/25)/(25/30)=0.240

關鍵字	對應字	支持度	信心水準
小孩、開始、爬	發跡	20/30=0.667	(3/20)/(20/30)=0.225

What 意圖與對應字之支持度及信心水準計算

關鍵字	對應字	支持度	信心水準
聽覺、障礙	介紹	28/30=0.933	(23/28)/(28/30)=0.919
關鍵字	對應字	支持度	信心水準
聽覺、障礙	定義	28/30=0.933	(25/28)/(28/30)=0.957
關鍵字	對應字	支持度	信心水準
聽覺、障礙	認識	29/30=0.967	(23/29)/(29/30)=0.820
關鍵字	對應字	支持度	信心水準
聽覺、障礙	情況	28/30=0.933	(15/28)/(28/30)=0.574
關鍵字	對應字	支持度	信心水準
聽覺、障礙	狀況	29/30=0.967	(13/29)/(29/30)=0.464
關鍵字	對應字	支持度	信心水準
聽覺、障礙	解釋	29/30=0.967	(22/29)/(29/30)=0.785

How 意圖與對應字之支持度及信心水準計算

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	解法	28/30=0.933	(25/28)/(28/30)=0.957

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	算法	28/30=0.933	(24/28)/(28/30)=0.918

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	答案	28/30=0.933	(23/28)/(28/30)=0.880

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	公式	27/30=0.9	(22/27)/(27/30)=0.905

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	步驟	28/30=0.933	(15/28)/(28/30)=0.574

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	方法	29/30=0.967	(17/29)/(29/30)=0.606

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	題庫	29/30=0.967	(5/29)/(29/30)=0.178

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	題材	28/30=0.933	(3/28)/(28/30)=0.115

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	類型	26/30=0.867	(4/26)/(26/30)=0.177

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	題目	25/30=0.833	(5/25)/(25/30)=0.240

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	方式	22/30=0.733	(4/22)/(22/30)=0.248

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	用途	29/30=0.967	(4/29)/(29/30)=0.142

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	公式	27/30=0.9	(16/27)/(27/30)=0.658

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	流程	28/30=0.933	(3/28)/(28/30)=0.115

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	辦法	26/30=0.867	(4/26)/(26/30)=0.177

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	情況	23/30=0.767	(1/23)/(23/30)=0.057

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	狀況	25/30=0.833	(2/25)/(25/30)=0.096

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	問題	28/30=0.933	(12/29)/(29/30)=0.444

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	認識	27/30=0.9	(2/27)/(27/30)=0.082

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	學習	29/30=0.967	(6/29)/(29/30)=0.214

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	教	29/30=0.967	(11/29)/(29/30)=0.392

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	算	26/30=0.867	(17/26)/(26/30)=0.754

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	輔導	25/30=0.833	(2/25)/(25/30)=0.096

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	輔助	28/30=0.933	(4/28)/(28/30)=0.153

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	運算	29/30=0.967	(22/29)/(29/30)=0.785

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	計算	29/30=0.967	(23/29)/(29/30)=0.820

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	練習	27/30=0.9	(5/27)/(27/30)=0.206

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	求得	26/30=0.867	(18/26)/(26/30)=0.799

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	求	25/30=0.833	(16/25)/(25/30)=0.768

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	解	29/30=0.967	(20/29)/(29/30)=0.713

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	教導	28/30=0.933	(5/28)/(28/30)=0.191

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	解釋	26/30=0.867	(7/26)/(26/30)=0.311

關鍵字	對應字	支持度	信心水準
計算、等腰三角形面積	教法	29/30=0.967	(9/29)/(29/30)=0.321

Who 意圖與對應字之支持度及信心水準計算

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	發明人	20/30=0.667	(7/20)/(20/30)=0.525

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	原著	28/30=0.933	(12/28)/(28/30)=0.459

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	作者	28/30=0.933	(13/28)/(28/30)=0.498

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	發明者	25/30=0.833	(10/25)/(25/30)=0.480

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	創作人	28/30=0.933	(11/28)/(28/30)=0.421

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	創造人	29/30=0.967	(13/29)/(29/30)=0.464

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	名字	29/30=0.967	(25/29)/(29/30)=0.891

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	姓名	29/30=0.967	(23/29)/(29/30)=0.820

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	姓氏	27/30=0.9	(20/27)/(27/30)=0.823

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	介紹	28/30=0.933	(21/28)/(28/30)=0.804

關鍵字	對應字	支持度	信心水準
12 屆、中華民國、總統	解釋	29/30=0.967	(16/29)/(29/30)=0.571

Where 意圖與對應字之支持度及信心水準計算

關鍵字	對應字	支持度	信心水準
南華大學	發源地	28/30=0.933	(15/28)/(28/30)=0.574

關鍵字	對應字	支持度	信心水準
南華大學	原點	28/30=0.933	(12/28)/(28/30)=0.459

關鍵字	對應字	支持度	信心水準
南華大學	原處	26/30=0.867	(11/26)/(26/30)=0.488

關鍵字	對應字	支持度	信心水準
南華大學	發生地	25/30=0.833	(18/25)/(25/30)=0.864

關鍵字	對應字	支持度	信心水準
南華大學	出處	28/30=0.933	(13/28)/(28/30)=0.498

關鍵字	對應字	支持度	信心水準
南華大學	方向	29/30=0.967	(20/29)/(29/30)=0.713

關鍵字	對應字	支持度	信心水準
南華大學	距離	27/30=0.9	(17/27)/(27/30)=0.700

關鍵字	對應字	支持度	信心水準
南華大學	位置	29/30=0.967	(21/29)/(29/30)=0.749

關鍵字	對應字	支持度	信心水準
南華大學	原生地	26/30=0.867	(10/26)/(26/30)=0.444

關鍵字	對應字	支持度	信心水準
南華大學	地點	25/30=0.833	(17/25)/(25/30)=0.816

關鍵字	對應字	支持度	信心水準
南華大學	方位	25/30=0.833	(13/25)/(25/30)=0.624

關鍵字	對應字	支持度	信心水準
南華大學	地址	29/30=0.967	(24/29)/(29/30)=0.856

關鍵字	對應字	支持度	信心水準
南華大學	地方	28/30=0.933	(21/28)/(28/30)=0.804

附　　錄　　二

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：如何計算等腰三角形面積?
意圖：How　　關鍵字：等腰三角形面積
對應字：解法、算法、答案　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.433</td></tr>
<tr><td>13 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.567</td></tr>
<tr><td>17 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.567</td></tr>
<tr><td>17 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：要怎樣製作網頁? 意圖：How　　關鍵字：製作、網頁 對應字：方法、步驟、流程　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 5 筆	0.167
兩個關鍵字	符合的資料 15 筆	0.500
三個關鍵字	符合的資料 20 筆	0.667
四個關鍵字	符合的資料 21 筆	0.700

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：要如何做蛋糕?

意圖：How　　關鍵字：做、蛋糕

對應字：做法、方法、流程　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.100</td></tr>
<tr><td>3 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.733</td></tr>
<tr><td>22 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.867</td></tr>
<tr><td>26 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：要怎樣才能製造出簡易機器人? 意圖：How　　關鍵字：製造、簡易、機器人 對應字：方法、流程、步驟　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 5 筆	0.167
兩個關鍵字	符合的資料 10 筆	0.333
三個關鍵字	符合的資料 11 筆	0.367
四個關鍵字	符合的資料 12 筆	0.400

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：要如何去九份?

意圖：How　　關鍵字：去、九份

對應字：方法、辦法、方式　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：什麼叫做聽覺障礙?
意圖：What　　關鍵字：聽覺、障礙
對應字：定義、認識、介紹　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.333</td></tr>
<tr><td>10 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.733</td></tr>
<tr><td>22 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.833</td></tr>
<tr><td>25 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：何謂長期記憶?		
意圖：What	關鍵字：長期記憶	
對應字：定義、認識、介紹	測試樣本：30 筆	
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料	0.267
	8 筆	
兩個關鍵字	符合的資料	0.500
	15 筆	
三個關鍵字	符合的資料	0.567
	17 筆	
四個關鍵字	符合的資料	0.633
	19 筆	

搜尋 Google 資料之準確率統計表

例子：什麼是語言發展? 意圖：What　　關鍵字：語言發展 對應字：介紹、認識、定義　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 11 筆	0.367
兩個關鍵字	符合的資料 14 筆	0.467
三個關鍵字	符合的資料 18 筆	0.600
四個關鍵字	符合的資料 21 筆	0.700

搜尋 Google 資料之準確率統計表

例子：高血壓是什麼? 意圖：What　　關鍵字：高血壓 對應字：定義、認識、介紹　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 10 筆	0.300
兩個關鍵字	符合的資料 14 筆	0.467
三個關鍵字	符合的資料 17 筆	0.567
四個關鍵字	符合的資料 17 筆	0.567

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：何謂為肝癌?
意圖：What　　關鍵字：肝癌
對應字：認識、介紹、定義　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.667</td></tr>
<tr><td>20 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.733</td></tr>
<tr><td>22 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.833</td></tr>
<tr><td>25 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：小孩什麼時候開始學爬?
意圖：When　　關鍵字：小孩、開始、學、爬
對應字：時間、多大、幾月　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.333</td></tr>
<tr><td>10 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.367</td></tr>
<tr><td>11 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：煮菜要何時開始放油?
意圖：When　　關鍵字：煮菜、開始、放油
對應字：時間點、時機、時間　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.167</td></tr>
<tr><td>5 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.200</td></tr>
<tr><td>6 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.233</td></tr>
<tr><td>7 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.267</td></tr>
<tr><td>8 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：台灣稻米何時開始播種?
意圖：When　　關鍵字：台灣、稻米、開始、播種
對應字：時間、時期、日期　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.100</td></tr>
<tr><td>3 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.133</td></tr>
<tr><td>4 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.267</td></tr>
<tr><td>8 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.233</td></tr>
<tr><td>7 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：百合什麼時候可以澆水?
意圖：When　　關鍵字：百合、澆水
對應字：時間點、時機、時間　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.133</td></tr>
<tr><td>4 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.267</td></tr>
<tr><td>8 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.333</td></tr>
<tr><td>10 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.367</td></tr>
<tr><td>11 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：國小何時開始學注音?
意圖：When　　關鍵字：國小、開始、學、注音
對應字：時間、時期、時間點　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.100</td></tr>
<tr><td>3 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.267</td></tr>
<tr><td>8 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.233</td></tr>
<tr><td>7 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.367</td></tr>
<tr><td>11 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：走哪邊可以去基隆? 意圖：Where　　關鍵字：去、基隆 對應字：地點、位置、方向　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 10 筆	0.333
兩個關鍵字	符合的資料 15 筆	0.500
三個關鍵字	符合的資料 18 筆	0.600
四個關鍵字	符合的資料 16 筆	0.533

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：南華大學在什麼地方?
意圖：Where　　關鍵字：南華大學
對應字：地址、地點、地方　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.467</td></tr>
<tr><td>14 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.633</td></tr>
<tr><td>19 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.667</td></tr>
<tr><td>20 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：哪邊可以去台北捷運站?
意圖：Where　　關鍵字：去、台北捷運站
對應字：地點、位置、方向　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.633</td></tr>
<tr><td>19 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.733</td></tr>
<tr><td>22 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：離台中火車站有多遠?
意圖：Where　　關鍵字：離、台中火車站
對應字：距離、地點、方向　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.300</td></tr>
<tr><td>9 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.400</td></tr>
<tr><td>12 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.433</td></tr>
<tr><td>13 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：台北 101 在何方?		
意圖：Where	關鍵字：台北、101	
對應字：方向、位置、方位	測試樣本：30 筆	
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料	0.333
	10 筆	
兩個關鍵字	符合的資料	0.433
	13 筆	
三個關鍵字	符合的資料	0.500
	15 筆	
四個關鍵字	符合的資料	0.367
	11 筆	

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：誰是第 12 屆中華民國總統?
意圖：Who　　關鍵字：第 12 屆、中華民國、總統
對應字：姓名、名字、介紹　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.333</td></tr>
<tr><td>10 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.433</td></tr>
<tr><td>13 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：著作三國志的人是誰?
意圖：Who　　關鍵字：著作、三國志
對應字：作者、原著、發明人　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.067</td></tr>
<tr><td>2 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.100</td></tr>
<tr><td>3 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.167</td></tr>
<tr><td>5 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.167</td></tr>
<tr><td>5 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：誰發明了飛機? 意圖：Who　　關鍵字：發明、飛機 對應字：姓名、介紹、發明人　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 8 筆	0.267
兩個關鍵字	符合的資料 10 筆	0.333
三個關鍵字	符合的資料 12 筆	0.400
四個關鍵字	符合的資料 9 筆	0.300

搜尋 Google 資料之準確率統計表

例子：電燈是由誰發明的? 意圖：Who　　關鍵字：電燈、發明 對應字：發明人、創造人、發明者　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 12 筆	0.400
兩個關鍵字	符合的資料 13 筆	0.433
三個關鍵字	符合的資料 13 筆	0.433
四個關鍵字	符合的資料 14 筆	0.467

搜尋 Google 資料之準確率統計表

例子：誰是微軟的創辦人?		
意圖：Who	關鍵字：微軟、創辦人	
對應字：姓名、名字、介紹	測試樣本：30 筆	
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料	0.500
	15 筆	
兩個關鍵字	符合的資料	0.533
	16 筆	
三個關鍵字	符合的資料	0.567
	17 筆	
四個關鍵字	符合的資料	0.467
	14 筆	

搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：為什麼會有智能障礙?
意圖：Why　　關鍵字：智能、障礙
對應字：原因、因素、成因　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.400</td></tr>
<tr><td>12 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.533</td></tr>
<tr><td>16 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.600</td></tr>
<tr><td>18 筆</td></tr>
</table>

搜尋 Google 資料之準確率統計表

例子：為何會常拉肚子? 意圖：Why　　關鍵字：常拉、肚子 對應字：原因、因素、成因　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 2 筆	0.067
兩個關鍵字	符合的資料 9 筆	0.300
三個關鍵字	符合的資料 10 筆	0.333
四個關鍵字	符合的資料 12 筆	0.400

搜尋 Google 資料之準確率統計表

例子：我為什麼會發生頭痛? 意圖：Why　　關鍵字：發生、頭痛 對應字：原因、因素、成因　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 12 筆	0.400
兩個關鍵字	符合的資料 19 筆	0.633
三個關鍵字	符合的資料 22 筆	0.733
四個關鍵字	符合的資料 23 筆	0.767

搜尋 Google 資料之準確率統計表

例子：他為什麼會產生心臟病? 意圖：Why　　關鍵字：產生、心臟病 對應字：起因、原因、因素　　測試樣本：30 筆		
關鍵字/搜尋引擎	Google	準確性
單一關鍵字	符合的資料 10 筆	0.333
兩個關鍵字	符合的資料 18 筆	0.600
三個關鍵字	符合的資料 20 筆	0.667
四個關鍵字	符合的資料 22 筆	0.733

表 36 搜尋 Google 資料之準確率統計表

<table>
<tr><td colspan="3">例子：小孩為什麼會得到腸病毒?
意圖：Why　　關鍵字：得到、腸病毒
對應字：原因、介紹、成因　　測試樣本：30 筆</td></tr>
<tr><td>關鍵字/搜尋引擎</td><td>Google</td><td>準確性</td></tr>
<tr><td rowspan="2">單一關鍵字</td><td>符合的資料</td><td rowspan="2">0.267</td></tr>
<tr><td>8 筆</td></tr>
<tr><td rowspan="2">兩個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">三個關鍵字</td><td>符合的資料</td><td rowspan="2">0.500</td></tr>
<tr><td>15 筆</td></tr>
<tr><td rowspan="2">四個關鍵字</td><td>符合的資料</td><td rowspan="2">0.567</td></tr>
<tr><td>17 筆</td></tr>
</table>