

以 5W1H 問句做網路資料探勘之機制架構設計

楊惠媚
大同技術學院
企業管理系
nancy@
ms2.ttc.edu.tw

陳育銘
南華大學
資訊管理學系
nec78557054@
yahoo.com.tw

林育弘
南華大學
資訊管理學系
u0111095@
yahoo.com.tw

王昌斌
南華大學
電子商務學系
cbwang@
mail.nhu.edu.tw

摘要

由於時代進步及網路的蓬勃發展，利用網路搜尋資料已漸漸成為時代趨勢，對於問句的描述方式往往是搜尋的關鍵，但其搜尋後的資訊，並無法符合使用者的意圖且直接影響到其準確性。因此要如何透過問句去了解使用者的意圖，是目前學者所追求的研究目標。

本研究主要將使用者的問句以中研院 CKIP 系統做快速的斷詞及詞性標記，且以 5W1H 系統化的歸納方法來搭配有限自動機 (Finite Automata) 的運作方式來比對出意圖型態、關鍵字及 5W1H 對應字，再結合關鍵字與對應字後於網路搜尋引擎中做搜尋。本研究實驗結果顯示我們所提出的加入 5W1H 對應字方法可改善其搜尋的準確率，以幫助查詢者在網路上找到符合自己意圖的答案。

關鍵詞：有限自動機、意圖、5W1H

Abstract

As a result of time progress and network's vigorous development, the question description often are the search keys, also immediate influence search material accuracy. The search website generally in is unable to change the search under the order, important and duplicated excessively many limits, causes to be unable to conform to inquiry intention and the answer, therefore how can improve conforms to the inquiry intention, is goal which the present scholar studies.

This research is mainly Chinese interrogative sentence by Academia Sinica CKIP system fast breaks the to sentence segmentation and Part-Of-Speech, and matches the Finite Automata machine by the

5W1H systematization's induction method the intenten compares to leaves the intention state, the key words and the 5W1H correspondence word, after the recombination key words and the corresponding word, makes the search in the search engine. This research experimental result showed we proposed joins the 5W1H correspondence character method to be possible to improve its search the rate of accuracy, helps the inquiry to find in the network conforms to myself intention answer.

Keywords: Finite Automata、Intention、5W1H

1. 緒論

1.1 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，太多的搜尋引擎並沒考慮到使用者所要問的意圖，然而其搜尋結果也無法優先呈現給使用者。因此，使用者往往需要花費更多精力在篩選所取得的資料。

自然語言的發展，是目前吸引許多學者持續投入相關模型的建構與開發，近年來由於統計式語言處理技術[24]的成熟，讓許多的難題有了新的理解及解決之道。透過學者們的研究，自然語言處理的技巧應用在許多領域上，包括資訊檢索[13]、文件分類[21]、文件摘要[4]、構詞語法[34]、詞義消歧[17]、字典編纂[32]、詞彙語意[7]、未知詞擷取[11]、Q&A等。

近來自然語言系統[6]的研究逐漸活絡起來，主要的原因是，透過觀察大量的語料庫[15][8]，可知道人們生活起居語言使用的情形。其次，電腦科技的突飛猛進，有快速的微處理器足以進行複雜的語言模型計算[24]；有容量的儲存媒體得以儲存訓練語料，使得早在 60 年代就已經存在的研究方法，再度呈現新

的生命。

1.2 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，大多以關鍵字為主，但因所檢索方式不同，所造成搜尋的資訊並無法符合使用者的意圖。然而現今的檢索系統大多都已加入了自然語言問句輸入功能，以大大改善檢索的效能。張鐘尹表示[14]，自然語言可協助我們找到明確的意圖及對應關係。

對此，本研究設計一系統，可讓使用者以自然語言方式輸入問句，經解析後之語意於網路上做相關的文件內容的搜尋與擷取，藉由文件建立知識間彼此的關聯，並將此知識內容建構成類似知識地圖。

本機制分為三大功能模組如圖 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。

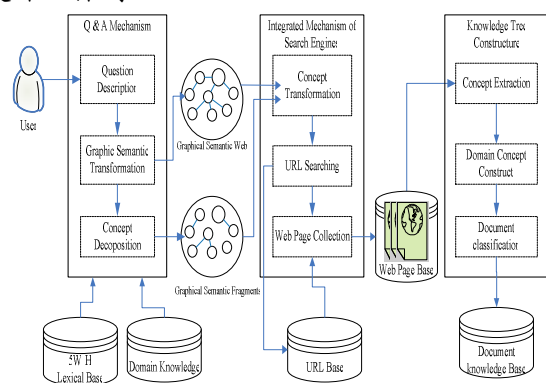


圖 1 整體架構

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[33]，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) [16] [30]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。

2. 文獻探討

2.1 5W1H

一個問句中可以作为協助檢索的分為兩部分，一個是「關鍵字 (Keyword)」，另外則是「意圖 (Intention)」，關鍵字部分代表構成

一個問句的關鍵詞，這關鍵詞有可能為一個詞或是數個詞所構成的集合；而意圖則代表該問句的意圖，在這意圖可由原問句裡的片語或是一些片語的組合而成。將於本章做意圖的詳細介紹。

根據學者張鐘尹[14]表示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，本研究於表 1 至表 6 分類出 5W1H 的意圖。

表 1 What 之例子

問句	關鍵字	意圖
什麼是四則運算?	四則運算	四則運算
何謂學習障礙?	學習障礙	學習障礙
什麼是數學?	數學	數學

表 2 Who 之例子

問句	關鍵字	意圖
誰知道輕度障礙?	輕度障礙	人名
誰會數學除法?	數學除法	人名
誰了解數學教學模式?	數學教學模式	人名

表 3 Why 之例子

問句	關鍵字	意圖
為什麼會有學習障礙?	學習障礙	原因
為什麼不了解四則運算?	四則運算	原因
你為何會發生車禍?	發生車禍	原因

表 4 When 之例子

問句	關鍵字	意圖
小孩何時開始學走路?	小孩學走路	時間
他離開是什麼時候?	離開	時間
他畢業是什麼時間?	畢業	時間

表 5 Where 之例子

問句	關鍵字	意圖
台北是在哪?	台北	地點
嘉義是什麼地方?	嘉義	地點
阿里山的地點?	阿里山	地點

表 6 How 之例子

問句	關鍵字	意圖
要如果取得駕照?	取得駕照	方法
要怎麼學數學?	學數學	方法
如何了解心臟病?	心臟病	方法

2.2 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) [27]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[29]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[23]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。

2.3 自然語言處理技術

2.3.1 字串比對

一般常用的語意分析模型為馬可夫語言

模型(Markov Language Model)，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[3]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較常使用。

2.4 答詢系統(Q&A System)

隨著網際網路資訊量的爆炸，長久以來答詢系統(Question and Answer system)便是人工智慧裡一個很熱門的研究領域。其中的TREC就曾針對答詢系統做了以下的說明：[25]

目前的資訊檢索系統只能找出使用者有興趣的文件，通常都還需要瀏覽過所有的文件才能找的使用者所關心的資訊，不僅浪費時間更耗費精力。因此，建立一套良好的系統來解決這樣的問題就極具迫切性，而答詢系統的設計便是因應這樣的需求而產生的。答詢系統可以依據使用者的問題來回答其所關心的資訊，而並不是輸出相關的文件。

從 TREC 的說明不難得知，答詢系統基本上主要由資訊檢索系統及資訊擷取系統所組成，資訊檢索系統先依據使用者的“問題”篩選出所有可能的文件，然後再交給資訊擷取系統擷取出使用者所關心的資訊。一般來說，答詢系統要能成功所要考慮的因素非常的多，從前端資訊檢索的成功與否，到後端資訊擷取的正確率都是環環相扣的，甚至對於使用者“問題”理解是否正確也是非常重要的因素。一般的答詢系統在前端都會有問題處理器(Question Processor)先行處理使用者的問題，其主要的工作是希望可以確認出『問題的種類(Information of asking point)』以及『建立關鍵字』[28]。以“Who is Julian Hill?”為例，問題處理器會先確認出問題種類為『PERSON』(表示希望系統找出跟“人”有關的資訊，如年齡、性別等)，並建立關鍵字『Julian Hill』。資訊檢索系統會依據關鍵字『Julian Hill』從語料庫中找出所有關於『Julian Hill』的文件，接著再透過排序演算法(Ranking Algorithm)縮

小文件處理的數目，最後再由資訊擷取系統擷取出關於『Julian Hill』的資訊。圖 2 即為此答詢系統的執行結果。

Q : Who is Julian Hill?
 A : name : Julian Werner Hill
 Type : Person
 Age : 91
 Gender : Male
 Position : research chemist
 Affiliation : Du Pont CO.
 Education : Washigton University ; MIT

圖 2 答詢系統執行結果

從圖 2 我們即可看出資訊擷取技術在答詢系統中所扮演的角色，倘若沒有好的資訊擷取技術，答詢系統也就無法發揮功用。因此我們可以說答詢系統的核心工作便是建立在良好的資訊擷取技術上。當然，答詢系統不單單只是在做資訊擷取而已，從問題理解、資訊檢索、到資訊擷取，每一個部份都非常重要而可以加以研究的課題。以問題理解為例，隨著各種語言的特殊及多變性，對於問題類型的確認就存在著模糊性(ambiguity)，而並不只是列舉問題的種類就可以加以解決，而這其中又以中文為甚，因為中文在前處理上必須予以斷詞，斷詞的成功與否，就會影響到接下來的處理。

此外，中文結構較為鬆散、多縮寫型式，想要對中文問題加以理解，可以說是一項非常嚴苛挑戰。儘管如此，資訊擷取仍然可以說是答詢系統中的核心工作，而積極開發出正確率高的資訊擷取系統也就成為了我們目前的研究目標。

2.5 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(Entities 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。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[16]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。

2.6 關聯規則(Association Rule)

聯規則 (Association Rule) 探勘方式首先由Agrawal[26]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 $I = \{ I_1, I_2, \dots, I_n \}$ 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 7 交易資料庫片段

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閾值項目組，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：

$\text{Support}(i) = \text{項目} I \text{ 出現次數頻率} / \text{資料庫中交易筆數}$ 。

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：

$\text{Confidence}(A \rightarrow B) = A, B \text{ 同時在資料庫中出現的次數} / \text{項目} A \text{ 在資料庫中出現次數}$ 。

例如：(milk $\rightarrow$ bread, support=0.2, 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。

3. 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。

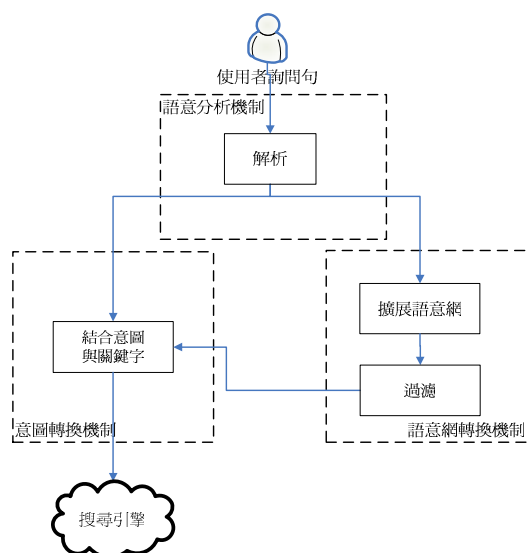


圖 3-1 圖形化語意網轉換機制架構圖

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。本機制除了提供使用者一個圖形化的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。語意分析機制中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，以前處理(Pre-Processing)後再透過解析將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利用 Domain Ontology 來進行擴展成語意網。意圖轉換機制主要將意圖與關鍵字結合後，於資料庫中找出所對應資料，經轉換後，於網路搜尋引擎上做查詢。

3.1 語意分析機制(Semantic Analysis Mechanism)

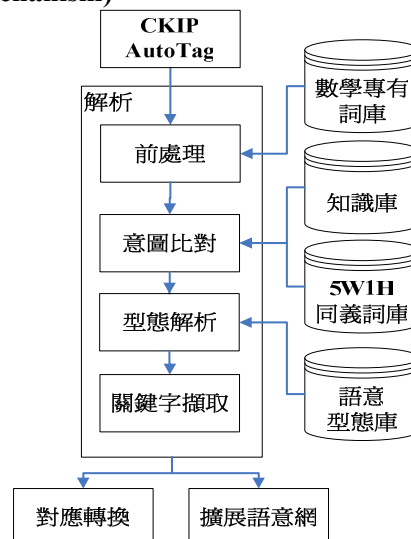


圖 3-2 語意分析機制

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 AutoTag

主要將使用者的問題斷詞及詞性標記，把我們需要的關鍵字標記出來，我們將以一例子「如何學習除法」來說明，以 CKIP[5]斷詞及標記後分為三部分「如何(D) 學習(VC) 除法(Na)」。

(2) 前處理(Pre-Processing)

將上步驟如未能斷詞完成的詞做結合，如「四則運算」等專有名詞，會斷成「四(Neu)

則(Nf)運算(VC)」三部分，透過此步驟可將「四則運算」合併為單一的專有名詞「四則運算(Na)」、「Nb」、「.....」等詞彙。

(3)意圖比對(Intention Matching)

將問句字串配合 5W1H 同義詞庫比對出所屬問句的意圖類型，其詞庫由楊宸彥及牛維娟[9] [2]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

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

(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) [12]演算法做型態解析，以建立有限自動機(Finite Automata)與語意型態庫做快速掃描，可判斷出型態類型，如圖 3-9 所示。

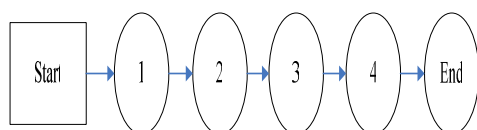


圖3-9 有限自動機流程圖

5W1H 之六種型態模型，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

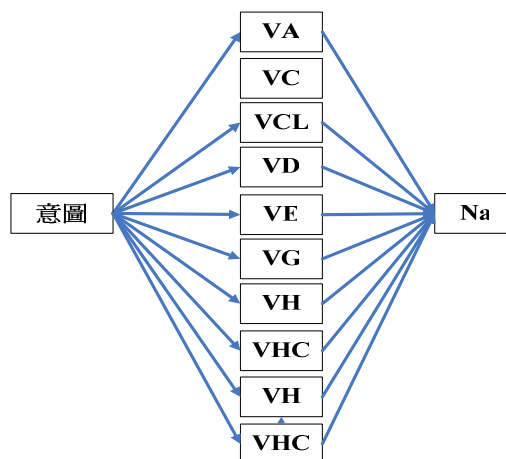


圖 3-10 類型 How 之第一型

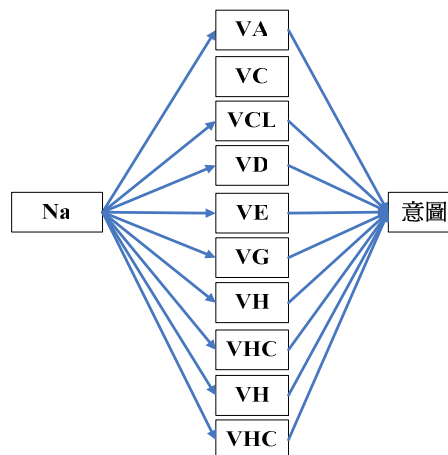


圖3-11 類型How之第二型

第二種型態What

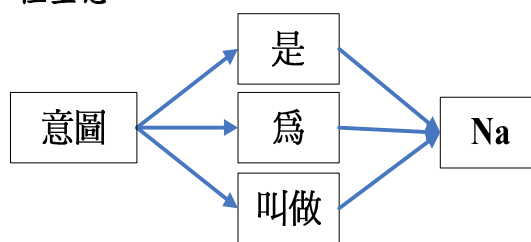


圖 3-12 類型 What 之第一型

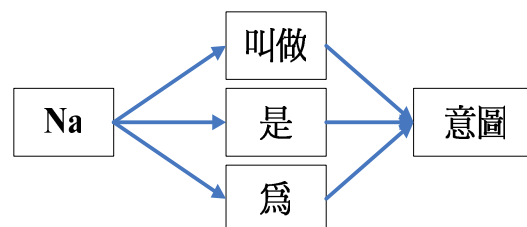


圖3-13 類型What之第二型

第三種型態Who

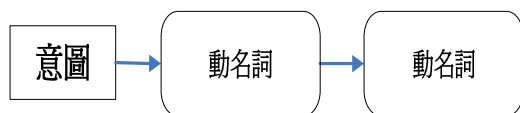


圖 3-14 類型 Who 之第一型

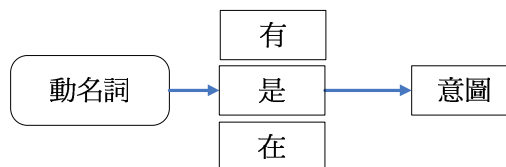


圖 3-20 類型 Where 之第二型

第四種型態 Why

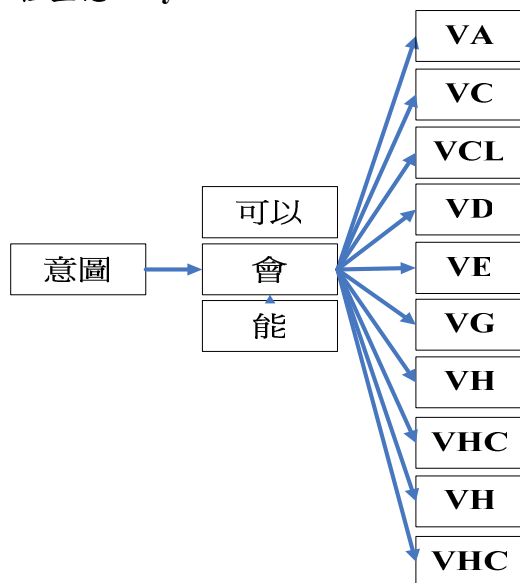


圖 3-15 類型 Why 之第一型

第五種型態 When

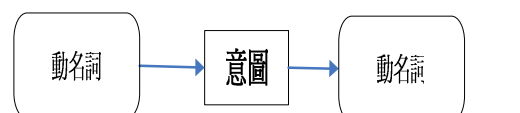


圖 3-16 類型 When 之第一型

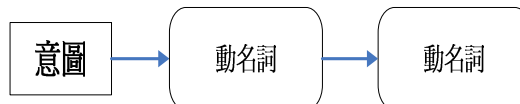


圖 3-17 類型 When 之第二型

第六種型態 Where

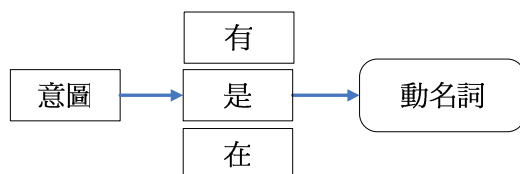


圖 3-19 類型 Where 之第一型

Kunth-Morris-Pratt演算法：

輸入： P及T，字樣及本文字串；f 陣列(失敗鍵陣列)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)

經上步驟後，問句字串中包含名詞「N」的字串擷取出來，如「Na」、「Nb」、「VA」、「VB」、「...」等來作為關鍵字，以提供下一步驟做擴展語意網。

3.2 語意網轉換機制(Semantic Net Transformation Mechanism)

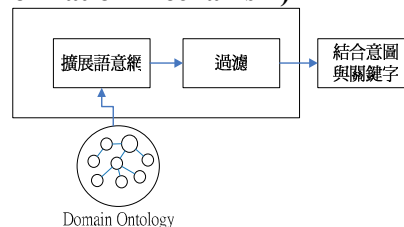


圖 3-22 語意網轉換機制

本機制包括三步驟如圖 3-22 所示，將包含名詞「Na」或「VA」等的關鍵字，透過 Ontology 技術擴展成語意網，目的為將語意轉

換成圖形化。其說明如下：

(1) 擴展語意網(Extended Semantic Net)

被延伸後的關鍵字，利用領域本體論(Domain Ontology)去延伸相關概念出來，形成語意網。以圖 3-23 為例：

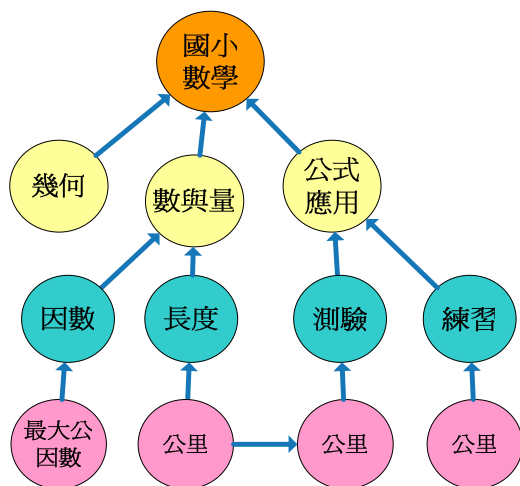


圖 3-23 國小數學領域之 ontology 架構

(2) 過濾(Filtering)

根據本體論技術，在系統上設立一參數門檻值，可依使用者所需作增加或減少。以下圖 3-24 為例：

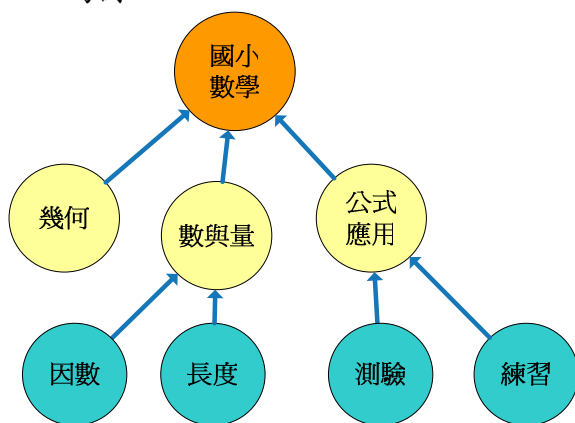


圖3-24 國小數學領域之ontology架構

3.3 意圖轉換機制(Intention Transformation Mechanism)

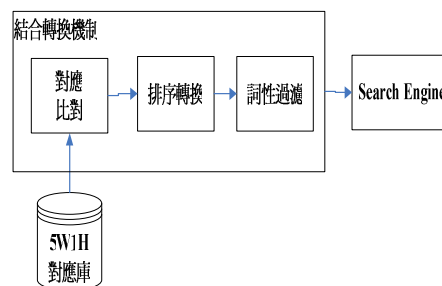


圖 3-25 意圖轉換機制

本機制目的是將意圖與關鍵字結合後，以 5W1H 對應庫的「動名詞」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 對應庫如表 3-2 的比對，將所對應的動名詞找出來。

表 3-2 5W1H 對應資料表

意圖類型	類型	Domian Keyword	對應名詞	對應動詞
why	數學類	數學知識	來源、源由、理由、原因、起源、介紹、因素、因子	發生、發現、過程
		等腰三角形面積		
		四則運算		
		數學的教學模式		
		平行四邊形		
		智能障礙		
when	數學類	數學知識	日期、起源、時間	發生、發現、發源、發跡
		等腰三角形面積		
		四則運算		
		數學的教學模式		
		平行四邊形		
		智能障礙		
what	數學類	數學知識	介紹	
		等腰三角形面積		
		四則運算		
		數學的教學模式		
		平行四邊形		
		智能障礙		
how	數學類	數學知識	解法、答案、公式、算法、教法、方法、步驟、題庫、題材、類型、題	教、算、輔導、輔助、運算、計算、練習、求得、求、解、教導、解釋
		等腰三角形面積		

		四則運算	目、方式、用途、公式	
		數學的教學模式		
		平行四邊形		
		乘法運算		
		智能障礙	情況、狀況、問題	
where	數學類	數學知識	發源地、原點、原處、發生地、原生地、出處	
		等腰三角形面積		
		四則運算		
		數學的教學模式		
		平行四邊形		
		智能障礙		
who	數學類	數學知識	發明人、原著、作者、發明者、創作人、創造人	
		等腰三角形面積		
		四則運算		
		數學的教學模式		
		平行四邊形		
		智能障礙		

(2) 排序轉換(Ranking Transformation)

經上一步驟後，從資料表中對應出來的「動名詞」配合關鍵字來做排序轉換，本研究經過文件訓練後如「四則運算」+「方法」=「名詞」+「對應字」，透過王亭雅[1]所提之方式可做直接結合轉換。以下圖3-29例子表示：

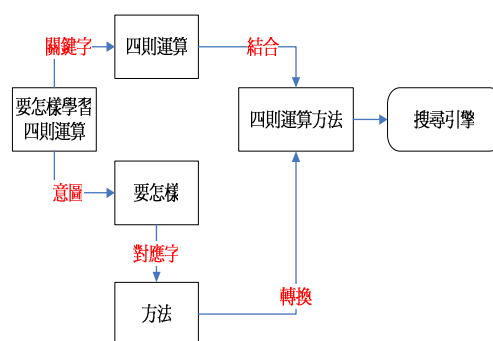


圖 3-29 轉換示意圖

(3) 詞性過濾(Filtering)

將轉換過後的字串做詞性過濾，讓下一步驟整合型網頁搜尋引擎機制做搜尋處理，以解決使用者之問題。

4. 實驗結果

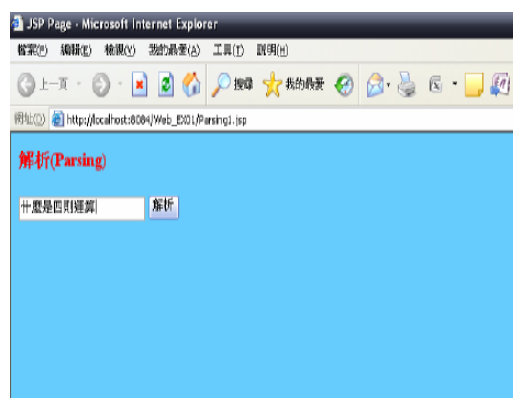


圖 4-1 解析頁面

上圖 4-1 為使用者以自然語言輸入一問句後，經解析於下一步驟。

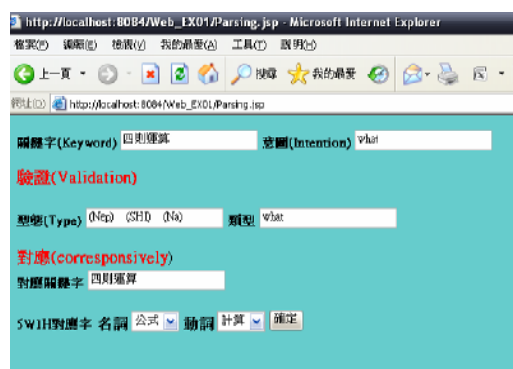


圖 4-2 驗證對應頁面

圖 4-2 主要將使用者問句經解析，可把關鍵字及意圖比對出來，透過一驗證機制將使用者問句之型態及類型比對出來。

經驗證後之意圖及關鍵字，可由意圖對應

到 5W1H 對應字動明詞。可選擇出所需之動名詞於網路搜尋引擎做搜尋。



圖 4-3 網路搜尋頁面

圖 4-3 主要是將關鍵字結合使用者所選之對應字於網路上做搜尋，利用此機制，期望解決使用者的問題。

圖 4-4 及 4-5 表示，本研究將關鍵字及對應字於搜尋引擎 Google 及 Yahoo 做搜尋，以 Precision 公式來做評估。

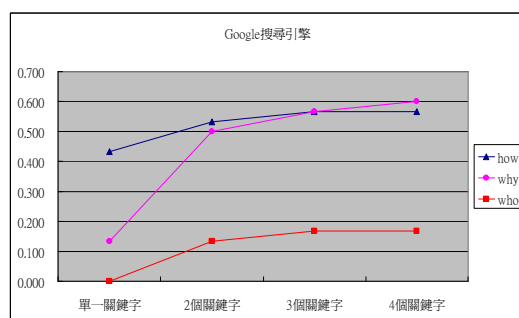


圖 4-4 搜尋 Google 資料準確率比較圖

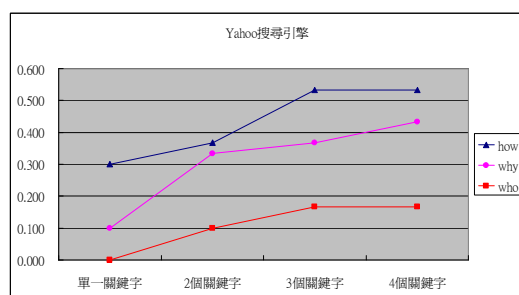


圖 4-5 搜尋 Yahoo 資料準確率比較圖

結論與未來展望

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、型態解析及問題轉換。而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：

- (1)可讓使用者更容易了解其所描述之問句的真正涵意。
- (2)圖形化語意網可應用於其他相關之領域，例如知識地圖、討論區等，能挖掘出豐沛的知識，幫助我們容易了解內容。
- (3)本研究歸納出 5W1H 之問句類型，能更快速及提高準確性。
- (4)透過語意網的擴展，能讓使用者找到更多所需知識。
- (5)轉換後的字串，可大大提高搜尋的準確率。
- (6)有限自動機(Finite Automata)的建立，可改善比對的效率及準確率。
- (7)在 5W1H 對應庫中，意圖類型的對應字眾多，可透過權重庫來設定權重值，可經過使用者的每一筆處理過後的資料來做訓練，形成一回饋機制，希望能在未來做深入探討，達到本研究最終目的。

參考文獻

- [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] 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。
- [3] 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。
- [4] 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。
- [5] 中央研究院中文斷詞系統，URL：
<http://ckipsvr.iis.sinica.edu.tw/>
- [6] 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。
- [7] 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。
- [8] 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。

- [9] 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。
- [10] 黃耀明(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。
- [11] 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。
- [12] Baase, S., and Gelder, A. V., "Computer Algorithms, Introduction to Design and Analysis", Addison Wesley Longman, 2000.
- [13] Baeza-Yates, R. and B. Ribeiro-Neto, Modern Information Retrieval, The ACM Press, Addison Wesley Longman Limited, Essex, England, 1999.
- [14] Chung-Yin Chang, "A Discourse Analysis of Questions in Mandarin Conversion," M.A. Thesis, National Taiwan University Graduate Institute of Linguistics, June 1997.
- [15] 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), 1994, pp. 234-241.
- [16] Gruber, T. R., "A translation approach to portable ontology specifications", Knowledge Acquisition, 5(2), 1993, pp. 199-220.
- [17] 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.
- [18] J. Han and K. Kamber, Simon Fraser University (2001), Data Mining: Concepts and Techniques, 5th ed., San Francisco: Morgan Kaufmann, pp. 230-235
- [19] Jordan, P.W., "Using Terminological Knowledge Representation Languages to Manage Linguistic Resources," cmp-lg/9605024 (URL : <http://xxx.lanl.gov/>), 1996.
- [20] Kao, M., Cercone, N. and Luk, W. (1988). Providing Quality Responses with Natural Language Interface: the Null Value Problem. IEEE Transactions on Software Engineering, 14(7), 959-984
- [21] 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.
- [22] Li, Charles and Sandra A. Thompson. 1981. "Mandarin Chinese: A functional reference grammar". Berkeley and Los Angeles University of California Press.
- [23] Mitri, M., 1993, "Combing semantic networks with multi-attribute utility models: An evaluative database indexing method," Proceedings, Ninth Conference on Artificial Intelligence for Applications, pp. 462.
- [24] Manning, C. D. and H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, London, England, 1999.
- [25] QA Track Specifications, URL : <http://www.research.att.com/>
- [26] ~singhal/qa-track-sepc.txt
- [27] R. Agawal, T. Imielinski and A. Swami. "Mining Association Rules between Sets of Items in Large Databases," Proc. Of ACM SIGMOD, pages 207 - 216, May 1993.
- [28] Ruan, W., Buerkle, T. and Dudeck, J., 2000, "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.
- [29] 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.
- [30] Smolentsev, S. V., 2002, "The identification and self-organization problems in dynamic semantic networks," IEEE International Conference on Artificial Intelligence Systems, pp. 35-39.
- [31] 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, 2001, pp. 3-9.
- [32] Salton, G., and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc, 1983.
- [33] Wilks, Y. A., B. M. Sator and L. M. Guthrie, Electric Words, The MIT Press,

London, England, 1996

- [34] 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.
- [35] Yarowsky, D. and R. Wicentowski, “Minimally Supervised orphological Analysis by Multimodal Alignment”, In Proceedings of the ACL-2000, pp. 207-216, HongKong, 2000.