網際網路 FAQ 檢索中意圖萃取與語意比對之研究

賴育昇，李坤霖，吳宗憲

國立成功大學資訊工程研究所

Page 135 ~ 155

Proceedings of Research on Computational Linguistics

Conference XIII (ROCLING XIII)

Taipei, Taiwan

2000-08-24/2000-08-25

網際網路 FAQ 檢索中意圖萃取與語意比對之研究

賴育昇、李坤霖、吳宗憲

國立成功大學資訊工程研究所
{laiys, leekl, chwu}@csie.ncku.edu.tw
Fax: +886-6-2747076

摘要

本論文之主要目的是希望能利用自然語言查詢來做為 FAQ 檢索的方式。一個完整的 FAQ 樣本必定含有一個問題與該問題的答案。藉由比較使用者的詢問句以及 FAQ 樣本的問句，如果兩者的語意相當接近，則該 FAQ 樣本的答案也就可能包含使用者想要的資訊。此外，一個 FAQ 樣本的答案也可能包含其他額外的資訊。因此，除了兩個疑問句的比對之外，使用者所需的資訊也可以透過比對詢問句與 FAQ 樣本的答案而得到。

透過語意文法以及停用詞的篩選，我們將問句分成兩個部分：「意圖區段」和「關鍵詞區段」。意圖區段傳達使用者主要的意圖，關鍵詞區段包含問句中所有的關鍵詞，問句句意的比對將建立在這兩部分各自的語意比對上。此外，我們採用向量空間模型來比較詢問句中的關鍵詞與 FAQ 樣本的答案。

經實驗驗證，本論文所提出的方法確實比單純使用關鍵詞查詢來得準確，使平均正確答案的排名從第 12.04 名提升到第 2.91 名，且使得前十名的召回率由 78.06%提升到 95.11%。

1. 緒論

1-1. 背景說明

目前資訊檢索(information retrieval)的技術已經廣泛使用在我們日常生活中。舉凡上圖書館借書、網路搜尋資料，我們常會需要一些資訊檢索的工具協助我們找出想要的資料。以目前的技術，資訊檢索的應用大多只提供由關鍵詞進行查詢，藉由關鍵詞的比對以找出相關的文章或資料。但是，只利用關鍵詞查詢有兩個缺點：(1)關鍵詞不能清楚且完整地表達使用者的意圖，以致相關的搜尋結果過多，使用者往往需要經過好幾次的來回修改關鍵詞或查詢方

式才能得到想要的結果。(2)當使用者想要查詢的資料不存在關鍵詞，或者使用者無法找到適當的關鍵詞，則甚至無法找到所需的資料。

相較於關鍵詞查詢，使用自然語言查詢是最能夠清楚表達使用者意圖的方式，也是最自然的方式。隨著網路的蓬勃發展以及自然語言處理技術的提昇，以自然語言為主的資訊檢索是一個正在興起的研究方向。目前已有幾個網站提供自然語言查詢的服務：在國外有 Ask Jeeves 網站[1]以及 FAQ Finder 系統[7]，國內有寶來證券的 E 博士[5]。但是由於目前電腦技術還不能做到完全理解自然語言的意義，以致使用自然語言來做資訊檢索的研究尚未成熟，但是這卻是未來資訊檢索必定要發展的方向。若能使之結合前端的語音辨識，直接利用語音查詢，將是更加便利且人性化的一種方式。

1-2. 研究動機與目的

在以自然語言查詢為主的資訊檢索應用中，FAQ (Frequently Asked Questions)檢索是一個不錯的方向。許多網站通常會針對該領域中常被問到的問題，經由人工整理這些問題及答案，提供給進入該網站的使用者直接閱覽，以節省詢問與回答重複或相關性問題的時間。但是隨著量的增加，使用者也愈來愈難藉由直接閱覽找到所需的答案，因此，現今許多網站也提供 FAQ 檢索的服務，讓使用者搜尋所需的資訊。本論文之主要目的便是希望能利用自然語言查詢來做為 FAQ 檢索的方式。

1-3. 研究方法簡介

一個完整的 FAQ 樣本必定含有一個問題與該問題的答案。藉由比較使用者的詢問句以及 FAQ 樣本的問句，如果兩者的語意相當接近，則該 FAQ 樣本的答案也就可能包含使用者想要的資訊。此外，一個 FAQ 樣本的答案也可能包含其他額外的資訊。因此，除了兩個疑問句的比對之外，使用者所需的資訊也可以透過比對詢問句與 FAQ 樣本的答案而得到。

FAQ Finder 系統利用 Word-Net 來衡量英文詞與詞的語意相似度，為整個系統發展語意相似度的基礎。但是在問句的相似度部分，則是單純地比對兩個問句中所包含的詞組，我們認為僅僅是比較詞組並不足以代表整個句意、有欠周延，而且也有明顯的缺失。例如：「肝癌會不會導致肝硬化？」、「肝硬化會不會導致肝癌？」，此二句有完全相同的詞組，但是在意義上卻是完全不同。

每個問句都有其意圖(intention)，該意圖唯一而且在句子裡扮演相當重要的角色。本研究

所提出的方法便是希望能有效地萃取出詢問句中所包含的意圖，並且藉由意圖來協助我們分辨兩個句子的語意。透過語意文法(semantic grammar)以及停用詞(stopping words)的篩選，我們將問句分成兩個部分：「意圖區段(intention segment, IS)」和「關鍵詞區段(keyword segment, KS)」，問句句意的比對將建立在這兩部分各自的語意比對上。此外，在關鍵詞的比對上，我們依舊保留目前被廣泛使用的關鍵詞查詢為基礎的資訊檢索技術—向量空間模型(vector space model, VSM)，用來比較詢問句中的關鍵詞與 FAQ 樣本的答案。

2. 系統架構

如圖 1 所示，本論文所提出之系統架構主要分為三大部分：「語意分析器」、「問句比對器」及「內文比對器」。以下針對這三個部分做一個簡單的介紹。

2-1. 語意分析器

透過語意分析器，我們可以從問句中萃取出 IS 及 KS，做為後續問句比對以及內文比對之用。語意分析器由下面幾個子部分所組成：(1) AutoTag，中研院 CKIP 小組發展的詞性標記系統，做為本系統的前處理器，將一個句子斷詞並標示詞性。(2) 關鍵詞萃取，由詞性的判斷以及停用詞的篩選，從斷詞後的句子中找出其 KS。(3) 意圖萃取，經由整理歸納的語意文法，從問句中找出其 IS。

2-2. 問句比對器

將使用者詢問句所萃取出來的 IS 及 KS 與 FAQ 的每一個問題的 IS 及 KS 逐一做比對。問句比對器可分為下面幾個子部分：(1) 剖析器(Parser)，將語意分析器萃取出來的 IS 剖析成剖析樹(IS parse tree)。(2) IS 相似度衡量，對於任兩個 IS parse tree，採用遞迴的方式配合一對一函數的最佳化，求取兩者的最大相似度。(3) KS 相似度衡量，透過比對兩個 KS 中所包含的關鍵詞相似度，配合一對一函數的最佳化，求取兩者的最大相似度。

2-3. 內文比對器

本論文採用向量空間模型，透過比對 KS 與 FAQ 答案，找出最適合回答該詢問句的答案。其中，在 Indexing 方面，以 TF×IDF 做為詞的權重，將每一個 FAQ 樣本的答案表示成實數向量。在 Content 相似度比對上，藉由向量相似的觀點，將 KS 所含的關鍵詞組與每一個 FAQ

答案所表示成的向量做比對，找出與 KS 最相關的答案。

除了上述的三大機制外，Ranking Strategy 將問句比對器及內文比對器所得到的結果，在此做一整合，最後將排名後的網頁超連結輸出。

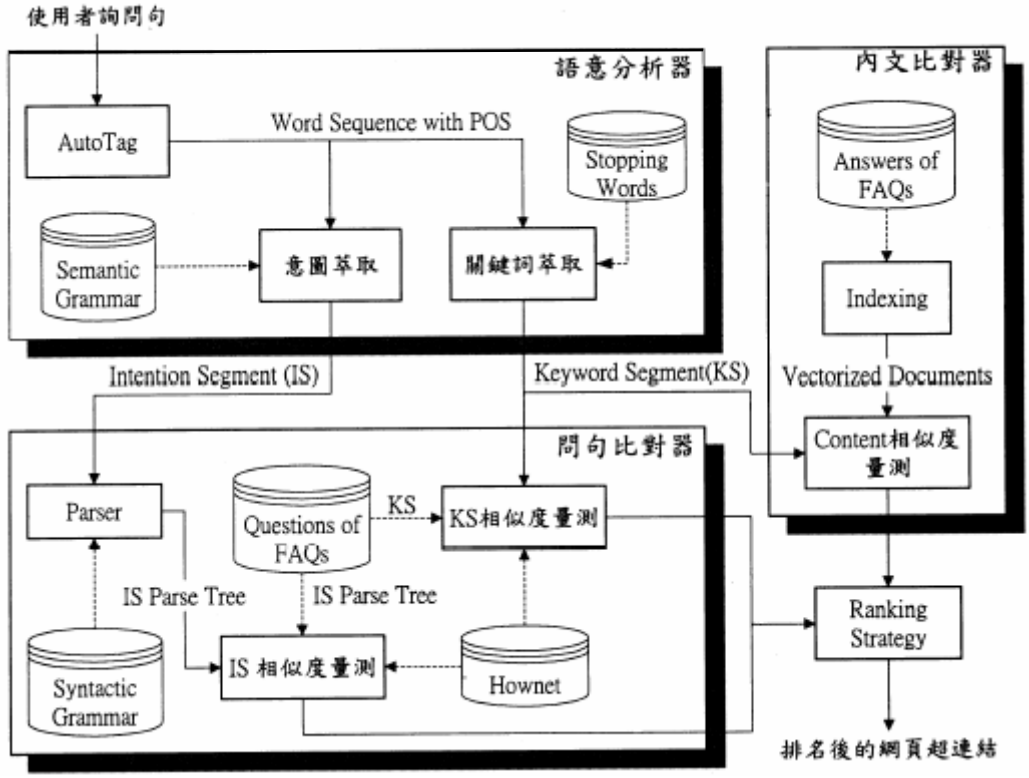

圖 1 系統架構圖

3. 問句的語意分析與處理

在大部分的情況下，關鍵詞有助於檢索出我們想要的答案，但是在符合關鍵詞比對的結果中，往往含有大量不是原來所期望獲得的答案，而其主要原因在於關鍵詞沒有辦法正確地傳達使用者的意圖。因此，我們希望透過對於問句的語意分析，能產生出問句的語意文法，進而萃取出包含在問句中的使用者意圖。

3-1. 疑問句分類

根據張鐘尹[3]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」等七個類型。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。這些功能成一線性分佈，從說話者的肯定度來看，分別表示說話者不確定性高的到不確定性低的；從訊息的角度來看，則表示說話者在尋求訊息的到傳遞訊息的疑問句。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。在網際網路上的問題也多以這三種形式存在，因此，本論文即針對此三種類型的問句來做分析。

3-2. 意圖區段(Intention Segment)的定義

對一個自然語言問句而言，我們認為除了關鍵詞之外，仍有其他因素可用來分辨問句間的差異。觀察下面三個問句：「怎麼治療感冒？」、「為什麼要治療感冒？」、「治療感冒的方法有哪些？」。如果只考慮關鍵詞，則「治療」和「感冒」都為以上三句的關鍵詞。如此一來，我們就無法從關鍵詞來判斷第一和第三句應該較接近，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「意圖區段」，我們將其定義為：「問句中所傳達最直接想獲得的答案，不需包含前提；IS 可以是問句之子句或片語，甚至結合其他特定片語而成。」透過對於問句的分析，意義相同卻以不同句型表現的問句，所萃取出來的 IS 應該能夠保持相同。如表 1 所示，透過的 KS 及 IS 的萃取，我們可以輕易地分辨上述例句的異同。

表 1 三個相似問句所對應之關鍵詞區段(KS)及意圖區段(IS)

問句	KS	IS
怎麼治療感冒？	治療、感冒	治療感冒的方法
為什麼要治療感冒？	治療、感冒	治療感冒的原因
治療感冒的方法有哪些？	治療、感冒	治療感冒的方法

3-3. 意圖的萃取

由上一個小節的說明得知，如果能從問句中正確的萃取出 IS，對於問句意圖的辨析有很大的幫助。從語言學的角度來看，問句的語意與問句的句型息息相關。我們針對三種最常被

用在網路上的問句類型進行分析，研究問句在各種句型結構下的意圖。

3-3-1. 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[3]。中文存在有許多疑問詞，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問句的意圖，例如問句中如果問到「為什麼」，幾乎可以想見的該句就是在問某件事情或現象的原因；但是，有些疑問詞會隨著在句子中的相對語法位置不同，其意義也不盡相同。如表 2 所示，「怎麼」這個疑問詞，若出現在副詞之前可做為詢問某件事情或現象的原因，但若出現在動詞之前卻做為詢問做某件事的方法[16]。

表 2 疑問詞「怎麼」的意圖因語法位置的不同而有所不同

問句	意圖
要怎麼治療口臭？	治療口臭的方法
你怎麼會(能/可以)離開？	離開的原因

3-3-2. 句末語助詞為「嗎」的問句

句末語助詞問句指句子末端帶有一個語助詞像是「嗎」、「吧」、「呢」、「啊」等。當語助詞為「嗎」時，該問句對於答案相當不肯定，而需要較多的外在訊息給予解答。這類型的問句在句子中通常會包含一個「法相(modality)副詞」[16]，如「會」、「可能」、「應該」。「法相」的定義是「說話者的對一個可能事件的看法或態度」，法相副詞的定義由語意規定，其所包含的詞性含有以往語言學分類中的大多數助動詞、部分動詞及動詞，但他們卻有許多共同的語法特色。而法相副詞之後所接的是動詞片語，我們認為此動詞片語即為其意圖所在。表 3 中列舉出部分句末語助詞為「嗎」的問句及其對應的 IS。此外，如果這類型問句不含有任何法相副詞，則以主要動詞片語作為 IS，如表 4 所示。

表 3 句末語助詞為「嗎」的問句及其對應的意圖區段(IS)

問句	IS
把脈能診斷出所有疾病嗎？	能診斷出所有疾病嗎
肝炎病人應戒酒嗎？	應戒酒嗎

表 4 不含法相副詞之句末語助詞為「嗎」的問句及其對應的意圖區段(IS)

問句	IS
急性 C 型肝炎可怕嗎？	可怕嗎
子宮切片的結果正確嗎？	正確嗎

3-3-3. 是非問句

是非問句是指包含具有 A-not-AB 或是 A-not-A 特性之詞組的問句，例如：「是不是」、「可不可以」、「是否」。是非問句和句末語助詞為「嗎」的問句，相對於英文便是由 be 動詞或是助動詞開頭的問句，這兩類問句在結構上是可以互換的。同樣地，表現在 IS 上面，相同語意不同句型的問句也會具有相同的 IS。對是非問句而言，我們認為意圖為接在 A-not-A 詞組之後動詞片語。表 5 列舉出部分是非問句及其對應的 IS。

表 5 部分是非問句及其對應的意圖區段(IS)

問句	IS
感染 B 型肝炎後會不會自動痊癒？	會自動痊癒嗎
哺乳的媽媽感冒可不可以服用藥物？	可以服用藥物嗎

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與語意的語意文法，當問句符合語意文法中某一則時，其相對應的 IS 之萃取方式也清楚的被規範著。表 6 列舉部分語意文法及其 IS 萃取方式，並舉例說明之。

表 6 部份語意文法及其例句

問句類型	問句	語意文法	IS
疑問詞問句	為什麼產後必須服用生化湯？	QW_1 NP Dba VP →IS=VP 的原因	服用生化湯的原因
句末語助詞問句	肝炎病人應戒酒嗎？	NP Dba VP →IS=VP	應戒酒嗎
是非問句	哺乳中的媽媽感冒可不可以服用藥物？	P Dba1 not Dba2 VP →IS=Dba2 VP	可以服用藥物嗎

3-4. 關鍵詞的萃取

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可從問句找出其 KS。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 做為斷詞及詞性標記的工具，此軟體為中研院資訊所 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。

一般在做關鍵詞查詢時，多半用的是名詞或動詞，所以斷詞後，我們先從句子中找出名詞及動詞的部分。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做關鍵詞用，如定詞(Ne)、量詞(Nf)、方位詞(Ng)以及代名詞(Nh)，我們把這些詞類的詞視為非關鍵詞。

另外，有些詞雖然符合以上規則，但是出現頻率卻相當高；相對而言，其重要性便降低，視為非關鍵詞。經由統計語料庫可得到一些詞頻，將高頻的詞經過人工篩選建立一個停用詞詞典(stopping word dictionary)，當一個詞出現在停用詞詞典中，便將之從關鍵詞組裡去除。表 7 列舉出部分問句及其對應的 KS。

表 7 問句及其相對應關鍵詞區段(KS)之範例

問句	KS
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)

4. 詞意比對

本論文中，詞意比對是所有語意比對方法的基礎，傳統語言學認為詞是構成語意的最小單元[19]，而目前計算語言學的趨向是把詞視為許多「語意成分」(semantic features)的組合。基於後者，我們利用知網(How-net)[17]作為詞意比對的知識庫。

4-1. 知網概述

知網是針對電腦設計的雙語常識知識庫，為創建人董振東先生研究十幾年的重要成果，提供了設計人工智慧軟體所需的知識。知網共收錄了 50220 個中文詞語，所涵蓋的概念總量達 62174 個，目前仍在擴充當中。做為一個提供中文計算需求的知識庫，知網詳盡地描述了概念之間的關係，概念所具有的屬性之間的關係，以及概念與所具有的屬性之間的關係。

對一個詞而言，在不同情況下可能代表不同的概念。知網將一個概念的定義表示成特徵及標識符號的組合。表 8 列舉幾個概念在知網中之定義，其中 W_C 為一概念，G_C 表示該概念的詞類，DEF 則為其定義。在定義中，特徵間以逗號區隔，第一個特徵稱為主要特徵，表示概念的類別屬性，具有上下位關係，如圖 2 所示；後面所接的特徵則為次要特徵，用來詳細規範概念的屬性。

表 8 How-net 定義範例

W_C	G_C	DEF
警察	N	human|人, police|警
病人	N	human|人, *SufferFrom|罹患, $cure|醫治, #medical|醫, undesired|莠
鮮花	N	FlowerGrass|花草

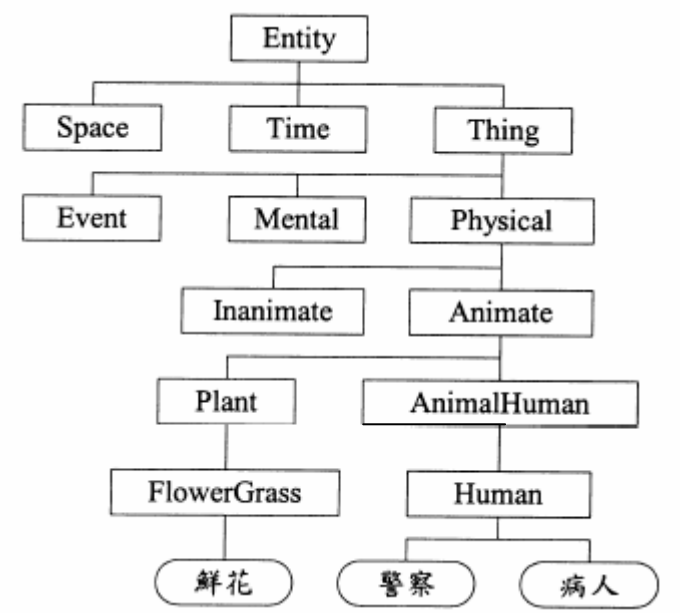

圖 2 How-net 主要特徵階層圖

4-2. 詞意相似度的量測

基於對知網的研究，我們利用知網對於每個詞彙完整的定義，量測兩個詞彙在語意上的相似度。同一個詞彙通常可表示一個以上的概念，所以兩個詞彙的相似度可以由個別的概念相似度求得，而概念相似度則是透過特徵的比對而來。如公式(1)所示，任兩個詞的語意相似度(Sim_{word})被定義成這兩個詞所有可能概念定義之間相似度(Sim_{def})的最大值。

$$Sim_{word}(w_1, w_2) = \max_{d_1 \in def(w_1), d_2 \in def(w_2)} Sim_{def}(d_1, d_2) \quad (1)$$

由於我們採用 AutoTag 做為詞性標記工具，所以可排除部份詞義混淆的情形。例如：當 w_1 同時包含名詞和動詞的概念時，若其詞性標記為動詞，則其他名詞類的概念將不予考慮。由於概念的定義是由主要特徵及次要特徵所共同描述，所以任兩個概念的相似度可定義如下：

$$Sim_{def}(d_1, d_2) = \beta \cdot Sim_{PF}(pf_1, pf_2) + (1-\beta) \cdot Sim_{SF}(sf_1, sf_2) \quad (2)$$

其中 pf_1 與 pf_2 分別是 d_1 與 d_2 的主要特徵，sf_1 與 sf_2 分別是 d_1 與 d_2 的次要特徵，Sim_{PF} 與 Sim_{SF} 分別是 d_1 和 d_2 的主要特徵與次要特徵的相似度，將會在下面的小節中做介紹，特徵結合係數 β 決定主要特徵相似度與次要特徵相似度間的權重。

4-3. 主要特徵相似度

對每一個概念，知網依照其類別屬性的不同，定義在主要特徵上，而類別屬性則構成一個階層式結構。在此階層式結構中，當兩個概念間的差異性越大，所屬的節點距離也越遠，反之則越接近。同樣的情況也發生在英文的 WordNet 上，在國外不少針對 WordNet 的研究也有類似的想法[10][12][13]。參考[8]的作法，我們將這個問題從兩個角度來思考。

一、節點(node)的觀點

從節點的角度來看，每個節點都代表唯一的概念，而且包含了特定程度的資訊量。要衡量兩個概念間的相似度，可以考慮它們所共同分享的資訊量。因此，對階層式架構中任兩個節點而言，其相似度便定義為其最接近的共同祖先節點之 information content (IC)。

根據資訊理論[14]，一個特徵 x 的 IC 可以表示成：

$$IC(x) = -\log p(x) \tag{3}$$

其中 $p(x)$ 表示具有特徵 x 或 x 的祖先特徵的這些概念在語料庫中出現的機率。對於知網的主要特徵而言，其機率值從上層單調遞減到下層的節點；而 IC 值恰與機率值相反，從下層單調遞減到上層，最上層的根節點，由於機率值等於 1，所以其 IC 值等於 0。

二、邊(edge)的觀點

從邊的角度來看，任兩個概念的相似度，可以由節點間的距離來量測；其中，兩個節點間的距離越長，其相似度越小。另外，我們也考慮深度對於任兩個相鄰節點間距離的影響，我們認為深度越大其距離越短；原因在於，對越深層的節點分類時，所描述的差別就越精細。

綜合以上兩個觀點，我們定義任兩個主要特徵的距離如下：

$$Dist(pf_1, pf_2) = \sum_{c_i \in \text{the shortest path}(pf_1, pf_2) - LSuper(pf_1, pf_2)} Cost(c_i, p_i) \tag{4}$$

其中 $LSuper(pf_1, pf_2)$ 表示 pf_1 與 pf_2 的最接近之共同祖先節點，c_i 表示 pf_1 與 pf_2 間最短路徑中除了最接近之共同祖先節點外的所有節點，p_i 則為節點 c_i 的父親節點，而 $Cost$ 代表任兩個相鄰節點間的距離，其定義如公式(5)所示。

$$Cost(c_i, p_i) = \left(\frac{d(p_i)+1}{d(p_i)} \right)^{\alpha} \left[IC_{PF}(c_i) - IC_{PF}(p_i) \right] \tag{5}$$

其中 $d(p_i)$ 表示節點 p_i 的深度(depth)，α 則為控制深度對於 *Cost* 的影響的參數。另外，由於相似度恰與距離的意義相反，因此定義主要特徵相似度如下：

$$Sim_{PF}(pf_1, pf_2) = 1 - \frac{Dist(pf_1, pf_2)}{\max_{i,j} Dist(pf_i, pf_j)} \tag{6}$$

4-4. 次要特徵相似度

知網對於一個概念的定義，除了主要特徵外仍有次要特徵用以輔助標示其屬性，但是次要特徵不具有階層式關係，而且一個定義通常包含不只一個次要特徵。因此可將次要特徵表示為二元向量(binary vector)，如此一來，次要特徵相似度就可藉由量測二元向量的相似度來得到。在向量空間中，對於二元向量相似度的衡量方法有下列幾種：Dice coefficient、Jaccard coefficient、Overlap coefficient、以及 Cosine 等[11]。由於每個次要特徵的重要性不一，如果某個次要特徵經常出現在各個概念定義中，則其辨別詞意的能力就較弱，反之則愈大。因此，我們結合 Dice coefficient 與 IC 為次要特徵相似度的量測方式，定義如下：

$$Sim_{SF}(sf_1, sf_2) = \frac{2 \times \sum_{f_i \in sf_1 \cap sf_2} IC_{SF}(f_i)}{\sum_{f_j \in sf_1} IC_{SF}(f_j) + \sum_{f_k \in sf_2} IC_{SF}(f_k)} \tag{7}$$

5. 語意比對

本研究中，語意比對可分為兩個部份，一個是問句與 FAQ 問題的比對，一個是使用者詢問句中所含的關鍵詞區段跟 FAQ 答案的內文比對。因此，一個問句 q 與一則 FAQ 樣本 p 之語意相似度，可定義成公式(8)。

$$Sim(q, p) = \delta \cdot Sim_{question}(q, q(p)) + (1-\delta) \cdot Sim_{content}(q, a(p)) \tag{8}$$

其中 $Sim_{question}$ 表示該問句與 FAQ 問題之相似度，而 $Sim_{content}$ 表示該問句與 FAQ 答案之相似度，並以一比對結合係數 δ 調整兩者間的權重。

5-1. 問句比對

在問句的比對上，因為每一個問句都由 IS 以及 KS 所組成，因此我們將分別量測 IS 相似

度以及 KS 相似度之後，最後再將這兩個相似度結合成問句相似度，如公式(9)所示。

$$Sim_{question}(query, q(faq)) = \gamma \cdot Sim_{IS}(IS_{query}, IS_{question}) + (1-\gamma) \cdot Sim_{KS}(KS_{query}, KS_{question}) \quad (9)$$

其中意圖-關鍵詞結合係數γ用來調整 IS 相似度(Sim_{IS})和 KS 相似度(Sim_{KS})間的權重。

5-1-1. 意圖區段相似度

經由語意分析器所萃取出來的 IS 通常是一個簡單的名詞片語或動詞片語，但是如何去量測兩個片語間的相似程度呢？考慮下面的例子：

P1：「吃心臟病藥」

P2：「吃治療心臟病的藥」

P3：「治療心臟病的藥」

如果將「吃」、「治療」、「心臟病」、「藥」視為關鍵詞，可以想見的我們將無法分辨 P1 與 P3 何者較為接近 P2，因為 P1 和 P3 同時擁有 P2 四個詞中的三個。但是經由對語言的理解，卻可以清楚的分辨 P1 應該比較接近 P2，甚至相同。

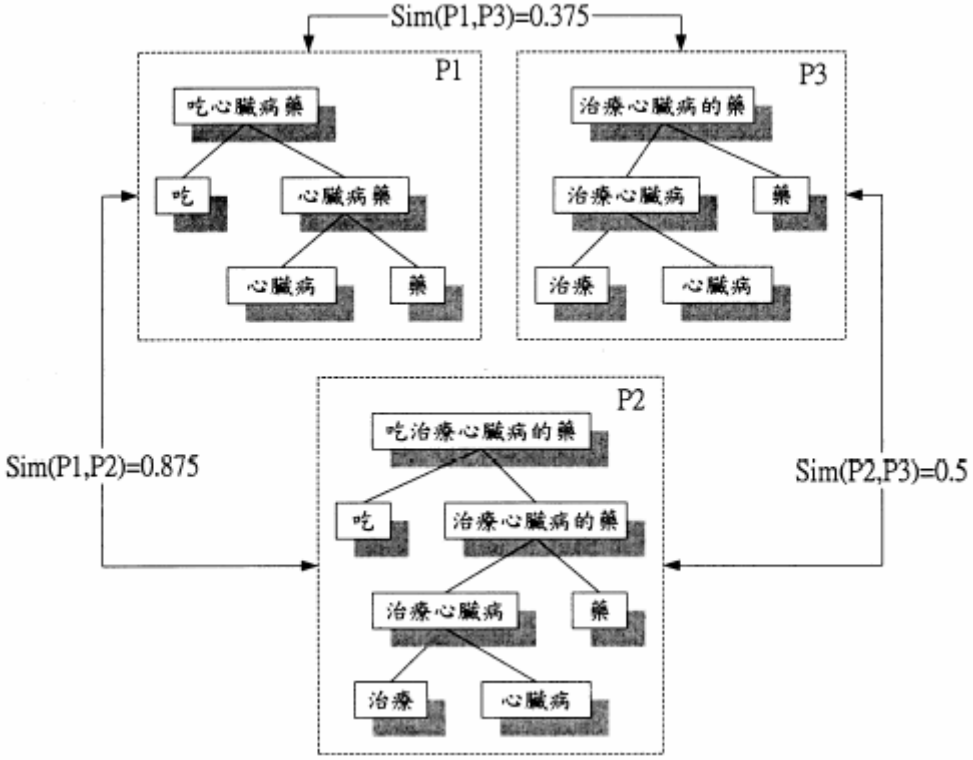

圖 3 IS 剖析樹比對示意圖

因此，我們將片語化成剖析樹(parse tree)，透過剖析樹的比對，來解決上述的問題。如圖

3，在分別建立三個片語的剖析樹之後，可以清楚地發現 P2 的「治療心臟病」這個動詞片語作為形容「藥」之修飾語，其地位相同於 P1 中的「心臟病」，亦為「藥」的修飾語。整體來看，P1 與 P2 都為動詞片語，其動詞都是「吃」，吃的對象都是「藥」，唯有在「藥」的修飾語上略有不同。另一方面，P3 恰為 P2 之子樹，相對而言，兩者之相似度應該小於前述之相似度。

我們參考 CKIP Tree Bank[2]整理部分的語法規則，再根據 Earley algorithm[6]建立一個語法剖析器。其中，若剖析樹之外部節點詞性為介詞或連接詞，則省略該分支以節省比對時間。對於任兩個 IS 剖析樹 T_1 與 T_2 ，我們定義比對公式如下：

$$
\begin{aligned}
&Sim_{IS}(IS_1, IS_2) \\
&= Sim_{tree}(T_1, T_2) \\
&= \begin{cases}
Sim_{word}(T_1, T_2), \text{若 } T_1 \text{ 和 } T_2 \text{ 都是單節點樹} \\
\dfrac{1}{|T_1|}\max\limits_i Sim_{tree}(T_{1,i}, T_2), \text{若 } T_2 \text{ 是單節點樹，而 } T_1 \text{ 不是} \\
\dfrac{1}{|T_2|}\max\limits_j Sim_{tree}(T_1, T_{2,j}), \text{若 } T_1 \text{ 是單節點樹，而 } T_2 \text{ 不是} \\
\max\{\dfrac{1}{|T_1|}\max\limits_i Sim_{subtree}(T_{1,i}, T_2), \dfrac{1}{|T_2|}\max\limits_j Sim_{subtree}(T_1, T_{2,j}), Sim_{subtree}(T_1, T_2)\}, \text{其他}
\end{cases}
\end{aligned}
\tag{10}
$$

其中 $Sim_{word}(T_1, T_2)$ 表示兩個單節點 IS 剖析樹間的相似度，$T_{1,i}$ 和 $T_{2,j}$ 分別表示 T_1 和 T_2 的子樹，$|T_1|$ 和 $|T_2|$ 分別表示 T_1 和 T_2 子樹的個數，$Sim_{subtree}$ 表示兩個非單節點 IS 剖析樹間的相似度，其定義如下：

$$
Sim_{subtree}(T_1, T_2) = \max_g \frac{\sum_{k=1}^{|T_A|} Sim_{tree}(T_{A,k}, g(T_{A,k}))}{|T_A|} \tag{11}
$$

其中 g 是一個從 T_A 到 T_B 的一對一函數，$T_{A,k}$ 表示 T_A 的一個子樹，$|T_A|$ 表示 T_A 子樹的個數。由於 g 為一對一函數，所以 $|T_A| \le |T_B|$，因此需要特別注意：若 $|T_1| \le |T_2|$，則設定 $T_A = T_1$ 且 $T_B = T_2$，否則設定 $T_A = T_2$ 且 $T_B = T_1$。

當 T_1 和 T_2 都是外部節點的時候，表示此二者皆為詞，對於兩個詞的相似度，就利用公式(1)所描述的詞意相似度來量測。當 T_1 或 T_2 其中之一為外部節點時，表示其中一個為詞另一個則為一個片語，此時則遞迴向下找出該片語中與該詞最相似的詞。當 T_1 和 T_2 都不為外部節點

時，就表示T_1和T_2都含有各自的子樹。此時，可以從三個方向來思考：最基本的想法，若兩顆樹的所有子樹都非常相似，則這兩顆樹可能是非常相似的，因此考慮$Sim_{subtree}(T_1,T_2)$作為T_1和T_2的相似度；另外，如果T_1相似於T_2的一個子樹，或是T_2相似於T_1的一個子樹，則根據分支的多寡來決定該相似度之權重。

5-1-2. 關鍵詞區段相似度

在量測兩個 KS 的相似度上，我們做了一個假設：對任一個關鍵詞而言，不會有兩個或兩個以上的關鍵詞與它對應。而這種對應關係恰可以一對一對應函數表示之，所以我們提出公式(12)來量測兩個 KSs $K_1=\{w_1,w_2,\cdots,w_m\}$和$K_2=\{t_1,t_2,\cdots,t_n\}$的相似度。

$$Sim_{KS}(K_1,K_2)=\max_f \frac{\sum_{i=1}^{|A|} Sim_{word}(a_i,f(a_i))}{|A|} \tag{12}$$

其中f是一個從A到B的一對一函數，a_i是A中的一個元素，$Sim_{word}(a_i,f(a_i))$表示關鍵詞a_i與其對應的關鍵詞的詞意相似度。如同前一小節，需特別注意：若$m\le n$，則設定$A=K_1$且$B=K_2$；反之，則設定$A=K_2$且$B=K_1$。

5-2. 內文比對

除了問句比對外，我們也利用問句與 FAQ 答案的比對來協助找出所需的答案，使用的方法則是目前被廣泛使用在資訊檢索應用的 vector space model (VSM)。VSM 主要分成兩個步驟：(1) 萃取特徵並以向量來描述之，(2)比較兩個特徵向量在向量空間中的夾角。本研究中，特徵向量是由每個關鍵詞的 TF×IDF 權重所構成。針對問句及 FAQ 答案求取個別的特徵向量$\vec{u}=\{a_1,a_2,\cdots,a_N\}$和$\vec{v}=\{b_1,b_2,\cdots,b_N\}$；然後利用餘弦公式計算其夾角，夾角愈小表示兩向量愈接近，以此做為該問句與 FAQ 答案的相關程度，如公式(13)所示。

$$Sim_{content}(\vec{u},\vec{v})=\rho_{\cos}(\vec{u},\vec{v})=\frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}} \tag{13}$$

其中N表示特徵向量的維度，也就是詞彙量。

6. 實驗結果與討論

本研究中，我們實驗使用的機器為 Pentium III 450 個人電腦，128 MB RAM，開發用的程式語言是 Microsoft Visual C++ 6.0。除了實驗測試之外，也透過 IIS 4.0 架設了一個網站，開放給網路上的使用者查詢，網址在 http://chinese.csie.ncku.edu.tw/faq/。在語料庫的收集方面，我們以人工在網路上收集了 1,022 則 FAQ，內容主要包括醫藥以及投資理財相關之 FAQ。

在系統評估方面，我們請 10 位非系統開發人員，並告知本網站所提供資訊的內容範圍，以人工的方式建立 185 則問句並標記與其相關之 FAQ。有別於關鍵詞資訊檢索，自然語言問句之意圖較明確，因此每則問句所對應的答案相當少，平均只有 1.36 則。因此，我們不使用精確率(precision rate)來衡量系統的效能，因為即使第一名就是正確答案，精確率仍會隨著名次增加而遞減。我們提出一個較恰當的評估方式—平均正確答案排名，其定義如下：

$$\text{平均正確答案排名}(AvgRank)=\frac{\sum \text{正確答案所在之名次}}{\text{正確答案個數}} \tag{14}$$

6-1. 意圖區段萃取實驗

根據語料庫中問句的語法型態，訂定了 85 條語意文法。為了測試根據該語意文法所萃取出來的 IS 的正確性，以人工建立 185 則問句來做測試，並以人工檢驗是否符合原本預期的結果。檢驗時，若其誤差不影響意圖的辨別，則視為正確萃取，經統計可達到 91.89%的正確萃取率，其中無法正確萃取的情況可分為以下幾種：

一、 屬於疑問詞問句、是非語句、句末語助詞為「嗎」之外的問句，由於並未在語意文法中定義其萃取方式，所以屬於「超越文法範圍 (out-of-grammar) 」而無法萃取。

二、 問句結構過於複雜甚至帶有兩個疑問子句，對於這類型問句目前仍無法處理。

三、 在 AutoTag 斷詞及標示詞性時已經出錯，導致後面意圖萃取無法正確判斷。

6-2. 基準系統

本實驗以關鍵詞查詢為基準(baseline)，與自然語言查詢做比較。因此我們令公式(8)中的係數 $\delta=0$，使得僅由內容比對來決定整體之相似度。經由統計每一條測試句之答案排名，結果獲得平均正確答案排名為 12.04 名，並得到前 N 名的召回率(recall rate)表列如下：

表 9 基線系統之前 N 名召回率

Top N	1	2	3	4	5	6	7	8	9	10
召回率 (%)	36.06	48.56	56.00	60.22	63.89	66.56	73.56	73.56	76.72	78.06

6-3. 詞意相似度之實驗

6-3-1. 主要特徵相似度之深度影響係數實驗

從公式(5)中得知，係數α決定深度對於任兩相鄰節點間距離(*Cost*)的影響程度，為了找出α之最佳值，我們固定公式(2)中的係數$\beta=1$，也就是完全以主要特徵相似度做為詞意相似度；公式(9)中的係數$\gamma=0$，表示不考慮 IS 對問句相似度之影響；公式(8)中的係數$\delta=1$，表示完全以問句相似度作為檢索的依據，然後根據平均正確答案排名來決定α之最佳值。

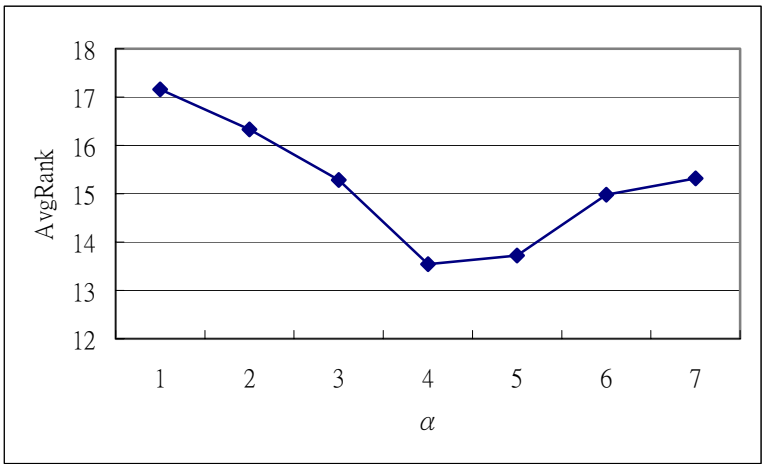

圖 4 係數α相對於平均正確答案排名之比較圖

如圖 4 所示，當$\alpha=4.0$時，其平均正確答案排名 13.54 為最佳結果，此結果與「深度越深則節點間距離越短」的觀點相符。

6-3-2. 次要特徵相似度計算方式實驗

本實驗比較四種二元向量相似度量測方式對系統效能的影響。我們固定$\beta=0$，也就是完全以次要特徵相似度為主，$\gamma=0$即不考慮 IS，$\delta=1$完全以問句比對來評估，結果表列如下：

表 10 比較各種二元向量相似度量測係數對系統平均正確答案排名之影響

	Dice coefficient	Jaccard coefficient	Overlap coefficient	Cosine
平均正確答案排名	6.28	6.29	7.61	6.51

表 10 顯示使用 Dice coefficient 之結果為最佳，所以在接下來的實驗都採用 Dice coefficient 來作為次要特徵相似度之量測方法。

6-3-3. 主要特徵與次要特徵之結合係數實驗

概念定義的相似度由主要特徵相似度及次要特徵相似度結合而來，因此本實驗的希望得到特徵結合係數 β 對系統效能的影響，同樣地，我們固定係數 $\gamma=0$ 與 $\delta=1$。如圖 5 所示，該實驗結果顯示，$\beta=0.3$ 使得平均正確答案排名達到 5.89 為最小。

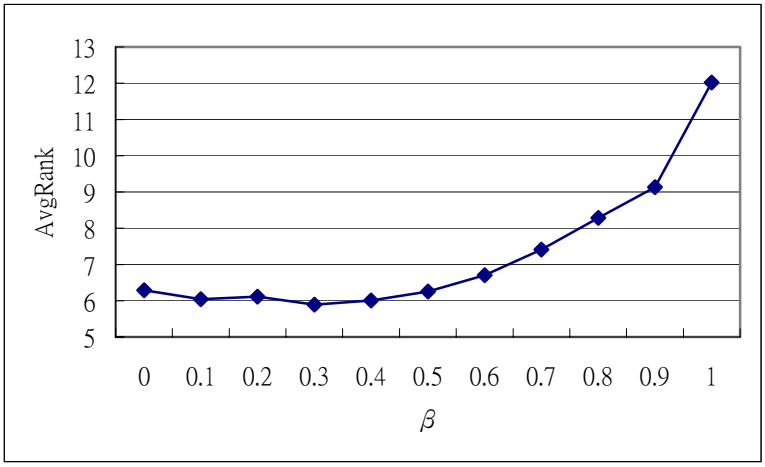

圖 5 特徵結合係數正確答案排名比較圖

6-4. 句意相似度實驗

由公式(9)，本實驗想了解意圖-關鍵詞結合係數 γ 對系統效能的影響，因此固定實驗值 $\alpha=4$ 與 $\beta=0.3$ 以及尚未實驗的 $\delta=1$。由圖 6 得知，當 $\gamma=0.3$ 時，其平均正確答案排名 3.59 為最佳結果。此外，當 γ 較大時，曲線迅速上揚，表示當 IS 相似度的比重過大時，其結果並不理想。這是因為 IS 僅包含問題的意圖，並未將前提包含進來；因此，IS 並不能完全取代 KS，而是相輔相成。

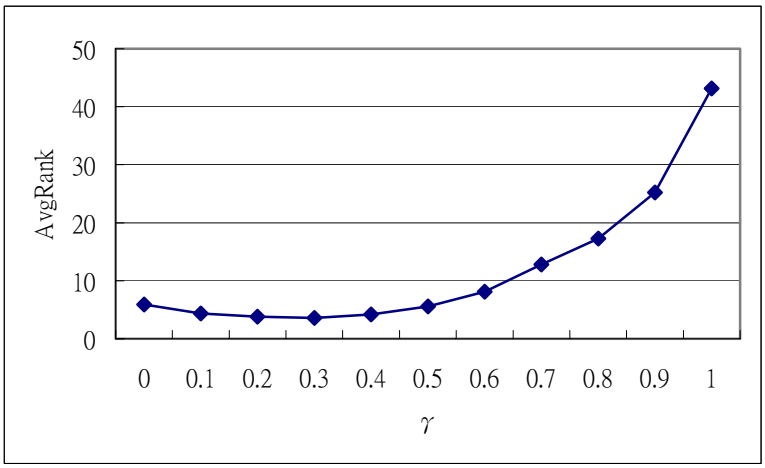

圖 6 意圖-關鍵詞結合係數γ之於平均正確答案排名比較圖

6-5. 問句相似度與內文相似度之結合係數實驗

由公式(8)，問句與 FAQ 樣本的比對由問句的相似度與內文相似度共同決定，因此本小節實驗其係數δ。實驗結果顯示，$\delta = 0.5$ 時，其平均正確答案排名落在 2.91 為最佳結果。觀察圖 7，δ在範圍[0.2, 1.0]中時，對系統效能的影響並不大；可得知，相較於內文相似度，問句相似度對系統效能的影響較大。

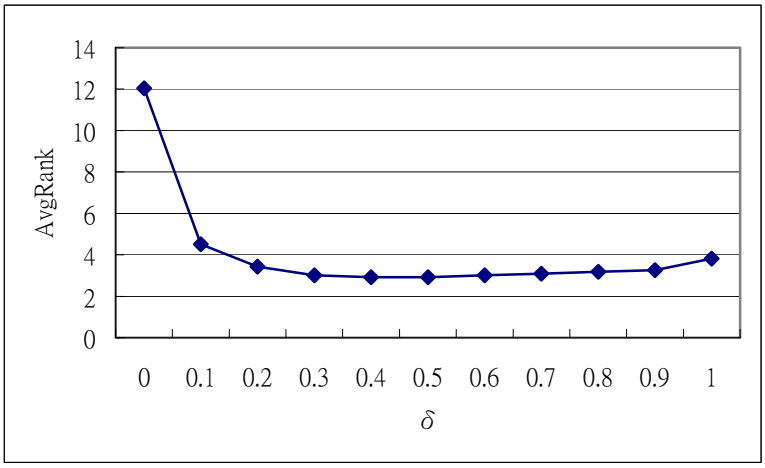

圖 7 比對結合參數δ之於平均正確答案排名比較圖

6-6. 實驗總結

最後，藉由控制參數，將各個方法以平均正確答案排名與召回率做一個比較。由圖(8)和圖(9)可以發現，無論從平均正確答案排名或是前 N 名的召回率來看，本論文所提出的方法明顯地改善了效能。相較於基準系統，平均正確答案排名約進步了 9 個名次。 第一名的召回率

從 36.06%提升到 64.67%，約提昇了 80%；而前十名的召回率也從 78.06%提升到 95.11%。

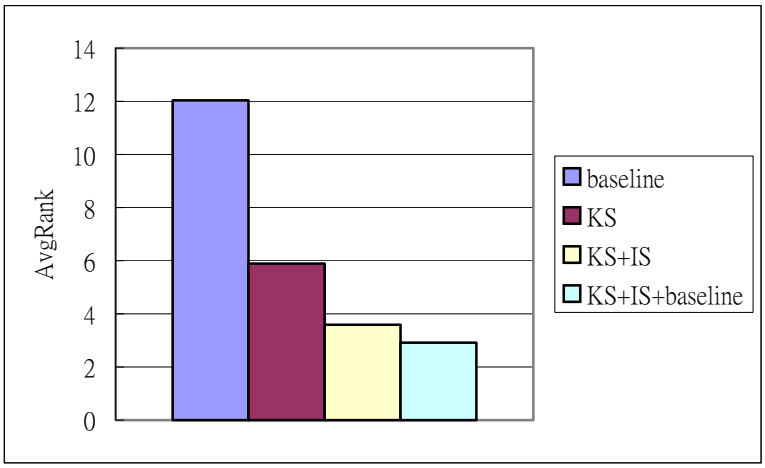

圖 8 系統平均正確答案排名比較圖，其中 baseline 表示只比較內文的關鍵詞，KS 表示只比較問句的關鍵詞，IS 表示只比較問句的意圖

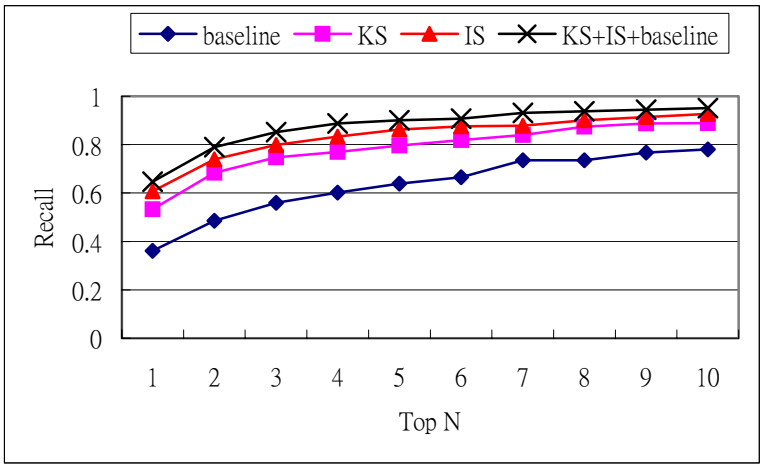

圖 9 系統召回率比較圖，其中 baseline 表示只比較內文的關鍵詞，KS 表示只比較問句的關鍵詞，IS 表示只比較問句的意圖

7. 結論與未來展望

本論文提出以問句意圖萃取以及語意比對的方法，應用到自然語言 FAQ 檢索上。經實驗驗證，該方法確實比單純使用關鍵詞查詢來得準確，使平均正確答案的排名從第 12.04 名提升到第 2.91 名，且使得前十名的召回率由 78.06%提升到 95.11%，但是其中仍存在一些待改進之處：

一、意圖萃取方面，雖然我們能處理 92%的語料，仍有許多問句的型態不在收集的範圍內，以及對於較複雜語法問句的誤判，可以藉由改善語意文法上來解決。

二、在語意相似度方面，我們採用知網做為詞意相似度量測的知識庫，但是知網中沒有定義的詞，則無法藉由它來量測詞意相似度。解決的方法有二：一是增加未定義詞到知識庫中，另一個是找出自動建立知識庫的方法。

三、在建立意圖區段剖析樹方面，對於剖析時普遍遭遇到詞性不明確的問題(ambiguity)，仍有困難無法克服。考慮現有資源，可以先建立機率剖析器[4][15]，進而建立包含語意之剖析器。

四、在自然語言理解方面，目前的系統並未具備推理能力，在許多情況下，詞語的組合可能引申另外的意義。這些會遭遇到但仍無法解決的問題，有待未來持續地研究。

參考文獻

[1] Ask Jeeves, http://www.ask.com.

[2] CKIP Tree Bank, http://godel.iis.sinica.edu.tw/CKIP/trees1000.htm.

[3] Chang, Chung-Yin, "A Discourse Analysis of Questions in Mandarin Conversion," M.A. Thesis, National Taiwan University Graduate Institute of Linguistics, June 1997, pp. 16-81.

[4] Collins, M. J., "Head-driven Statistical Models for Natural Language Parsing," Ph.D. Thesis, University of Pennsylvania, Philadelphia, 1999.

[5] Dr. E, http://drdai.polaris.com.tw.

[6] Earley, J., "An Efficient Context-free Parsing Algorithm," Communications of the ACM, vol. 6, no. 8, 1970, pp. 451-455.

[7] FAQ Finder, http://faqfinder.ics.uci.edu:8001.

[8] Jiang, Jay J. and David W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," Proceedings of the ROCLING X, 1997, pp. 19-33.

[9] Li, Charles and Sandra A. Thompson, "Mandarin Chinese: A functional reference grammar," Berkeley and Los Angeles: University of California Press, 1981.

[10] Lin, D., "An Information-Theoretic Definition of Similarity," Proceedings of the International Conference on Machine Learning, July 1998.

[11] Manning, Christopher D. and Hinrich Schütze, "Foundations of Statistical Natural Language Processing," The MIT Press, 1999, pp. 296-303.

[12] Markman, A. B. and D. Gentner, "Structural Alignment During Similarity Comparisons," Cognitive Psychology, vol. 25, 1993, pp. 431-467.

[13] Resnik, P., “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, August 1995, pp. 448-453.

[14] Ross, S., “A First Course in Probability,” Macmillan, 1994.

[15] Stolcke, A., “An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities,” Computational Linguistics, vol. 21, no. 2, 1995, pp. 165-202.

[16] 張麗麗, “現代漢語中的法相詞”, CKIP Technical Report, no.93-06, June 1993, pp. 1-16.

[17] 董振東, 董強, “知網”, http://how-net.com.

[18] 蔡維天, “The Hows of Why and the Whys of How”, 台灣語言學的創造力學術研討會, 2000, pp.1-27.

[19] 謝國平, “語言學概論”, 三民書局, 1996, pp.189-197.