電信服務業務 FAQ 查詢系統之研製

梁 婷[1]　　歐陽彥隆[2]
[1]國立交通大學資訊科學研究所　[2]中華電信公司
correspondence[1]: tliang@cis.nctu.edu.tw

摘要

FAQ 是資訊檢索的一個應用方向，這些經過人工整理，以一問一答方式呈現的資料，如能有效應用，將可提高使用者進行資訊檢索的滿意度。在本論文中，我們將針對電信服務業務相關查詢作業，研製一個口語化的 FAQ 查詢系統。此系統可透過簡單操作之滑鼠選單及鍵盤輸入介面，結合動畫的活潑化及語音聲效之回應，來作客戶服務，便利使用者查詢資料。另一方面本系統也可提供客服中心客服人員一個簡便之查詢輔助工具。藉由系統之多種查詢方式，可迅速查客戶到所要之資料，回答客戶之問題。此種以口語化查詢方法來取代關鍵詞查詢，可避免以關鍵詞查詢找到過多非相關資料的缺點，提高查詢資料之資訊檢索的滿意度，同時透過業務分類、選單、鍵詞瀏覽等輔助功能提高系統的便利度。

關鍵詞: FAQ, 資訊檢索, 分類, 自然語言, 容錯, 對談管理。

1. 緒論

FAQ(Frequently Ask Question)中文稱為常見問題集、答客問或你問我答等。FAQ 查詢是資訊檢索的一個應用方向。例如電信、金融、保險及證券等擁有眾多客戶之企業或公司，基於其業務需要，將客戶常常會詢問到的一些問題及業務說明內容，以一問一答的條列方式，來作業務說明或回答客戶問題。客服中心之客服人員在回答客戶相類似之問題，就有標準答案可迅速回答以提高客戶滿意度。另方面客戶亦可就這些 FAQ 中快速找到答案，獲得相關資訊。

目前國外之 FAQ 系統如 FAQ Finder 及 Ask Jeeves 等系統，都是一個使用自然語言的 Web-Based FAQ 查詢系統[1, 2, 19]。國內部份如網際智慧公司所建立之寶來 E 博士證券業務 FAQ 查詢系統，使用者可以輸入口語化句子來查詢相關證券業務 FAQ [18]。此種口語化查詢的查詢介面，若再結合語音辨識，將有助於系統的親和力及使用率。

設計口語化查詢機制主要包括輸入理解及對談管理處理。輸入理解是了解使用者話語所表達之意念，經由斷詞、詞類標註、語法剖析及語意分析等產生語意框架[13, 14, 15]。對談管理則是提供一個人機對談管理的機制[5]。對談管理將依照語意框架內各個 slot 之資料值，判斷是否有足夠條件資料，向後端資料庫擷取資料。當條件資料足夠時，對談管理將此框架資料轉成適當之資料庫指令，向資料庫查詢資料。當條件資料不足時，對談管理會提示使用者，請使用者補充輸入資料，以便系統有足夠條件資料，向後端資料庫提出查詢指令。同時在系統擷取資料後對談管理依照資料的屬性及資料量多寡，直接顯示資料或提示使用者再度輸入篩選條件，以便找出更符合使用者所想要之資料。對談管理的設計需要考量回應的策略及回應的順序 [3, 5, 6, 9, 11]。

在本論文中我們將探討如何設計一個口語化 FAQ 查詢機制，以便利使用者使用及查詢資料。我們針對中華電信公司電信業務實作一個 FAQ 查詢系統。此系統可透過簡單操作之滑鼠選單及鍵盤輸入等人機介面，結合動畫的活潑化及語音聲效之回應，來作客戶服務，便利使用者查詢資料。另一方面本系統也可提供客

服中心客服人員一個簡便之查詢輔助工具。藉由系統之多種查詢方式，可迅速查客戶到所要之資料，回答客戶之問題。此種以口語化查詢方法來取代關鍵詞查詢，可避免以關鍵詞查詢找到過多非相關資料的缺點，提高查詢資料之資訊檢索的滿意度，同時透過業務分類、選單、鍵詞瀏覽等輔助功能提高系統的友善度。系統設計將包函 FAQ 資料的蒐集、FAQ 意圖的分類、鍵詞搜詢、同音容錯、FAQ 問句與查詢間相似度的計算、對談管理控制及語境資料結構。細節在其他段落說明。

本論文其他段落結構為:第二段說明本系統的架構及發展步驟，第三段說明本系統之功能及比較，第四段為結論與未來研究。

2. 系統架構

圖 1 為本系統之架構流程圖。關鍵詞搜尋模組利用本系統所建立之關鍵詞庫，以最長詞匹配原則，將使用者查詢句子斷成詞串，並標註每個詞的詞類屬性及其所屬 slot。搜尋出來的關鍵詞若有同義詞，則系統會將該關鍵詞替換成一致之同義詞來代表，以便搜詢 FAQ 資料、計算 FAQ 問句與使用者詢問句之相似度及排秩。

容錯處理模組主要作用是當使用者查詢句無法與關鍵詞庫中鍵詞完全匹配時，則執行同音字容錯檢查。同音字容錯是先將使用者查詢句透過同音字轉換程序，轉成為同音代表碼的同音句，再與關鍵詞庫中鍵詞同音詞比對，並計算同音句與鍵詞同音詞的編輯距離。當編輯距離低於一門檻值時，則視之為鍵詞。同音字轉換程序則利用同音字庫作為轉換的基礎，將所有相同發音之中文字全部轉成一致的同音代表碼。另外系統會在關鍵詞搜尋後產生一個查詢框架，依照關鍵詞的詞類屬性，將內容填入框架中對應的 slot 欄位內。

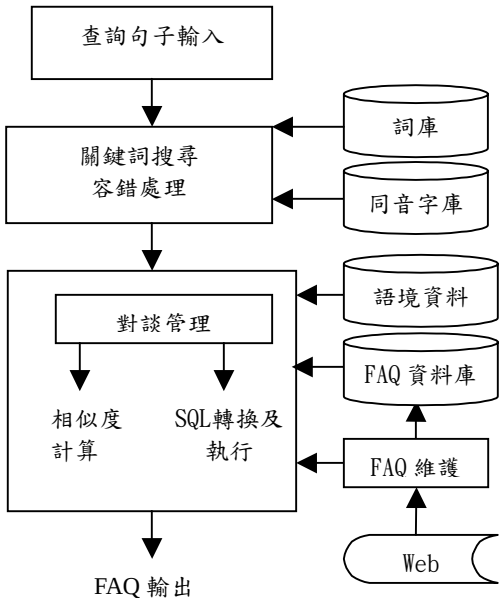

圖 1:系統流程架構

對談管理主要作用在於控制流程。根據使用者輸入之查詢句子及系統目前已有的資料，判斷是否有足夠資料向資料庫查詢。若資訊充足，則對談控制將使用者所欲查詢之內容轉成相關的資料庫查詢語言，向資料庫提出需求。如果使用者所提供之訊息資料不夠充份時，則透過聲音、動畫之提示，請使用者補充輸入系統所需的資料。SQL 轉換及執行模組依照框架中各 slot 欄位之資料，轉成資料庫對應之 SQL 指令，向後端資料庫提出 SQL 執行指令，執行資料庫查詢。相似度計算模組用來計算使用者查詢句子與查詢結果的 FAQ 兩者間之相似度。系統再依照相似度大小，由大至小降冪排秩。另外 FAQ 資料維護模組主要是擷取網頁上原始的 FAQ 資料並將這些原始資料轉入 FAQ 資料庫，同時負責 FAQ 資料及索引檔案建檔及維護的工作。

2.1 資料蒐集及分類

首先自中華電信公司企業網站首頁(http://www.cht.com.tw)上蒐集所要做實驗的 FAQ 範例問題資料，共 485 則中英文混合問題及解答。每筆 FAQ 問句長度統計資料(其長度

包括標點符號)問句平均長度為 34 bytes。FAQ 意圖有單一及多重意圖兩種，因系統對每個 FAQ 只能處理單一意圖問題，因此希望每個 FAQ 都是只有一個問題。若一個 FAQ 中同時詢問兩個以上的問題(如例 2-1)，則將此 FAQ 問句拆成兩個 FAQ 問句，因此目前共有 512 個 FAQ 問句。

(例 2-1): 原始問句:

行動電話的「話中插接」，如何申請？ 如何計費？

拆成兩個問句:

A:行動電話的「話中插接」，如何申請？

B:行動電話的「話中插接」，如何計費？

系統為了提供 FAQ 分類查詢的功能，參考中華電信公司現有業務的分類，將之分成 9 個大業務類別(如表 1)，再將每個大業務類別依其業務性質再分成若干子業務類別，共分成 55 個子業務類別，表 2 為其部份資料。

表 1: 電信業務種類一覽表。

項次	業務種類
01	市內電話
02	國內長途電話
03	行動通信
04	國際通信
05	數據通信
06	整體服務數位網路
07	衛星通信
08	智慧型網路
09	公用電話

表 2: 電信業務種類及子業務項目部份資料。

項次	業務種類	業務子種類
0100	市內電話	
・・・	・・・	・・・
0110		話中插接
0111		三方通話
0112		勿干擾
0113		電話信箱
・・・	・・・	・・・

考量公司未來業務的擴展性，我們將業務類別及子業務類別分別以 2 位數及 4 位數代碼來代表，因此我們就最多可有 99 個業務類別及 99 個子業務類別，方便本系統未來的擴展。

2.2 詢問句意圖分類

詢問句意圖分類就是客戶對電信業務服務，可能會發問的問題的分類。例如使用者可能想會問「如何申請某項業務」、「如何申告障礙」、「如何使用」、「如何申訴」等問題類型，我們將這些問題都歸屬於「如何」這個大意圖類型，而「申請」、「申告障礙」、「使」、「申訴」都是屬於「如何」這個大意圖類型下的小意圖類型。使用者陳述同一類意圖的說法，可有多種表達方式，因此系統最主要的就是要找出使用者所要詢問的意圖種類。例如一個使用者對「三方通話」這項業務不太了解，於是可用下面這幾種說法來向系統查詢：

“什麼是三方通話？”

“三方通話是什麼？”

“解釋一下什麼是三方通話？”

三種說法雖略有不同，但我們由”什麼是”、”是什麼”、“解釋”這些詞來幫我們判斷出使用者的主要目的是希望系統能對”三方通話”這個業務名詞作一解釋，因此我們將之歸屬於「名詞解釋」<Q-EXPLAIN>這類意圖類型分類，其代碼為 0301。

我們對意圖種類的分類，以 5W1H(Who，When，Where，What，Why，How)為大分類核心。每個大分類下依照業務性質再酌分出幾個小類別，例如<Q-HOW>這類大意圖之下又分出<Q-HOW-APPLY>、<Q-HOW-INQUIRE>、<Q-HOW-INFORM>等等這些小分類，目前共分成 60 餘種意圖分類。我們分別以 2 個數字代號來代表意圖之大類型及小類型，當有新的 FAQ 問題時，可很方便的擴展系統編碼。表 3 為其部份意圖類型資料。

表 3:句子意圖類型資料。

類型代碼	次類型代碼	句子意圖表達類型	句子表達意義	用詞範例
・・・	・・・	・・・	・・・	・・・
03	01	<Q-EXPLAIN>	業務或名詞解釋	什麼是 | 是什麼 | 解釋 | 說明一下 | 何謂 | 意思
04	00	<Q-HOW>	如何	
	01	<Q-HOW-APPLY>	如何申請	<如何>申請
	02	<Q-HOW-INQUIRE>	如何查詢	<如何>查詢
	03	<Q-HOW-INFORM>	如何申告	<如何>申告

表 4: 意圖類型分佈。

意圖代碼	意圖	資料筆數	比率	意圖名稱
01	<Q-COST>	70	13.67%	多少費用(固定費用)
02	<Q-COMPARE>	8	1.56%	比較差異 有何不同 | 有何差異| 有什麼不一樣 | 那裡
03	<Q-EXPLAIN>	69	13.48%	業務或名詞解釋 什麼是是什麼 | 解釋 | 說明一
04	<Q-HOW>	189	36.91%	如何解決問題
05	<Q-LOC>	8	1.56%	地點位置 到那裡 | 什麼地方 | 在那裡 | 到哪裡 | 向
07	<Q-TIME>	10	1.95%	時間日期 什麼時候 | 那一天 | 哪一天 | 何時
08	<Q-WHAT>	97	18.95%	文件證明 何種證件| 什麼證件 | 什麼東西 | 那些證件
09	<Q-WHO>	1	0.20%	找誰 找誰 | 找什麼人 | 向什麼人 | 對象為何
10	<Q-WHY>	27	5.27%	原因 為什麼 | 怎麼這樣 | 為何 | 為什麼 |原因
11	<Q-YES-NO>	27	5.27%	是否 是不是 | 對不對 | 好不好 | 有無限制 | 是否
90	<Q-MISC>	6	1.17%	無法歸類 | 雜項

我們分析現有 FAQ 問句問題內容，得到表 4 問題種類分佈狀況表，其中以意圖<Q-HOW>所佔 36.91%最多，其次分別為<Q-WHAT> 18.95%、<Q-EXPLAIN> 13.48%及<Q-COST>13.67%等意圖。所以目前 FAQ 主要問題範圍是功能說明(名詞解釋)、操作方法、申請手續(文件)及費用等四大類。

2.3 FAQ 資料儲存表示

因為 FAQ 資料為非結構化之資料，我們要將 FAQ 資料轉成結構化資料，以便利用資料庫來處理。首先我們要將每一 FAQ 問句找出足以代表此 FAQ 之屬性、特徵，以便查詢、比對及計算相似度。我們總共以「CATALOG」、「SUBCAT」、「INTENTION」、「IDXTERMS」及熱門指數權重「WEIGHT_H」欄位。

「CATALOG」、「SUBCAT」這兩個欄位是用來代表某個 FAQ 是屬於那一類業務及子業務，主要用途是系統會將所有業務類別及子業務類別資料先載入列示盒，使用者可透過列示盒(ListBox)瀏覽選單，快速選到某一項業務或子業務之所有 FAQ 資料。「INTENTION」欄位代表這個 FAQ 之意圖類型代碼。「IDXTERMS」欄位是用來作為 FAQ 問句索引鍵詞建檔及相似度計算。「WEIGHT_H」讓系統管理者可隨時調整 FAQ 問句之熱門指數權重。當有某類 FAQ 問題較熱門時，客戶詢問此類 FAQ 比率會較高，因此在系統一執行時，系統便會把熱門指數較高之 FAQ 事先載入，並顯示於挑選視窗供使用者瀏覽挑選，同時也因熱門指數有較高的權重值，系統在排秩回答候選 FAQ 時，這些較熱門的 FAQ 便可排在較前面，以利使用者查詢。

2.4 關鍵詞庫建立

在處理句子時我們首先需要建立一個系統關鍵詞庫對使用者輸入之句子作關鍵詞搜尋及鍵詞容錯搜尋的工作。這詞庫共包含專業鍵詞、意圖詞及一般對話用詞等三大類用詞。同時在關鍵詞庫中必須儲存每個鍵詞的一些屬性(表 5)，以便系統找出關鍵詞及意圖詞及其所含各種屬性後，作更進一步語意分析。

專業鍵詞隨著特定 FAQ 領域之不同，而有明顯不同。意圖詞雖然 FAQ 領域之不同，但意圖分類仍可依 5W1H 分類原則來增加意圖分類，僅需增加部份意圖分類詞即可。專業詞及意圖詞因對句子斷詞、關鍵詞搜尋有較大影響，故比較重要。本系統對關鍵詞之搜尋策略

是關鍵詞庫中不存在之詞則略過不處理，而一般對話用詞只是讓系統在斷詞時確認它是一個詞而已，因此其重要性相對較低。

表 5:關鍵詞庫檔案結構。

欄位	欄位名稱	用途	資料格式範圍	例子
TERM	鍵詞名稱	鍵詞	X(24)	勿干擾
PHONETIC	鍵詞同音詞	同音字容錯	X(24)	驚嚇擾
TYPE	鍵詞種類	鍵詞詞類	X(10)	Nb
CATALOG	鍵詞所屬類別	業務類別	X(02)	01
SUBCAT	鍵詞所屬子類別	子業務類別	X(04)	0112
SlOT_NAME	鍵詞所屬欄位	{Catlog | SubCat |intention |idxterms}	X(10)	
SYNONYM	同義詞	同義詞替換	X(24)	勿干擾
MARKER	鍵詞用途標籤	'F'：冗綴詞(filler) 'M'：意圖(Intention) 'V'：關鍵詞(Key Phrase)	X(1)	'F':filler 'M':Marker 'V':variable
INT_CODE	意圖代碼	鍵詞所屬意圖代碼	X(4)	

關鍵詞庫的建立是利用中研院資訊所之 CKIP Autotag 自動斷詞程式 [16, 17]，將所有 FAQ 問句作一斷詞。斷詞後之每個詞都含詞名稱及詞類，並以下面原則將詞資料填入關鍵詞庫。

原則一：去掉半型、全型個位數之數字(如 0,1,2..9)。

原則二：去掉半型、全型之標點符號(如，、；。：！？「」『』等等)。

原則三：連接詞、介係詞、代名詞、定詞、副詞等較不重要詞類，則設 Marker 欄位值為"F"，代表一般詞(冗贅詞)。

原則四：關鍵詞主要以名詞為主。非關鍵詞之名詞則設為一般詞(冗贅詞)。

原則五：動詞則可能為意圖詞。

原則五：英文字 母全部轉成大寫。

專業鍵詞之萃取，可以用字與字間之結合力關係，將這些萃取出來的專業鍵詞加入關鍵詞庫，同時也可依鍵詞在 FAQ 資料中出現次數及出現位置，來判斷是屬於專業鍵詞或是一般用詞。意圖詞用語則以人工整理得之。

2.5 FAQ 索引檔建立

每筆 FAQ 問句資料中有一個「IDXTERMS」欄位是存放代表 FAQ 之同義鍵詞。利用這些鍵詞來建立索引檔，以加快資料找尋速度，其主要步驟流程如下：

1. 取出 FAQ_REP 檔案中每筆 question 欄位內容。
2. 去掉半型、全型個位數之數字(如 0,1,2..9)。
3. 去掉半型、全型之標點符號(如，、；。：！？「」『』等等)。
4. 英文字母全部轉成大寫
5. 利用已建好之關鍵詞庫來對 FAQ 問句斷詞
6. 找出最長匹配詞及其屬性
7. 去掉關鍵詞庫中 Marker 欄位值為"F"之匹配鍵詞。("F":代表本詞為冗贅詞不需建索引)
8. 利用關鍵詞庫之同義詞欄位內容，作鍵詞替換。
9. 增加鍵詞索引資料到索引檔案內。

2.6 鍵詞權重計算

另外為了計算鍵詞在各個 FAQ 子類別之重要性，我們存放各個鍵詞在各個 FAQ 子類別之出現次數的資料及其比重。詞頻比率 TFR_i 是鍵詞 i 在 FAQ 子業務類別 j 中平滑後之比率值(如公式 1)，其值範圍為０～１之間。若 TFR 值越大，表示此鍵詞在某個 FAQ 子類別內的重要性越大 [4]。

$$TFR_{i,j} = \frac{C_{i,j}}{C_j} \qquad (1)$$

$C_{i,j}$：鍵詞 i 在 FAQ 子業務類別 j 中發生次數

C_j：子業務類別 j 中所有 FAQ 數

除了用 TFR 來表示某鍵詞在各個 FAQ 子

類別之重要性外，我們也計算某鍵詞 i 在所有 FAQ 子類別之分佈情形，以反轉文件頻率 IDF_i 來代表(公式 2)[4]。鍵詞 i 在子業務類別 j 中之權重 $W_{i,j}$ 為 $TFR_{i,j} \times IDF_i$。

$$IDF_i = \log \frac{S_t}{S_i} \qquad (2)$$

S_i：鍵詞 i 分佈在所有 FAQ 子類別中的次數

S_t ：所有 FAQ 子類別數

2.7 鍵詞查詢與鍵詞結合意圖詞查詢

關鍵詞庫中有「CATALOG」及「SUBCAT」兩個欄位。這兩個欄位的主要用途在於我們找到某個關鍵詞時，能立刻獲得這個關鍵詞在各個FAQ業務類別及FAQ子業務類別中出現及分佈情形，並可用來設定 S_Catalog 及 S_Subcat 這兩個系統變數，以便計算 FAQ 相似度。

關鍵詞查詢有兩個缺點：(一)關鍵詞不能清楚表達使用者想要詢問的內容，以致搜尋的結果過多。(二)當使用者所要查詢的資料不存在使用者所用的關鍵詞時，或是使用者無法正確描述出適當的關鍵詞，則系統無法找到所需要的資料。一般使用者日常所使用之句子內容可分成關鍵詞(Key Phrase)、意圖詞(Intention Phrase)及冗綴詞(filler) 等三個主要種類[8]。如下面(例 2-2)中，(如何申請)是一個意圖詞，而[勿干擾]是一個業務關鍵詞，其餘的字都是冗綴詞。我們除了可使用關鍵詞來幫我們找資料外，如能透過意圖詞之輔助，則能更迅速幫助使用者找到他想要的資料。

(例 2-2)：我想查一下(如何申請)[勿干擾]？

(如何申請)	：	意圖詞
[勿干擾]	：	關鍵詞
我想查一下	：	冗綴詞

2.8 關鍵詞搜尋

本系統所採用的關鍵詞搜尋流程是詞庫式最長詞匹配(longest matching)找尋法，其步驟如下

1. 查詢句子 Q=($C_1C_2C_3C_4....C_n$)
2. P 是一個指標，指到查詢句子 Q 中的第一個字 C_1。
3. 檢查關鍵詞庫中是否有以 C_1 開頭的詞彙，若沒有則指標 P 移至 C_2，重複步驟 3，若有則執行下面步驟。
4. 將所有以 C_1 開頭的詞彙，放入 Terms 陣列內。
5. Terms={ <C_1>、<C_1C_2>、<$C_1C_2C_3$>、<C_1C_4>、<C_1C_5>、}
6. 將 Terms 陣列依詞彙長度降冪排序。
7. Terms={ <$C_1C_2C_3$>、<C_1C_2>、<C_1C_4>、<C_1C_5>、<C_1>}
8. 將 Terms 中之詞依序一一取出，與指標 P 所指到之字串互相比對，是否完全匹配？
9. 若 Terms 中之詞沒有一個與指標 P 所指到之字串完全匹配，則指標 P 移至 C_2，重複步驟 3。否則執行下面步驟。
10. 鍵詞<$C_1C_2C_3$>與 P 所指到之字串完全匹配，斷出 $C_1C_2C_3$。
11. 指標 P 移至 C_4，重複步驟 3，直到 P 指標所指到之字串全部處理完。

因為使用者輸入之句子可能中英文混合，所以在關鍵詞搜尋上有中文鍵詞、英文鍵詞及英文中文混合鍵詞等三種找尋狀況，關鍵詞找尋規則分成中文、英文關鍵詞採最長詞匹配，英文中文混合關鍵詞則採最長詞匹配，當無法找到時，則放寬為部份詞匹配。

2.9 查詢句處理

我們對使用者輸入之查詢句子之處理流程共有下面四個步驟，各個步驟說明如下：

步驟 1. 關鍵詞搜尋

關鍵詞搜尋利用本系統所建立之關鍵詞庫，以

最長詞匹配原則，將使用者查詢句子斷成詞串，並標註每個詞的詞類屬性及其所屬 slot。如果關鍵詞庫內有同義詞，則將此鍵詞以同義詞替換。

步驟 2. 容錯查詢

容錯處理主要是當使用者查詢句無法與關鍵詞庫中鍵詞完全匹配時，則執行同音字容錯檢查。先將使用者查詢句透過同音字轉換程序，轉成為同音代表碼的同音句，再與關鍵詞庫中鍵詞同音詞這個欄位比對，並計算同音句與鍵詞同音詞的編輯距離。當編輯距離低於一門檻值時，則視之為鍵詞。同音字轉換程序利用同音字庫(由常用的 5401 個字所組成)將所有相同發音之中文字全部轉成一致的同音代表碼。

步驟 3. SQL 轉換及執行。

依照各鍵詞之屬性，產生 Frame 結構，並將各鍵詞資料填入對應 slot。依照框架結構中之各個 slot 值，轉成適當之 SQL 指令，找出所有可能之 FAQ。

步驟 4. 相似度計算。

計算查詢句子與 FAQ 問句之相似度並依相似度計算結果，降冪排序，顯示 FAQ 內容資料。

2.10 詢問句與 FAQ 問句之相似度計算

當使用者一次輸入或陸續輸入查詢句子後，則透過適當查表，將查詢詞彙指派到使用者查詢 Frame 之適當 Slot 內，再計算使用者查詢 Frame 與 FAQ 問句之相似度 S（公式 3），並由大至小降冪排序，將結果顯示給使用者。考量當使用者在問某一類業務問題時，常常會希望同時得到此一類業務的相關問題，所以加上類別及子類別這兩個歸屬值之權重，讓使用者得到同一類別之相關 FAQ。

$$S=MIN((C_1\times S_1+C_2\times S_2+C_3\times S_3+C_4\times S_4)+X,1) \quad (3)$$

其中

S_1：類別之歸屬值。

S_2：子類別之歸屬值。

S_1 及 S_2 由鍵詞所屬類別及子類別得到，比對詢問句之鍵詞所屬類別及子類別是否與 FAQ 問句相同。若類別歸屬相同則 $S_1=1$ ，否則 $S_1=0$。若子類別歸屬相同則 $S_2=1$ ，否則 $S_2=0$。

S_3：詢問句鍵詞 q 與 FAQ 問句 f 兩者間向量相似度(公式 4)。

$$\cos(\vec{q},\vec{f})=\frac{\sum_{i=1}^{n} q_i f_i}{\sqrt{\sum_{i=1}^{n} q_i^{\,2}}\sqrt{\sum_{i=1}^{n} f_i^{\,2}}} \quad (4)$$

S4：鍵詞數之包含率(公式 5)。

$$S4=\frac{C_m}{C_t} \quad (5)$$

C_m：使用者鍵詞出現於 FAQ 問句內的數目

C_t：FAQ 問句所有鍵詞數。

X: FAQ 熱門指數附加權重(值域為 0~0.05)，X=(H-0.5) x 0.1。

H：FAQ 熱門指數，預設 0.5。

C_1：類別之歸屬值之權重，預設為 0.1。

C_2：子類別之歸屬值之權重，預設為 0.1。

C_3：鍵詞之向量相似度權重，預設為 0.5。

C_4：鍵詞數之包含率之權重，預設為 0.3。

$C_1+C_2+C_3+C_4=1$。

2.11 容錯查詢

本系統具有容錯查詢能力，提供一個強健查詢方式，對使用者誤字、多字及漏字等現象皆能處理。本系統容錯是使用計算編輯距離之大小，找出各關鍵詞與使用者問句間編輯距離最小，且編輯距離低於門檻值之關鍵詞。實際使用經驗顯示中文字串之多字現象比較容易偵測出來，而漏字現象可能會造成句子斷詞錯誤，無法正確找出資料。

關鍵詞容錯功能允許查詢句子內鍵詞中有部份錯誤存在。在鍵詞與查詢句子無法完全匹配時，則再執行容錯鍵詞查詢。若未斷詞之第一個字是英文字，則其英文鍵詞尋找過程是

找出詞庫中所有以該英文字母開頭之鍵詞，其編輯距離與詞庫鍵詞長度比值為最小，且低於預設之門檻值，則將之視為鍵詞。

本系統也提供同音容錯處理功能。其關鍵技巧在於將使用者查詢句與關鍵詞都使用同音字轉換程序，轉成一致的同音代表碼後，再計算兩者間之編輯距離。另外容錯鍵詞找尋有兩個問題存在如下:

問題一：查詢句子 Q 中要取幾個位元組(Bytes)，來與詞庫之鍵詞 T_i 比較呢？

問題二：找到可能之鍵詞 T_i 後，查詢句子指標 P 要向前幾個位元組，繼續剖析未完之 Q？

系統利用一些簡單判斷規則，來解決上面兩個問題。判斷規則過程如下:

規則一：找出關鍵詞 T_i 之最後一個中文字 C_n，是否在待剖析句子 Q 中出現(Window Size 為 T_i 長度之 1.5 倍)？若 C_n 有出現，則取出 Q 自開頭位置至 C_n 之部份字串 S，計算 T_i 與 S 之編輯距離 EDIT。否則執行規則二。

規則二：找出 T_i 之最後第二個中文字 C_{n-1}，是否在 Q 中出現？若 C_{n-1} 有出現，則取出 Q 自開頭位置至 C_{n-1} 之部份字串 S，計算 T_i 與 S 之 EDIT。否則執行規則三。

規則三：找出 T_i 之最後第三個中文字 C_{n-2}，是否在待剖析句子 Q 中出現？若 C_{n-2} 有出現，則取 Q 自開頭位置至 C_{n-2} 之部份字串 S，計算 T_i 與 S 之 EDIT。否則 T_i 非關鍵詞，比較下一筆關鍵詞 T_{i+1}。

規則四：若 T_i 與 S 之 EDIT 低於門檻值，則視之為鍵詞，指標 P 向前移動 S 的長度，繼續剖析 Q。否則非關鍵詞。

2.12 系統狀態圖

本系統的對談流程控制是使用系統狀態圖來控制。系統設有一個 State 變數，內含 4 個位元(Bit)，每個位元代表一個 slot。以這個位元之開啟或關閉，來代表此 slot 是否已經有資料(如圖 2)。系統再依目前 state 狀態作適當回應。例如 state="0001"代表使用者已經有輸入鍵詞，但類別、子類別及意圖詞這三個 slot 尚未有資料。另外我們對每個鍵詞已事先計算

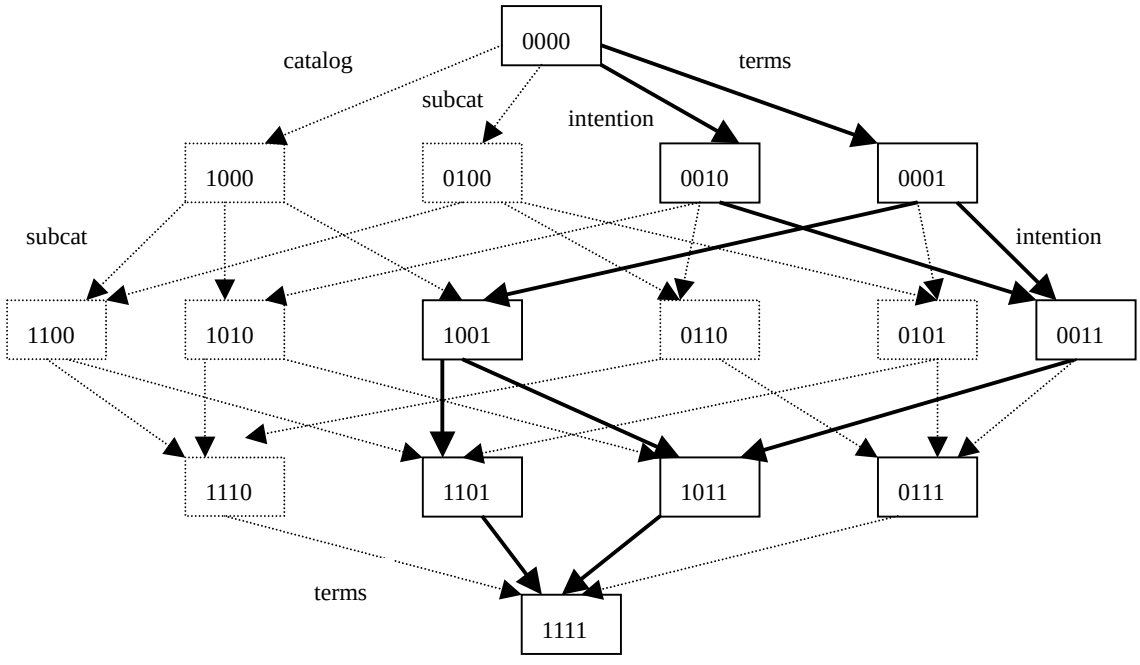

圖 2:系統狀態

是否只出現在某些特定類別 FAQ 中，所以可由鍵詞得到所屬類別、子類別的資料。例如 state="0001"表示使用者輸入之鍵詞並無法讓系統來判斷鍵詞的類別、子類別，也就無法設定系統的類別、子類別的 slot，所以系統要回應使用者，請使用者補充輸入查詢資料。因為系統的事先計算鍵詞所屬類別，所以圖中有些 state 狀態(虛線表示)將不會發生。目前每個 FAQ 只以四個 slot 來代表，而一般對談系統之資料庫多屬關連式資料庫，會有較多的 slot 欄位，因此對談管理之功效在本系統中不易顯現。

2.13 語境資料

表 6: 語境資料結構。

欄位	欄位名稱	欄位說明	欄位型態
1	time_stamp	時間郵戳	X(8)
2	old_stamp	舊時間郵戳，用於回溯使用者查詢過程(back tracking)	X(8)
3	state	系統狀態 “0000”～”1111”	X(4)
4	catalog	業務類別代碼	X(2)
5	subcat	子業務類別代碼	X(4)
6	intention	意圖代號	X(4)
7	terms	鍵詞	X(60)
8	user_query	使用者鍵入之查詢句子	X(100)
9	source	資料來源 'K' 使用者使用鍵盤鍵入 'M' : 使用者使用滑鼠點選 'S' : 系統產生	X(1)
10	FAQId	FAQ 檔案代號	X(5)
11	RecCount	找到資料筆數	9(5)

語境資料主要用途是協助對談管理流程控制，語境資料 (如表 6)。記錄使用者之查詢過程。每筆語境資料都有 Time Stamp，以便回溯查詢過程。系統將使用者的查詢過程及目前 state 狀態及系統變數等資料產生一筆對應的查詢過程資料。對談管理利用此一過程檔案與查詢句子比對，可判斷使用者是否切換查詢議題，也可來協助系統將使用者陸續輸入之查詢句子轉成 frame 之 slot 欄位。例如使用者先輸入鍵詞“申請”來查詢資料，接著再輸入“勿干擾”，則系統由語境資料之輔助，會判斷出使用者之意圖為” 申請”、”勿干擾”，所以會找出“如何申請勿干擾?”之 FAQ，同時系統也會找出與“勿干擾”相關之 FAQ，但不會找出含有鍵詞“申請”的 FAQ。

3. 系統功能與比較

本系統主要功能為 FAQ 查詢，有選單方式及口語化兩種查詢。另外可選擇設定系統動畫人物及聲音種類。系統提供 FAQ **與**詞庫的資料維護作業，如新增、修改、刪除、查詢等功能。FAQ 索引維護作業則用來建立 FAQ 索引檔，以便系統加快資料搜詢的速度。網頁 FAQ 擷取作業，則用來自網站擷取原始 FAQ 資料後，將這些網頁 FAQ 資料經格式剖析後轉入本系統 FAQ 資料庫內。FAQ 查詢分析則用來統計、分析各類 FAQ 查詢流量及使用者的查詢過程。本系統所用人工合成聲音檔由貝爾實驗室的「中文語音合成系統」預錄產生。

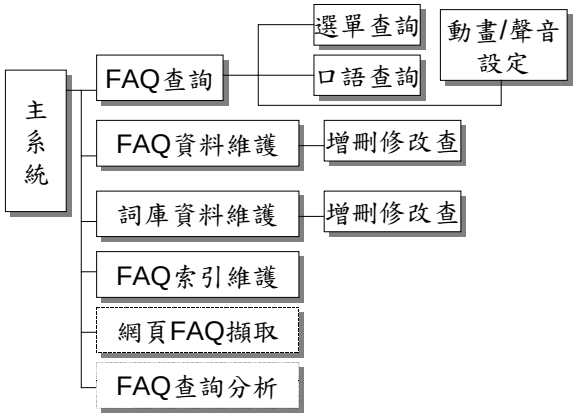

圖 3: 系統功能。

本系統的設計都以使用者的查詢方便性來考量，具有聲音及動畫回應功能，提供使用者一個友善的人機介面，便利的操作使用方法。本系統具有「選單瀏覽查詢」及「口語化

查詢兩」大類查詢方法。「選單瀏覽查詢」共有「類別」、「子類別」、「意圖種類」、「關鍵詞」及「熱門 FAQ」等五種瀏覽查詢種類，使用者可分別依「業務類別」、「子類別」及「意圖種類」等三種分類方式來查詢某類別的 FAQ 資料。也可使用「關鍵詞」瀏覽檢視功能來作鍵詞查詢檢索，另有「熱門 FAQ」供使用者快速瀏覽較熱門之 FAQ 問題。

「選單瀏覽查詢」的主要目的，列出各種查詢類別選項，方便使用者透過滑數自行挑選檢視資料。避免對本系統較陌生之使用者一進入系統時，茫茫然不知所措之現象。分類查詢的方式讓使用者可依分類項目來查出該查詢類別之所有 FAQ，使用者可就這個類別中之所有 FAQ 問句，再更進一步檢視其 FAQ 解答，另外系統提供鍵詞檢索功能，將所有鍵詞依照字母順序排序，透過滑鼠，供使用者瀏覽挑選含有某鍵詞之所有 FAQ 資料。

「口語化查詢」讓對系統較熟悉之使用者，利用鍵盤直接輸入口語化之查詢句子，快速得到相關資料。系統同時具有容錯功能，對於常見的多字、漏字及誤字等錯誤情形都能檢出，另外對於中文輸入中最常見的同音詞錯誤一樣可處理，允許使用者在誤、漏字輸入時，仍然能正確剖析查詢句子，找出正確而相關之資料。

圖 4:系統啟動時顯示較高熱門指數之 FAQ。

如圖 4 當系統啟動時，會先顯示較高熱門指數之 FAQ。

在一般查詢時系統可共有「類別」、「子類依使用者的查詢過程，找出使用者想要之資料。例如使用者先輸入"申請"這個鍵詞，共查到 88 筆含有"申請"這個詞的 FAQ 資料，接著使用者再輸入"勿干擾"，系統只會查到 6 筆"勿干擾"之 FAQ 資料，同時會把"如何申請勿干擾"顯示在第一筆的位置，並把這個 FAQ 的解答資料顯示於最下面視窗內(如圖 5)。

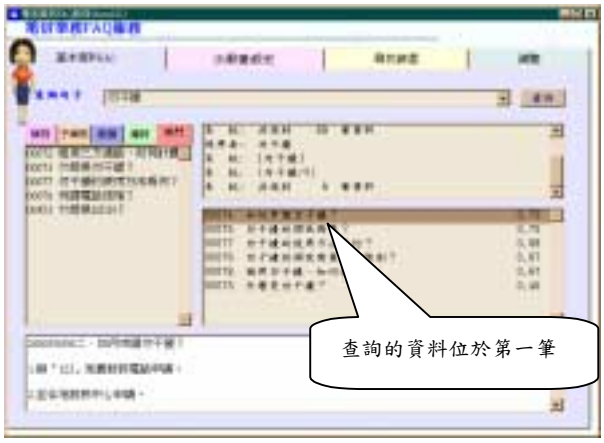

圖 5: 使用者再輸入"勿干擾"的查詢畫面。

若使用者直接輸入"勿干擾如何申請"，系統只會把"勿干擾"這項業務的所有 FAQ 資料顯示，同時會把"如何申請勿干擾"顯示在第一筆的位置 (圖 6)。

圖 6:使用者輸入"勿干擾如何申請"之畫面。

系統也可以直接輸入單一鍵(”勿干擾”)來查詢資料。

表 7 為本 FAQ 查詢系統與其他類似系統之功能比較一覽，顯示本系統是一個功能完備的 FAQ 查詢系統。

表 7:本系統與其他系統比較。

項目	本系統	寶來 E 博士	Ask Jeeves
語言	中文	中文	英文
關鍵詞查詢	○	○	○
口語化查詢	○	○	○
熱門 FAQ 瀏覽查詢	○	○	X
類別瀏覽查詢	○	X	○
關鍵詞瀏覽查詢	○	X	○
意圖瀏覽查詢	○	X	X
聲音、動畫	○	X	X
相似度排秩	○	○	○
查詢結果	列式盒(ListBox)顯示,可上下捲動	條列式(List)顯示	分頁(Page)顯示
同音字容錯查詢	○	X	X
條件查詢	X	X	○

4. 結論與未來研究

本論文提出並實作了一個電信服務業務之 FAQ 查詢系統，系統具有選單查詢及口語化查詢兩大類查詢方法。使用者可分別依業務類別、子類別及意圖種類等三種分類方式及關鍵詞檢索功能來作選單瀏覽查詢，另有熱門 FAQ 供使用者快速瀏覽。口語化查詢的功能讓使用者以自然的口語化句子向系統查詢資料，配合聲音、動畫的活潑性反應，達到友善的人機介面。另外本系統提供的容錯功能，讓系統在鍵詞多字、漏字及誤字等錯誤情形下時仍能正確斷詞找出資料。同時經由整個查詢過程的記錄系統可作查詢流量統計。後續研究將加強於快速對談管理模組的發展及口語化剖析技術的研究。

參考文獻

[1] Robin Burke, Kristian Hammond, Vladimir Kulyukin “Natural Language Processing in the FAQ Finder System: Result and Prospects” 1997 AAAI Spring Symposium Technical Report SS-97-02。

[2] K. Hammond, R. Burke, C. Martin, S. Lytinen, “ FAQ finder: a case-based approach to knowledge navigation” Proceedings of 11th Conference on Artificial Intelligence for Applications, 1995, pp. 80 - 86。

[3] Wolfgang Minker ;Alex Waibel and joseph Mariani ”stochastically-based semantic analysis” Kluwer academic publishers 1999。

[4] G. Salton, Automatic Text Processing. Addison-Wesley。

[5] S. Seneff and J. Polifroni, "Dialogue Management in the Mercury Flight Reservation System," presented at Satellite Dialogue Workshop, ANLP-NAACL, Seattle, April 2000。

[6] C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi and V. Zue, "Muxing: A Telephone Access Mandarin Conversational System," Proc. 6th International Conference on Spoken Language Processing, Beijing, China October 2000。

[7] Ruifeng Xu and Daniel Yeung “Experiments on the Use of Corpus-based Word BI-gram in Chinese Word Segmentation “ Systems, Man, and Cybernetics, IEEE International Conference, vol.5, 1998, pp. 4222 - 4227。

[8] Yen-Ju Yang; Lee-Feng Chien; Lin-Shan Lee “Speaker intention modeling for large vocabulary Mandarin spoken dialogues”, Proceedings of Fourth International Conference on Spoken Language, Vol. 2, 1996 , pp. 713 -716。

[9] V. Zue and J. Glass, "Conversational Interfaces: Advances and Challenges" Proceedings of the IEEE, Special Issue on Spoken Language, Vol. 88, August 2000。

[10] 林一中，江東輝，”中文口語交談系統架構初探”，電腦與通訊，第 66 期，pp.3-12，1998 年。

[11] 彭崇銘，王慧明，林一中，江東輝，張照煌，”基於分散式主從架構之中文口語交談系統的設計”，電腦與通訊，第 76 期，pp.15-23，1999 年。

[12] 黃韻璆，曾怜玉”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碩士論文 民國 88 年 6 月。

[13] 廖康任，曾怜玉”自然語言應用於具可移植之電話總機之研究” 國立中興大學應用數學碩士論文 民國 88 年 6 月。

[14] 李憲育，曾怜玉”模擬電話總機之自然語言處理” 國立中興大學應用數學碩士論文 民國 87 年 6 月。

[15] 李坤霖，吳宗憲”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碩士論文 民國 89 年 6 月。

[16] http://godel.iis.sinica.edu.tw/CKIP/ws/ 中央研究院詞庫小組 CKIP 中文自動斷詞系統。

[17] 中文詞知識庫小組 “中文詞類分析” 中央研究院資訊科學研究所 1993 年 6 月。

[18] http://drdai.polaris.com.tw/default.asp。

[19] http://www.ask.com 。