應用問答技術於電腦領域論壇檢索之研究

Applying Question and Answering Technique to Computer Hardware Forum Practice

黃純敏
國立雲林科技大學資訊管理系
Department of Information Management
National Yunlin University of Science and Technology
huangcm@yuntech.edu.tw

江志銘
國立雲林科技大學資訊管理系
Department of Information Management
National Yunlin University of Science and Technology
g9223719@yuntech.edu.tw

呂盛興
國立雲林科技大學資訊管理系
Department of Information Management
National Yunlin University of Science and Technology
g9523707@yuntech.edu.tw

摘要

線上論壇(BBS)已成為人與人之間資訊交流與分享的重要管道，如何搜尋出較符合使用者預期的成果是目前檢索系統探討的重要議題之一。本研究針對電腦領域專有名詞設計問答檢索系統，本系統包括：詞性合併、答案類型偵測、候選答案評分機制三個子系統。本機制可依據使用者所提出的問題，快速回覆可能的候選答案組合，最後透過評分的機制挑選出最佳的答案。實驗結果發現，整體系統的精確率為40.48%，召回率為 42.93%，與採用相似度查詢方式比較，精確率與召回率分別提升9-22%以及 19-34%。此外，在專有名詞的辨識率則達七成七。

關鍵詞：問答技術，詞性合併，答案類型偵測，候選答案評分

Abstract

The On-Line Forum becomes an important channel of communication for general publics. To search a variety of articles from the forum, the most straightforward way is to use keyword matching. Traditional search engine accepts keyword input, and then replies a list of ranked results. From those ranked results, users still need to examine each article to find out the answer. The computer forum differs from traditional Q&A news reports in writing, many specific terms like ASUS P5GD1 Pro or ASUS EN6600GT/TD/128M are frequently used in the articles. In order to recognize these terms correctly, we employ a merging method to join each possible word together with a phase using the CKIP's tags. In this study, we propose a Q&A system that users can ask a question about computer hardware, and have possible answers in return. The system was evaluated by using a QA set about computer hardware. Compared to TFIDF, the performance of precision and recall increases 9-22%, 19-34% respectively. Besides, the average accuracy of the POS merging subsystem reaches 77%.

Keyword: Question Answering, Part of Speech Merging, Answer Type Detection, Candidate Answers Scoring

一、前言

線上論壇中專業領域內的問題解答一直是使用者頻繁搜尋的目標，舉凡：硬體技術文件、使用手冊、安裝方法…等。這些文件集所蘊含的專業術語與專有名詞的比例相當高，例如：「Kinston DDR 533 RAM」、「GA-8SR 533」、「Geforce 4 MX 440」等，此類詞彙通常在文件集中具有一定的識別特徵(Features)，若能辨識這類型的專有名詞將有助於提升問答系統內資訊檢索模組的效能。

本研究以 Google 線上論壇[9]內的問答集做為研究對象，經由觀察發現，我們發現在電腦硬體論壇內的文章具有下列幾點特性：

1. 電腦專有名詞的詞彙十分普遍，如：Intel P4 3.0CG、P4P800SE。
2. 文章重點都環繞在專有名詞之前後文資訊。
3. 討論區內的文章篇幅不長，文章的品質無法根據詞彙數來決定。
4. 討論區的文章用語與寫作風格迥異、不同文章討論的主題不同。

本研究提出一電腦專有名詞問答檢索系統(Computer Domain Proper-Nouns Answer Finding System)其所要實現的系統結合資訊檢索與資訊擷取的技術，希望能幫助使用者更快速、有效率的從非結構化的文件中搜尋符合使用者需求的答案。

首先，使用者可以輸入問句「那塊主機板可以支援 Intel P4 CPU 處理器？」，藉由問句剖析(Parsing)將問句拆解成關鍵詞組(Keywords)和問句意圖(Intending)。即：<意圖詞：那些主機板>、<關鍵詞組：支援、Intel、P4、CPU、處理器>。接著，將剖析的問句意圖進行答案類型偵測找出所對應的答案類型，樣版：<那塊主機板→主機板類型>。將答案類型(Answer Type)與關鍵詞組，進行資訊檢索模組搜尋，找出所有相關的文件集合，並對所有候選答案評分與排序(Ranking)取出前五名回傳給使用者。回傳的答案會依分數高低傳回一文章段落、句子或特定的專有名詞。例如：答案：「文章：865PE 支援 p4 CPU ，而且最快速的速率可達…」、「句子：p4 CPU 支援 P4P800-X 主機板」、「特定專有名詞：8IPE1000G 主機板」。

二、文獻探討

問答系統有別於一般的資料搜尋，一般的搜尋只是針對使用者輸入的關鍵字詞加以搜尋，從大量資料中比對出內含該關鍵字的資訊呈現給使用者，但是使用者還是必須自行過濾判斷其中真正有用的資訊。問答系統可以根據使用者的問題找出確切的答案[4]，也就是說問答系統能夠根據使用者的意圖來正確的回答使用者想要的答案。

目前問答系統的研究可分為三大類：基於常問問題集的問答系統，基於百科知識的問答系統，開放領域的問答系統[5]。中文問答系統的實作也由其困難之處，而主要的困難點也都源自於中文字詞處理的原罪〔例如：斷詞的困難、詞性的判斷等〕，另外中英文夾雜的情形也使得中文檢索技術更為困難[10]。基於常問問題集以及百科知識所架構而成的問答系統大多的目的都是回答一般性的問題，但是要將回答一般性問題的問答系統應用在特定領域，例如本研究中的電腦硬體領域，也很難見其成效，因為要回答特定領域的問題所使用的資料應該與該領域高度相關，如此才能有較高的回答正確率。

在前言中本研究歸納了幾項關於電腦硬體論壇中的文章特色，而受限於現有的搜尋引擎或是資訊檢索技術的能力，使用者在查詢有關電腦硬體的資料時，往往無法根據自己本身的問題來得到精確的解答。在問答系統的研究中，使用網際網路的資源來輔助問答系統的效能已經被證實是有效的方式[6]，因此本研究所提出

的電腦專有名詞問答檢索系統將透過納入此方法，彌足現有檢索技術對電腦硬體資料查詢的不足之處，希望能幫助使用者更快速、有效率的從非結構化的文件中搜尋符合使用者需求的答案。

三、研究架構

本研究提出一套電腦專有名詞問答檢索系統，其所要實現的系統結合資訊檢索與資訊擷取的技術，輔以答案類型偵測的方式找出所需要的答案。主要的目的是希望幫助使用者更快速、有效率的找出電腦硬體類型的專有名詞，系統可分為二大部分：(1)前置處理：討論區的文章進行剖析、斷詞與詞性標記、詞性合併與答案類型偵測；(2)後端處理：將剖析的問句資訊進行資訊檢索，找出符合問句意圖的候選段落或句子，最後，對文章進行評分與排序。

圖 1 為本研究系統架構，共分成四個部分：分別為「詞性合併」、「答案類型偵測」、「資訊檢索」、「候選答案評分」。以下針對各部分研究步驟進行說明。

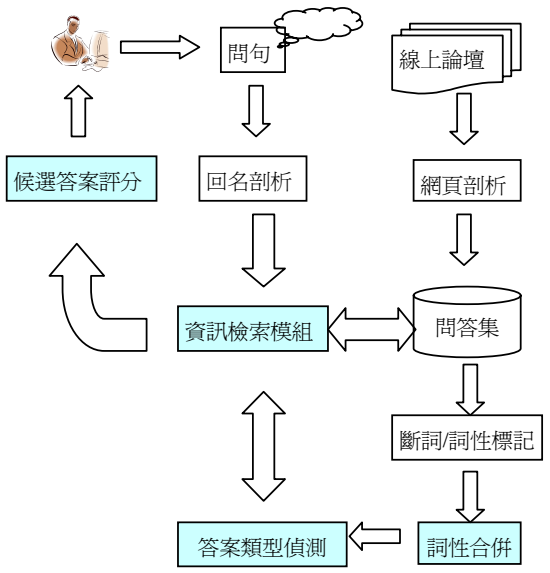

圖 1、系統架構圖

1. 首先由使用者提出查詢問句，透過問句剖析器取得問句中的關鍵詞與疑問詞(Question Word)。前者將進行資訊檢索模組，後者則進行問句/答案類型配對模組(Question/Answer Type Matching)。
2. 資訊檢索模組會找出同時出現關鍵詞組的文件集。
3. 另一方面，透過問句/答案類型配對模組可以取得疑問詞與其對應之答案類型(Answer Type)。
4. 完成 2,3 步驟後，系統將從回傳的文件集中偵測可能的候選答案，隨後進行候選答案評分機制(Candidate Answers Scoring)，最後將其結果排序取出前五名(Top 5 Rank)答案回傳給使用者。

（一）斷句與斷詞處理

由於線上論壇內的文章寫作風格不盡相同，為避免造成段落句子不易分辨的情況發生，因此，在進行斷詞之前，必須先將論壇內的文章給予適當的處理，在此，我們藉由標點符號用法[8]，將文章重新斷句，並排除其他過於頻繁的標點符號以及停用詞。

根據中研院詞性標記表[1]將詞性分為四十七種詞類，其中以名詞及動詞所代表的訊息最具意義且比例最高，另外，由於電腦硬體論壇內的單位詞相當頻繁，因此，英文及數量詞亦在我們分析的範圍內。至於其他的詞類如語助詞與感嘆詞，如：哦、啊…等，在文件中大致上都只負責修飾、連接、表達語氣或態度的功能，我們將予以排除。值得注意的是，副詞(D)中包含疑問性副詞，如：為什麼、何時、如何..等，對文件內容有決定性的影響，故不可輕易忽略之。

在斷詞方面，本系統採用中研院詞庫小組開發的中文斷詞與詞性標註系統[2]，CKIP 系統在未知詞的偵測上具有相當的成效，諸如：技嘉、金士頓、華碩、微星…等專有名詞皆能有效的斷出詞彙，但對於電腦專有名詞的辨識上效果不

佳，以 P4P800SE 為例：CKIP 系統會將其斷成 P 4 P(FW) 8 0 0 (Neu) SE(FW)，主要的原因在於 CKIP 系統對於英文字皆給予外文詞性標記(FW)，對數字則給予數量詞性標記(Neu)，一旦遇上英文與數字同時出現時，這樣的斷詞結果將影響後續在資訊檢索模組中計算詞彙權重的誤差(表 1)。

表 1、詞性分類(本研究整理)

分類	詞性	個數
名詞	Na Nb Nc Ncd Nd	5
動詞	VA VAC VB VC VCL VD VE VF VG VH V VL HC VI VJ VK	16
量詞	Neu Nes Nep Neqa Neqb Nf Ng	7
副詞	Da Dfa Dfb Di Dk D	6
停用詞	DE SHI V_2 A Nh I P T Caa Cab Cba Cbb	12
英文標記	FW	1

（二）詞性合併

為了加強電腦詞彙的鑑別力，本研究歸納出相關的詞性規則，針對硬體領域的專有名詞進行詞性合併，以詞性構成要素為原則，輔以經驗法則得知在某些情況下某些詞性共同出現的頻率相當高，並以「長詞優先合併」以及「單位詞優先合併」二大原則，歸納出下列近 18 條合併規則(表 3)，並自定四個詞性(表 2)做為合併完成後的詞性，詳細說明如下：

1. 長詞優先

在專有名詞中通常其資訊具有不可分割的特性，例如：「中央研究院」與「中央」、「研究院」其資訊量較高且較具代表性。

2. 單位詞優先

為了從合併的詞彙組合中正確判斷量詞及單位詞的型態，我們蒐集國語教育委員會「國語辭典第三版之量詞表」[7]，這也是為了彌補 CKIP 系統不足的地方，CKIP 系統針對數量詞或是單位詞並沒有特別的處理，而本研究將數量與單位的量特地擷取出來對於之後的詞彙解析有正面的效果，詳述於系統實作與評估一節。找出與價格單位有關的詞彙，如：元、個、顆、條、塊…等，並作為量詞類型判斷的依據。常見電腦量詞如：數量詞：一個、二顆、三條…、價格：$1,400, 三千元、六百八十塊…、單位詞：32 位元、65 度、12 公分…。

表 2、自定詞性類型

Proper Noun(PN)： 專有名詞	Number Unit(NU)： 數量單位
Question Word(QW)： 疑問詞	Modify(MF): 修飾詞

表 3、詞性合併結果(部分)

CKIP 詞類標記			構成詞彙	詞性
(FW)	(Neu)	(FW)	Sp2600+,V9999GT	(PN)
(FW)	(FW)	(Neu)	Pioneer-A09,BenQ-FP791	(PN)
(Nes)	(Neu)	(Nf)	每一個、另一種	(QW)
(Nep)	(Neu)	(Nf)	這兩塊、哪一張	(QW)
(Neu)	(FW)		264 MB、333 MHz	(PN)
(FW)	(Neu)		AMD 64、NVIDIA 5200	(PN)
(FW)	(FW)		INTEL-M-P4CPU,K8N4-E-Deluxe	(PN)
(Na)	(Na)		記憶體插槽、液晶顯示器	(PN)
(Dfa)	(VH)		非常高、最快	(MF)
(Da)	(Neu)		約 100、最多 4096、、總計三萬	(NU)
(Neu)	(Nf)		一千多塊、100 元、五萬元	(NU)
(Nep)	(Nf)		這顆、這張、那個、那塊	(NU)
(Neu)	(Neqa)		一部份、兩部份	(NU)

（三）答案類型偵測

答案類型偵測的目的在判斷合併後的詞彙其所屬的類別，例如：P4P800SE 屬於「主機板類」、DDR 400 512 MB 屬於「記憶體類」等，這些詞彙在電腦領域出現的次數非當頻繁，為了正確的辨識出各詞彙所屬的類型，我們提出一答案類型偵測技術，即透過線上搜尋引擎(Yahoo Kimo Search)擷取網頁中重要的關鍵詞彙。我們觀察目前檢索技術，如：page Ranking，發現詞彙間共同出現的機率可作為答案類型判斷的依據。偵測流程如圖 2 所示。

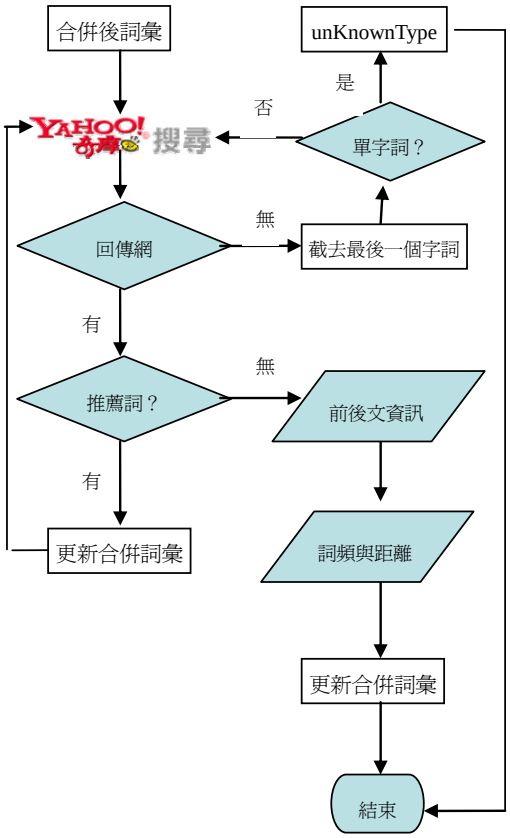

圖 2、答案類型偵測模組流程圖

處理流程解說如下：

1. 我們將合併後的詞彙進行搜尋引擎的檢索，並剖析回傳網頁的前十筆文件。
2. 以合併的詞彙為中心點(focus)，找出中

心點前後各取五個出現的詞彙集合，並且優先擷取出名詞(Na)的詞彙組合，若擷取出的詞組中出現我們定義的類別相關詞組時，即予以計算權重。

3. 查詢的詞彙若無回傳網頁代表合併的詞彙不具代表性或合併過多資訊，因此，我們截去詞彙的最後一個單字詞，並重新檢索該詞彙。直到截取至最後一個字詞為止，若最後一個字詞仍無回傳網頁則系統會給予未知類別(unKnownType)。此外，若回傳網頁內包含 Yahoo Kimo 提供的推薦詞彙，則以推薦詞優先並更新合併詞彙，重新進行檢索。
4. 從檢索後的網頁文件中，找出詞彙的前後文資訊，接著，開始計算各查詢詞彙與類別詞彙間共同出現的頻率(Co-Occurrence)以及彼此詞彙間的距離(Distance)，最後附以權重，取出權重值最高者視為該關鍵詞之類別關鍵詞(Category)。
5. 電腦硬體的分類本研究蒐集名人三 C 電腦賣場所提供的電腦硬體報價單[3]，將報價單上的硬體分類作為研究中的分類關鍵詞(表 4)。

表 4、電腦硬體分類與相關特徵詞彙

類別詞及相關詞組		
<主機板&Motherboard &Mainboard>	<硬碟&HD&Hard Disk>	<螢幕&Monitor&LCD>
<記憶體&RAM& Memory>	<音效卡&Audio&Sound>	<掃瞄器&Scanner>
<顯示卡&VGA&AGP>	<燒錄機&DVD RW&CD RW>	<印表機&Printer>
<處理器&CPU>	<光碟機&CDROM>	

在此我們以關鍵字:”P4P800SE”為例：

1. 以詞彙”P4P800”為中心點(focus)，找出中心點前後各取五個出現的詞彙集合(主機板、華碩、規格說明、SOCKET478、支援…)，並且優先擷取出名詞(Na)的詞彙組合：主機板(Na)、華碩(Na)、規格說明(Na)。分別計算詞彙與類別相關詞組的權重。
2. 在此我們假設兩詞彙出現的次數愈多、距離愈近，表示兩者關係愈緊密。$term_i$ 為合併後的詞組，$term_j$ 為類別詞組。Category 為我們所定義的類別關鍵詞，以 C_i 表示 $term_i$ 的所屬類別，$|C|$ 代表類別詞組的個數。f_{ij} 代表 $term_i$ 與 $term_j$ 之間共同出現的次數，共同出現的次數愈多其權重愈高；d_{ij} 代表 $term_i$ 與 $term_j$ 間的距離位置，兩詞彙的距離愈近其權重愈高。如公式(1)所示：

$$C_i = \arg\max_{j=1}^{|C|} f_{ij} \sum (2^{d_{ij}})^{-1} \qquad (1)$$

3. 如表 5 所示，查詢詞彙(P4P800)與類別詞組(主機板)在共同出現 3 次，各別距離(不包含標點符號)為 2,0,4。另一類別詞組為(顯示卡)，共同出現 2 次，各別距離為 4,4。其權重分別為：

$$C_{主機板} = 4*(1/2^2 + 1/2^0 + 1/2^4) = 3.9375$$

$$C_{顯示卡} = 2*(1/2^4 + 1/2^4) = 0.25$$

4. 計算出各類別詞組與查詢詞彙的權重後，取其權重最高即查詢詞彙(P4P800)予以歸類到「主機板類別」。

表 5、Yahoo Search 查詢"P4P800"網頁結果

標題	檢索結果
產品詳細資訊	>> 主機板 \ 華碩 - P4P800 規格說明: Socket478, ... 我的主機板 P4P800 有支援 on board LAN 但是我想用我的 BOOT ROM 開機 所以插入另...
Download	支援與服務 技術資料 主機板支援 ... P4P800 Deluxe 及 P4P800 BIOS 1018 更新 BIOS 之前務必詳閱詳細 ... P4P800 Deluxe 及 P4P800 BIOS 1017 更新 BIOS 之前務必詳閱 ...

（四）候選答案評分

一旦偵測出電腦專有名詞及其所屬類別後，透過資訊檢索找出相似度最高的文件，再從這些文件中尋找定義的樣版特徵，這樣作法可以召回較廣泛範圍的文件集合，並進行縮小範圍的比對查詢，以找出可能的候選答案並加以評分。

首先，我們將問句分解成二個部份：疑問詞及關鍵詞組。問句：那些主機板支援 K8 CPU？疑問詞：那些主機板(which type)。關鍵詞組：支援、K8、CPU。藉由答案類型偵測出該疑問詞所對應的答案類型樣版<那些主機板→主機板型號>，搭配關鍵詞進行文件檢索，找出相似度最高的文章，並從中擷取出相對應的解答。

對文件中某一個潛在可能的答案而言，在距離查詢詞彙的範圍內，問句中的詞組(支援、K8、CPU)與答案類型詞彙共同出現的組合，愈靠近答案類型者貢獻的分數愈高。排序加權後的分數，取出前 5 個做為候選答案(candidate answers)。在此我們共分為三個層面予以加權：

1. Match Degree(MD)：代表查詢詞彙(query i)間共同出現在同一篇文章的次數。
2. Pattern Degree(PD)：代表查詢詞彙與可能的答案類型(answer j)共同出現的頻率。
3. Distance Degree(DD)：代表查詢詞彙與可能的答案類型間的距離。

計算公式如下：

$$Score_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n}(MD_{ij}\cdot PD_{ij}\cdot 1/2^{DD_{ij}}) \quad (2)$$

舉例說明計算過程如下：

- 問句：那些主機板支援 K8 CPU？
- 查詢問句的關鍵詞組(支援、k8、CPU)同共出現在同一篇文章，Match degree=3
- 藉由樣版得知：答案類型為<那些主機板->主機板類型>
- 文章內共同出現關鍵詞組與答案類型的詞彙組合為：＜主機板&支援＞、 ＜主機板&k8＞ ＜主機板&CPU＞ ＜k8&CPU＞，包含答案類型的組合共有三組，故 Pattern degree=3
- 關鍵詞組與答案類型詞組間的距離為：＜P4P800(主機板)&支援=3＞、＜GA-7VTXE(主機板)&k8=5＞ ＜MSI 694D PRO(主機板)&CPU=1＞ ，Distance degree 分別為 $1/2^3, 1/2^5, 1/2^1$

$$Score_{p4p800,支援} = (3*3*1/2^3) = 1.125$$

$$Score_{GA-7VTXE\&kk8} = (3*3*1/2^5) = 0.28125$$

$$Score_{MSI694DPRO\&CPU} = (3*3*1/2^1) = 4.5$$

- Score Ranking=<第一組答案：MSI 694D PRO>, <第二組答案：P4P800 >,<第三組答案：MSI GA-7VTXE>

（五）資訊檢索系統

透過中研院未知詞詞性標註系統將文章進行斷詞處理與詞性合併後，配合模版比對擷取出電腦領域的專有名詞，接著，建立各詞彙的文件之特徵向量，向量索引值的建立方式即將文件內容轉化為個別的文件特徵向量，這些文件向量再存入向量索引檔中以提供後續的文件相關度比對使用。如公式(3)所示：

$$\cos(q,d) = \sum_{i=1}^{n} q_i d_i \Big/ \sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2} \quad (3)$$

qi: 問句 (query) 內出現的詞彙之 TF*IDF，di:文件(document)內出現的詞彙之 TF*IDF，qidi 共同出現的詞彙其 TF*IDF。

四、系統實作與評估

（一）系統實作

本研究的資源來源以 Google 線上論壇與華碩網站的問答集為主要蒐集的對象，我們總共下載了 3,000 篇問答集，必須注意的是，每個問題都必須對應到回覆的文章，換句話說，沒有人回覆的文章將被排除在外，因此，去除無人回應的文章 204 篇、重覆的文章 236 篇以及廣告信 495 篇，真正有效的問答集共 2,065 篇。其中 google 論壇有 1,545 篇，華碩網站問答集占 520 篇。

透過中研究未知詞與詞性標記系統進行斷詞處理，總共斷出 23,588 個詞彙，平均每一個問題中含有 17.784 個字(不含標點符號)、7.235 個詞彙；平均每一篇文章有 38.015 個詞彙、5.483 個句子。整體來看，問答集內所討論的文章篇幅稍短且字數偏少。

在問答系統中剖析問句的意圖以及對應的答案類型將攸關整體系統的精確率與召回率，若能正確的判斷一個問句的答案類型將有助於資訊檢索在搜尋上的效能。因此，本實驗的目的在瞭解答案類型的辨識對系統檢索的影響程度。

首先，我們針對詞性合併的效果進行統計，數據如表 6。從統計數據可以看出，詞性合併後的詞彙在討論區中占有一定的比例 23,588 / 4,207=17.84%，如果能夠正確的辨識出合併的詞彙將有助於資訊檢索上的搜尋。

表 6、詞性合併個數

詞類	個數	占總合併詞數比例%
專有名詞	2,284	9.68%
量詞	1021	4.33%
疑問詞	178	0.75%
修飾詞	724	3.07%
總共合併詞彙	4,207	17.84%

我們觀察發現電腦專有名詞的組成大多以英文及數字的組合為主，如：GA-8SR533、Kinston DDR 400，因此，我們以此規則進行專有名詞的的答案類型偵測，嘗試辨識出各專有名詞所屬的類別。

我們總共訓練 2,284 個專有名詞，並依據詞組共同出現的頻率與最小距離位置，計算其權重值，在訓練的過程中，可能因為專有名詞合併過多資訊而形成未知詞彙，若系統無法辨識該未知詞將會給予一個”未知類型(UnKnownType)”。此外，我們發現 Yahoo Kimo 搜尋引擎會提供這類的推薦詞大多是在拼錯字的情況時被辨識出來。例如：查詢關鍵字”AUSU P5G1”，Yahoo 的推薦詞為”ASUS P5G1”，AUSU 已被更正為 ASUS。表示這個詞彙通常較廣泛被使用且網頁數量較多。若遇到這類的情況，系統將會優先擷取以確保該詞彙的完整性，除了可以過濾因過度合併所造成的未知詞彙之外，亦有更正系統辨識錯誤的效果。

在辨識錯誤率方面，我們蒐集名人三 C 電腦賣場所提供的電腦硬體報價單[3]，其內含電腦硬體名詞共 11,233 個，並與本系統所辨識出的專有名詞進行比對，若詞彙出現在報價表內且屬於同一類別則視為正確的詞彙，反之則視為錯誤。其計算方式如下：

辨識正確的個數=總合併的個數 - 辨識錯誤的個數 - UnknownType 的個數。

推薦詞的部分，僅做為輔助合併詞彙在進行搜尋時取得較佳的網頁結果，故推薦詞不納入辨識正確個數的計算中。實驗結果如表 7 所示。

表 7、答案類型偵測實驗數據

詞類	總合併個數	錯誤個數	UnKnownType 個數	推薦詞個數	正確個數
專有名詞	2284	362	354	201	1568

我們觀察實驗結果發現以下幾點特性：

1. UnknownType 的詞彙

合併的規則並無法適用於所有的專有名詞組合，造成部分詞彙合併過多或太少的資訊，影響系統辨識的正確率，例如：合併過多資訊” KM 18G PROVER2”，正確應為：青雲主機板型號” KM 18G”、合併太少資訊”Radeon 9600”、正確應為：青雲顯示卡”Radeon 9600 Pro”。

2. 辨識錯誤的詞彙

在辨識錯誤的情況中，合併的字詞會受到標點符號的影響而造成誤判，例如：Plextor 716>a?。在系統將會被合併成”Plextor 716”以及單字詞”a”，正確應為燒錄機：”Plextor 716a”。

3. 推薦詞的詞彙

我們透過 Yahoo Kimo 所提供的推薦詞彙的確具有修正詞彙的效果，例如：合併詞彙「AUSU P4S 533 MS」，推薦詞為「" ASUS P4S533"」，AUSU 已被更正為 ASUS，這類的詞彙大部分為拼錯字的情況。至於對於合併過多資訊的部分，藉由截去最後一個單字詞的策略，可以將多餘的雜訊排除，承如上例：” AUSU P4S 533 MS”，因為最後一個單字詞”MS”不具代表性，加以排除後，成功的取得詞彙”ASUS P4S533”。

4. 辨識正確的詞彙

長詞優先合併的詞彙確實具有代表性的意義，例如：合併後詞彙 Ti 4200 AGP 8X 與 Ti 4200，前者的資訊較後者完整且

較具資訊含量較高，因此，在剖析網頁時將 Ti 4200 AGP 8X 辨識為「顯示卡」類別時權重較高。此外，”單位詞優先合併”亦有增加正確詞彙的效果，例如：文章內容：「硬體報價：華碩主機板 P5GD1 3399 元，有議者請洽…..」，斷詞與詞性標記「硬體(Na)報價(Na)：(COLONCATEGORY)華碩(Nb)主機板(Na)P5GD1(FW)Pro(FW)3400(Neu)元(Nf)…」，合併規則分別為：{”(Neu),(Nf)”,”(FW},(FW)”,“(FW),(Neu)”}先合併單位詞組：3400 元，再合併 P5GD1 Pro，則可避免合併出 Pro 3399 的錯誤詞彙。

我們分別統計各類別正確與錯誤的辨識率，如表 8 所示。從實驗結果我們得知：「主機板」、「顯示卡」、「硬碟」類別的辨識率較其他類別高，觀察其文章內容發現，這三類中的文章內容在描述專有名詞時都相常詳細且正確，例如：文章：「我有一顆硬碟是 WD 120G SATA 7200rpm…」、「…ASUS A7N8X-X 主機板支援…」、「目前評價較高的顯示卡像 FX5700XP-TD128、…」。透過詞性合併規則，這些文章內的詞彙都被系統正確的辨識成功並且各自歸類到所屬的類別。

表 8、各類別答案類型辨識率

類別	錯誤率	正確率
主機板	14.16%	85.84%
處理器	42.86%	57.14%
硬碟	17.47%	82.53%
記憶體	34.03%	65.97%
顯示卡	10.28%	89.72%
光碟機	22.63%	77.37%
燒錄機	19.91%	80.09%
音效卡	21.05%	78.95%
營幕	24.05%	75.95%
掃瞄器	20.59%	79.41%
印表機	23.08%	76.92%

然而，「處理器」與「記憶體」類別的錯誤率偏高，歸結其原因在於這兩類的專有名詞大多包含某些特定的標點符號，例如：P4 CPU 2.5G 與 DDR400 512MB*2，標點符號造成此兩個類別無法正確的合併，進而影響到後續的答案類型偵測上。

整體來看，合併的專有名詞平均正確辨識率為 77.26%，錯誤率為 22.74%，若能改善上述的例外情況，應該可以有更佳的辨識效果，將於未來研究中再以說明。

（二）系統評估

為了有效評估系統精確率與召回率，我們將文件集內的問答集整理與分類，從本研究所定義的十一的類別中分別中抽出 20 篇問答集，再加上有關詢問價格、單位名詞以及網址的問答集各 20 篇，共計十四類，合計 280 篇，並且針對各類別設計出三個問題，共 42 個測試問題(附錄三)，每個問題以前 5 個答案分別計算其系統的精確率與召回率，計算方式如下：

1. 精確度(Precision)=正確答案的筆數 / 回傳答案的篇數
2. 召回率(Recall)=正確答案的筆數 / 文件集內可回答問題的篇數

此外，我們採用 TREC QA 中用來衡量問答系統的 RAR(Reciprocal Answer Rank))指標，RAR 的值愈高，代表系統能夠更快速、有效率的提供使用者確切的答案。如公式(4)、公式(5)所示：

$$RAR = 1/Rank_i \quad \text{i=1,2,3,4,5} \tag{4}$$

$$MRAR = (1/N)\cdot\sum_{i=1}^{n} RAR_i \quad \text{N=42} \tag{5}$$

以本實驗為例：本實驗共有 42 個問句 N=42，每個問句會回傳前 5 筆答案 n=5，若答案出現在第一筆則 i=1, RAR=1/1=1 分，另一答案出現在第二、

三、四、五筆，其 RAR 分別為 1/2=0.5 分, 1/3≒0.3 分, 1/4=0.25 分, 1/5=0.2 分。MRAR(Mean Reciprocal Answer Rank)為 RAR 之加總平均。

從實驗結果(表 9)發現,有部分問題在文件集找到答案的比例偏低，尤其以「音效卡」類別僅找到三個答案。可能的原因是抽取的文件內討論音效卡的文章不多，導致系統在答案評分階段無法給予適當的加權值，影響了系統的精確度。此外，我們分析錯誤答案與找不到的文章，發現查詢關鍵字與類別詞組間的距離太遠，造成雜訊過多，影響候選答案在評分上的精確性。

表 9、採用答案類型輔助資訊檢索之效能

	答案在第一筆	答案在第二筆之後	ΣRAR	MRAR
價格	3	2	3.6416	0.728
數量	2	2	2.6416	0.660
網址	2	4	3.2832	0.547
主機板	5	5	6.604	0.660
處理器	3	3	3.9624	0.660
記憶體	2	5	3.604	0.515
硬碟	3	6	4.9248	0.547
顯示卡	2	7	4.2456	0.472
音效卡	2	1	2.3208	0.774
燒錄機	2	6	3.9248	0.491
光碟機	1	4	2.2832	0.457
營幕	1	3	1.9624	0.491
掃描器	2	3	2.9624	0.592
印表機	2	2	2.6416	0.660

相較之下，「主機板」與「顯示卡」與「硬碟」以及「燒錄機」這四類的 RAR 值較其他類別高，代表這四類能較快速、有效的找到正確的答案，探討其原因可能是：

1. 抽出的文件集中討論這四類的問題居多，通常回覆的文章很明確的回答到問題的本質，例如：選那一塊主機板比較不容易爆漿？答案：建議你買 K8N NEO2 比較不容易爆漿….。
2. 這四類的答案類型偵測的辨識率較高，在召回的文件數相對增加。
3. 有關價格與網址方面的問題，這類的問題回覆多半較制式，幾乎都能夠正確的找出答案，例如：Lite-on 燒錄機多少錢？答案：光碟商場報價 lite-on $1900 是目前館內最價宜的燒錄機…。
4. 除此之外，本系統能夠正確辨識出「否定問句」的問題類型，如問句為：不容易過熱 CPU 的有那些？答案:p4 3.0 以下的 cpu 一般來說比較不容易有過熱的情況，但是要看…..。

為了有效評估系統的精確度與查全率，我們另外設立一組採用相似度計算

(TFIDF)的方式進行比較，查詢問句與回覆答案間的相似度門檻值分別以 0.8, 0.6，分別計算其精確率與召回率，其實驗結果如表 10 所示。

表 10、本系統與採用相似度計算的精確率與召回率

系統	精確率	召回率
本系統	0.40476	0.42929
TFIDF(0.6)	0.30519	0.23737
TFIDF(0.8)	0.19355	0.09091

整體來看，採用答案類型偵測輔助的精確率與召回率都較未採用的系統其效能有明顯提升，與門檻值為 0.6 做相似度比較，在精確率與召回率上分別提升 9.95%與 19.19%；與另一組門檻值為 0.8 比較，其精確率與召回率大幅提升到 21.12%以及 33.83%，我們觀察發現：回傳答案的文章絕大多數皆在討論與主題相關的議題，以致於進行候選答案評分加權時，分數較高的段落或句子皆能有效的回答使用者的問題。

五、結論與未來研究方向

（一）結論

傳統的問答系統在專業領域的檢索上，會以分類階層架構輔助查詢，目的在縮小文件檢索的範圍，再從中找出與問句相似度最高的文章，但對於電腦類的專有名詞，例如：文章內提及：P4P800SE 支援 DDR 533 雙通道…，其辨識上明顯不足，以致於在計算段落或句子時，無法適當的賦予權重而影響系統的精確度。

本研究發現電腦類別的專有名詞構成要素以英文(FW)與數字(Neu)二種詞性居多，我們總共歸納出二十條的經驗法則進行詞性合併，從實驗結果得知：詞性合併可以有效的擷取大部分的特徵值，辨識率在七八成之間。

除此之外，本研究提出一個以搜尋引擎為基礎的答案類型偵測機制，藉由偵測問句的答案類型，輔助資訊檢索搜尋並且從回傳的文件中找出候選答案，實驗結果顯示採用答案類型偵測輔助資訊檢索的方式其精確率在四成左右，相較於相似度計算的方式其精確率提升了 9-21%，而召回率提升了 19-27%，在評估系統的效能上，RAR 值最裔的四類「主機板」、「硬碟」、「顯示卡」、「燒錄機」分別為 6.604、4.9248、4.2456、3.9248，代表系統在回答這四類的問題時的其效能較其他類別佳。

（二）未來研究

1. 語意分析處理

由於本研究僅處理針對定義的問句類型進行詞彙處理，造成無法對答案文件做更深一層的語意解析，建議在未來研究中可結合專業領域的 Ontology 架構，透過領域內定義的從屬關係，輔助字詞上的語意判斷事必可提升答案的精確度。

2. 詞性的文法結構

在本論文針對答案類型擷取的判斷機制上，考量僅查詢詞彙與和答案詞彙間共同出現的頻率關係與距離，未來，在擷取文章或句子的過程中可考量兩者出現的先後順序，以排除詞彙之間的重疊性，避免過度的加權。

3. 相關回饋

由於答案構成因素可能是一個專有名詞、段落、句子或文章，本論文僅儲存使用者查詢的問句與回傳的答案，並未對其內容進行分析，未來可考慮從系統回傳的結果進行剖析，藉由使用者對答案的評分來提升系統的精確度。

4. 答案偵測辨識率

在訓練答案類型的過程中所定義的十一組類別關鍵詞，若能增加特徵值的詞組，如：<主機板&晶片組&南北橋晶片…>，將有助於判斷詞彙所屬的類別，因為特徵值共同出現的頻率愈高，代表該詞組具有一定相關性與鑑別性。然而，電

腦產業的快速發展也令類別關鍵詞將會隨時更新，因此如何使系統能夠不至於延遲於電腦產業的發展之外，也是未來的研究課題。

六、參考文獻

[1] 中文詞知識庫小組, *中文詞類分析*,

[2] 中研院 CKIP 詞庫小組, *Team of Chinese Knowledge Information Processing(CKIP)*, http://godel.iis.sinica.edu.tw/CKIP

[3] *名人 3c 電腦量販*, http://www.mren.com.tw/

[4] 李季 and 孫冀俠, "一個簡單的中文問答系統." vol. 6: 維普資訊, pp. 64-66, 2004.

[5] 秦兵, 劉挺, 王洋, 鄭實福, and 李生, "基於常問問題集的中文問答系統研究." vol. 35: 維普資訊, pp. 1179-1182, 2003.

[6] 崔桓, 蔡東風, and 苗雪雷, "基於網路的中文問答系統及資訊抽取演算法研究." vol. 18: 維普資訊, pp. 24-31, 2004.

[7] *重編國語辭典修訂本*, http://www.sinica.edu.tw/~tdbproj/dict/

[8] 劉玉琛, *標點符號用法*, http://myweb.hinet.net/home11/kuangten/wskill/ch10.htm

[9] *Google*, http://groups.google.com.tw/group/cn.bbs.comp.hardware

[10] G. T. Huang and H. H. Yao, "Chinese Question-Answering System." vol. 19: 萬方資料資源系統, 2004.