使用詞夾子建立中文典籍分析加值服務

Term Clips to Build Chinese Ancient Books Analyses and Value-Added Services

謝育平 楊龍廉 趙建宏

Yuh-Pyng Shieh **Lung-Lian Yang** **Jian-Hung Jau**

黃銘立 古馮文 林郁智

Ming-Li Huang **Feng-Wen Gu** **Yu-Jr Lin**

銘傳大學資訊工程學系

Department of Computer Science and Information Engineering,
Ming Chuan University

摘要

當我們在一般搜尋引擎針對問題，進行搜尋動作時，回傳結果的單元數少則兩百，多則上千，需要花上一些時間閱讀分析，才能引導自己繼續往下搜尋真正想要的資訊。所以搜尋結果的自動閱讀分析代理程式，就顯得相當的重要。目前通用型搜尋引擎未提供此類功能的主要原因，是設定搜尋的領域太廣造成搜尋結果分析困難。本文主要選用 432 部中文典籍，實作搜尋引擎及自動閱讀分析代理程式，供中文領域研究學者使用，並作為中文典籍領域的研究基礎建設。

在自動閱讀分析實作上，本文主要使用詞庫頻率分析呈現，在其中 96 部典籍中，使用詞夾子萃取 139 個分類詞庫，計 21538 個詞。處理一個分類詞庫平均約需 14 分鐘的機器時間及 35 分鐘的人工時間，平均新增 64 個詞。

1. 緒論

中文典籍研究歷史悠久，研究人員眾多，純粹喜愛的讀者也多，單就典籍內容來說，就已經是不得了的傑作。近年來，由於資訊科技的發達，使得中文典籍的研究也邁入了數位化的時代，但是之中最重要的基礎建設是「數位化」，古騰堡計畫[26]及開放文學網[33]為了這樣的基礎工程提供了無版權糾紛的原始碼供各界加工使用。目前中文典籍的搜尋網站，仍以單一典籍服務為大宗，例如紅樓夢網路教學研究資料中心[30]、《紅樓夢系列》[21]、紅楼梦谭[29]、中華典籍網路資料中心－紅樓夢網路教學研究資料中心[24]、紅楼梦网[28]、三國演義網上版 Romance of the Three Kingdoms Online[20]，尚無完整有系統的搜尋後分析服務。本文目前主要使用開放文學網的所有的 432 部中文典籍進行加工處理，期望提供搜尋功能及閱讀分析功能，給中文典籍的研究學者及普通讀者使用。(註:2008 年 8 月時開放文學網有 432 部中文典籍，2008 年 12 月已增至 480 部。)

一般搜尋引擎的搜尋引擎是為了服務所有的使用者，而導致搜尋後結果相當廣闊論。一般說來，搜尋資料牽涉到的領域越廣，後分析處理的難度就越高，所以本文選擇中文典籍領域來降低難度並創造價值。

自動閱讀分析功能範圍很廣，有自動摘要文章系統，有標籤雲系統，有相關文章或搜尋詞推薦等。自動摘要文章系統主要對同質性的文章(例如:新聞文章)進行摘要，萃取部分段落當成該文章的摘要。自動摘要文章系統對中文典籍說並不合適，主要原因是各典籍文體不同，作者不同，風格不同，想要使用相同的系統，進行成功的自動摘要文章系統並不容易。而標籤雲系統，主要使用標籤對各單元進行描述，在搜尋結果中進行標籤出現的頻率，進而顯現某些標籤的重要性，在這樣的設計下，各標籤對各單元描述的基礎工程就變成一項基礎工程，這項基礎工程將決定系統的成功與否。Flicker[31]是一個相片搜尋網站，他的成功建立在讓廣大的使用者幫忙建置標籤工程。，讓使用者幫忙標籤工程，這對本文初建系統來說是困難的。相關文章推薦，主要是根據一篇文章進行相關文章推薦，而相關蒐尋詞推薦則是根據搜尋結果進行相關蒐尋詞推薦。對於中文典籍搜尋暨分析系統來說，領域比較聚焦，所以可以有的分析也比較可以客製化。基於以上的討論，本文使用標籤雲系統，但是標籤雲的顯現方式略嫌雜亂(如圖 4)，所以改採條狀頻率呈現(如圖 8)。所以接下來的工程是有兩樣，一是建立中文典籍搜尋引擎，二是建立自動閱讀分析系統。

建立中文典籍搜尋引擎主要考量搜尋的單位，Google 搜尋引擎的搜尋單位是網頁，新聞檢索系統的搜尋單位是一則新聞。中文典籍的搜尋應該以段落為單位，主要原因是段落通常是一個情境的描述，類似電影或電視劇中的場景切換，具有獨立可閱讀的意涵，另一個原因是以章回為單位則範圍太大，例如一部《紅樓夢》只有 120 回，搜尋「寶玉」將回傳 116 回的資料、那跟叫使用者看完整部典籍沒有不同；而在《施公案》100 回中搜尋「施公」也傳回 91 回的資料。所以 432 部中文典籍的分段工程及搜尋引擎的實作，將是這部分的重點。

建立自動閱讀分析系統主要考量讀者或研究學者需要哪些分析，本文目前主要採用標籤雲系統，但是標籤雲本身並未被分類，所以實用性對學者來說略嫌不足，想像對三國演義讀者來說，搜尋結果的標籤雲中，期待可以分類為人名、地名、武器名、坐騎名等，這樣的分類可以讓使用

者在組織後分析標籤中更佳的方便。432 部中文典籍，文體不一，內容不一，所已針對性的客製化應該被採用。例如紅樓夢讀者期待標籤雲可以分類為人名、官名、器皿等，卻絕對不會有武器名及坐騎名等。本文既然採取標籤雲系統，標籤從何而來將是一個最大的問題。龐大的標籤應該自動抽取，自動抽取的標籤應該要有品質保證。本文的作法是針對每一部典籍進行分類詞庫的建立，例如三國演義則建立人名、地名、武器名、坐騎名等詞類，本文使用謝育平老師指導學弟張尚斌建置的詞夾子演算法[11]，使用人工半自動從三國演義本文中萃取各類之詞。這樣的人工半自動萃取工程，每部每詞庫所花時間不一。然後在搜尋結果集中使用該典籍各類詞庫進行標籤標記及標籤分類導讀。為了瞭解中文領域學者可能關心哪些詞類，我們特別蒐尋中文領域論文以做研究。

本文組織如下，第二章為中文典籍文獻探討及實驗，第三章為詞夾子與搜尋引擎，第四章為中文典籍與分類詞庫，第五章為結論與討論，第六章為參考文獻。

2. 中文典籍文獻探討及實驗

本文主要目的是針對現存中文古典典籍，製作後分析處理的搜尋引擎。在當我們針對一部典籍進行研究時，無可避免地需要對文章內的人、事、時、地、物進行了解；以賴采蘋「《搜神記》中的動物類型研究-以動物與人類的關係為中心」為例，論文中將 246 種動物作為觀察對象，再細分成 5 大類以此來觀察牠們自古與人類之間的關係,如此數目繁多的動物跟人物角色當作關聯對象，我們具有分類標籤呈現的搜尋引擎，能方便作者藉由動物與人物辭庫，來做動物與人的關聯，利於他完成整篇的論文。又以胡玉珍《西遊記中的精怪與神仙研究》[8]為例，該論文主要針對特定地區居民對於鬼神崇拜的現象來做分析，作者除了精讀西遊記外，也須蒐集及其它有關特定地區的相關資料，所以在 Google 搜尋引擎上，搜尋「西遊記 火焰山」是有可能的，進而搜尋西遊記中火焰山出現怪物的相關資料，但在這樣的蒐尋過程中，為了避免資料遺漏，逐一閱讀搜尋結果是免不了的，但這樣卻可能得看過上千篇的搜尋結果，這是非常消耗時間跟人力，此時我們構想出如果有一個代理程式,能先幫作者看過整個搜尋的結果並整理出作者可能想要的資訊跟關聯的資訊，再利用程式明顯標出怪物在有火焰山的那篇幅，節省她閱讀分析的寶貴時間。

在我們小組研究中文領域的相關論文後，發現中文領域學者所研究的範圍相當的廣闊，有最基本的人物探討，再者思想研究、女性形象，更遠及服飾配件的研究五花八門。以下列出學者對於中文典籍中的某些物件類別(例如:鬼怪)特別感興趣並撰寫而成的論文。

(表 1)中文典籍論文所關心的詞庫

書名	詞庫	對映論文
水滸傳	女性	《水滸傳》中的女性及其影響[14]與《水滸傳》女性研究[5]
搜神記	動物	《搜神記》中的動物類型研究—以動物與人類的關係為中心[16]
三國演義	男子服飾	《三國演義》中男子服飾的角色刻劃效應—以曹操、關羽、諸葛亮為中心的比較研究[19]
聊齋誌異	女妖	聊齋誌異婦女形象研究 [8]與《聊齋誌異》女妖故事研究[12]
聊齋誌異	鬼狐仙妖	《聊齋誌異》鬼狐仙妖研究[9]
儒林外史	女性	論《儒林外史》女性形象塑造[6]

紅樓夢	賈府女性	《紅樓夢》中賈府女性人物論[7]
紅樓夢	十二金釵	《紅樓夢》十二釵命運觀之研究[2]
紅樓夢	飲食	《紅樓夢》飲食情境研究[17]
紅樓夢	丫鬟與小姐	丫鬟與小姐之互動關係研究—以《紅樓夢》為主的論述[13]
紅樓夢	女性	《紅樓夢》的女性認同[3] 、《紅樓夢》與《鏡花緣》的才女意義析論[4] 與 父權社會下的女兒國-《紅樓夢》女性研究[15]

以實際操作來當例子，當我們想知道關羽為何擁有赤兔馬的關係時，我們可以在書籍清單中選擇三國演義，並輸入「關羽 赤兔馬」做搜尋，會呈現「關羽 赤兔馬」共找到 25 段，並在右邊人名統計顯示出前兩名曹操 11 段 21 次、呂布 7 段 25 次，得知如果我們想要更了解赤兔馬和關羽必須得再針對曹操和呂布做些研究，我們就在利用右邊分頁點選曹操跟呂布作進一步的關聯式搜尋，分序搜尋「關羽 赤兔馬 → 呂布」並過濾出 6 段關羽、赤兔馬和呂布的段落，其中「關羽 赤兔馬 → 呂布」的→表示在「關羽 赤兔馬」的搜尋結果中再搜尋呂布的結果。在第三回十三段中的對話「告布曰：『此是董公久慕大名，特令某將此奉獻。赤兔馬亦董公所贈也。』」，第二十五回 第八段「公曰：『莫非呂布所騎赤兔馬乎？』操曰：『然也。』遂并鞍轡送與關公。」，很明顯了解到呂布的赤兔馬是董卓所贈與，之後曹操在二十五回中為了討好關羽就將赤兔馬送給了他；在分序搜尋「關羽 赤兔馬 → 曹操」並看過出來的段落的文章可以找到第七十七回第四段「關公既歿，坐下赤兔馬被馬忠所獲，獻與孫權。權即賜馬忠騎坐。其馬數日不食草料而死。」最後在關羽死掉後赤兔馬因其喪失主人也死掉了。

總結所有在經過後分析的結果我們很輕易的了解到董卓贈馬給呂布、曹操送馬關羽和關羽死赤兔馬死的關係，在搜尋例子中搜尋到的段數只有全三國演義書中的 25 段，卻是從第三回、第二十五回、第七十七回所得到的結果，如果我們不靠機器利用人工我們得從第三回提到赤兔馬開始閱覽直到第七十七回赤兔馬死才能夠的到結果，這種幫助使用者閱讀並節省其閱讀時間的功能就是本研究的目的。

(圖 1)關羽 赤兔馬 與 曹操 呂布 關聯式分析呈現畫面

我們運用搜尋引擎來做中文古典典籍的分析和研究，想讓使用者不必花大量的時間整理中文古典典籍的資料，例如：以胡玉珍《西遊記中的精怪與神仙研究》[10]為例，該論文其中有針對特定地區所出現的人間精怪來做分析，作者除了精讀西遊記外，也須蒐集及有關特定地區的相關資料，而且資料可能不齊全，如原先論文中出現在西遊記中地點的精怪，經由我們系統再對西遊記中地點進行搜尋分析，利用角色顏色標示的方法，迅速的找出在居住在這地點的所有怪物，不但補全了原論文中遺漏的精怪還額外找出新住所及新的居住精怪，如(表 2)所示之，這溫馨的功能能幫助使用者節省大量的閱讀時間。

(表 2) 論文中所列精怪

住所	該論文中所列精怪	本系統搜得出以該地點為住所的精怪
雙叉嶺	老虎精 熊羆精 野牛精	揭諦
黑風山 黑風洞	黑大王 蒼狼怪 白花蛇怪	雷公
火雲洞 鑽頭號山	紅孩兒	牛魔王 四海龍王 山神 金蟬子 雷公 羅剎女 南海龍王 牛王 電母 西海龍王 北海龍王 老龍王
鍾南山	黃毛虎 白毛角鹿 羚羊	烏雞國王
破兒洞 解陽山	如意真仙	紅孩兒
火焰山 芭蕉洞	牛魔王 玉面公主 羅剎女	聖嬰大王 雷公 六丁六甲 風魔 巨靈神 木叉 揭諦 靈吉菩薩 山神 托塔李天王 鐵扇仙 紅孩兒
柳林坡 清華洞	美后	比丘國王
陷空山 無底洞	金鼻白毛老鼠精	托塔天王 哪吒 山神 揭諦 護國天王

隱霧山 折岳連環洞	蒼狼怪 豹子精	揭諦
豹頭山 虎口洞	黃獅精 狻猊獅 博象獅 白澤獅 伏狸獅 猱獅 雪獅	雷公 猴王 九頭獅子
青龍山 玄英洞	辟寒大王 避暑大王 避塵大王	四值功曹 增長天王 太白金星 二十八宿 揭諦
七絕山 稀柿同	紅鱗大蟒	揭諦
衡陽峪 黑水河神府	小龘龍	四海龍王 西海龍王 玉帝 雷公
通天河 水龘之地	老龘	
毒敵山 琵琶洞	蠍子精	
亂石山	九頭蟲 萬聖龍王	
盤絲嶺 盤絲洞	蜘蛛精 蜈蚣精	
白虎嶺	白骨夫人	四海龍王
系統額外發現新住所及新居住精怪		
峨眉山 清涼洞		勝至金剛
崑崙山金霞嶺		永住金剛
五台山秘魔巖		潑法金剛
須彌山摩耳崖		大力金剛
毛穎山		山神
平頂山 蓮花洞		金角 銀角
朱紫國		東宮太子
百腳山		蜈蚣精
獬豸洞		賽太歲
蟾宮		玉兔
獅駝洞		獅王
亂石山		奔波兒灞 灞波兒奔 斯魚怪 黑魚精 萬聖公主 九頭駙馬 象王 大鵬

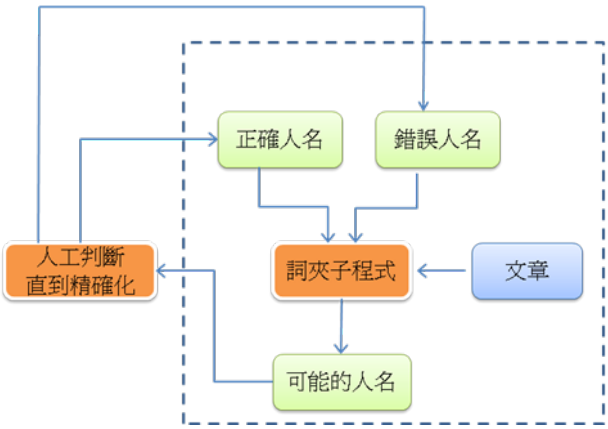
3. 詞夾子與搜尋引擎

詞夾子演算法是本文第一位作者在台灣大學擔任博士後研究時指導學弟研究而成[11]。詞夾子演算法是用來在某一文件集中萃取某類資料(例如:在西遊記中萃取地名),我們使用的是半自動的人工介入方式。程式啟動時，人工先提供文件集中少數正確的地名(火燄山)，及文件集中少數會引起誤解的地名；然後程式會自動產生關於正確地名集的詞夾子，並使用詞夾子在文件集中抓取更多候選詞，但是由於是機器抓的，所以還是需要人工過濾處理，處理後將正確的詞放入正確集中，錯誤的詞放入錯誤集中，再使用程式進行萃取，此時程式將更清楚知道我們心中想要的詞類，因為程式是不懂什麼叫做地名的，他接受的指令是尋找很多跟火焰山有相同性質的詞，所以正確樣本集越多，則建立的詞夾子就越能貼切我們心中想要的詞類。詞夾子演算法在 David

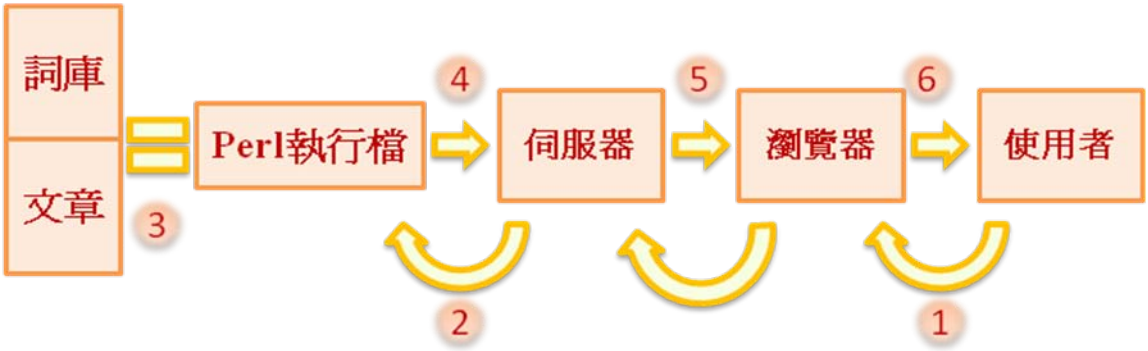
Nadeau, Satoshi Sekine [20]的分類中是屬於半指導式的學習演算法（Semi-supervised learning）中的交錯是學習(bootstrapping)。

「詞夾子」主要概念是利用文章寫作上的一些特定習性與字辭之間的耦合關係，來找出專有名詞。先給予樣本詞，然後找出和樣本詞相關的 **詞夾子** ，並利用這些 **詞夾子** 找出與樣本詞類似的候選詞出來，之後以迭代方式不斷的產生 **詞夾子** 和候選詞。當我們希望機器在文件集中尋找地名時，我們很難精準的形容何謂地名，如果查字典，得到的結果也許是“區域之名稱”，但這樣的定義對萃取來說並沒有幫助。所以我們對地名的定義可能是那些 1. 後方可以接行政單位名的名詞：例如XX省、XX縣、XX區。2.前方可以接“去”、“到”、“往”、“上”、“下”等行動動詞的名詞：例如去XX、到XX、往XX等。當然還有更多的規則用以定義地名，卻寫也寫不完。所以我們採取的策略是使用學習的機制讓機器能自動學習定義“地名”的規則。至於定義“地名”的規則，我們則以詞夾子的概念呈現。所謂的詞夾子係指一個候選詞出現位置的前文、後文、前 綴、後綴等特徵要求。以“到中山區吃宵夜”為例，“到”是“中山區”的前文，“吃”、“吃宵”及“吃宵夜”為“中山區”的三個後文，“中”、“中山”則為“中山區”的前 綴而“區”、“山區”則為“中山區”的後綴，為了機器自動操控，我們定義空字串 “” 為當然的前文、後文、前綴及後綴。以更數學的方式來討論，一個詞夾子c=(f，p，s，r) 即為四個字串的組合分別代表(前文，前綴，後綴，後文)的要求。例如c=(去，，，吃)則表示要求候選詞出現位置的前文必須是“去”，後文必須是“吃”而對其前綴、後綴並沒有特殊的要求。在我們設定一個候選詞(k=“中山區”)之後，在一個文章字串d中可能出現好幾個候選詞，我們就使用自然數來編號是別之(即以k_1表示k的第一個出現，k_i表示第i個出現)。我們說候選詞k的一個出現k_i滿足一個詞夾子c=(f，p，s，r)，如果f，p，s，r分別為k_i的一個前文、一個前綴、一個後綴、一個後文。我們說候選詞k滿足詞夾子c，如果候選詞k的一個出現k_i滿足詞夾子c。而我們說一個詞夾子在文章字串d中夾中了候選詞k，則表示候選詞k滿足詞夾子c。

以候選詞k= “中山區”且文章字串d=“到中山區吃宵夜，然後再去中山區最著名的KTV唱歌”為例，候選詞k= “中山區”有兩個出現(occurrence)，前一個我們以k_1稱之、後一個以k_2稱之	
k_1之前文	**空字串、到**
k_1之前綴	**空字串、中、中山、中山區**
k_1之後綴	**空字串、區、山區、中山區**
k_1之後文	**空字串、吃、吃宵、吃宵夜**
k_2之前文	**空字串、去、再去、後再去、然後再去**
k_2之前綴	**空字串、中、中山、中山區**
k_2之後綴	**空字串、區、山區、中山區**
k_2之後文	**空字串、最、最著、最著名、最著名的…**
K 滿足之詞夾子	**(到，，吃)、(去，，，)、(去，，，最)、(到，，區，)**
K 不滿足之詞夾子	**(到，，最)、(去，，，吃)、(往，，，去)、(到，，，去)**



(圖 2) 詞夾子程式萃詞示意圖



(圖 3) 本文搜尋引擎架構圖

本文建立的搜尋引擎資料路徑大致如下。當①使用者透過瀏覽器對伺服器下達搜尋的指令時，②伺服器即呼叫Perl程式執行，程式根據使用者所下的查詢典籍及查詢詞句要求，③即時讀取並篩選所需的段落資料，然後使用典籍分類詞庫比對段落建立資料，④並將結果包裝為輕量化的資料格式，⑤再透過網路傳送給使用者，當瀏覽器收到伺服器所回傳的資料，⑥即直接包裝為便於使用者瀏覽的資訊來呈現。這樣設計的搜尋引擎並未使用索引檔，純粹是動態進行全文檢索及字串比對，雖然說這樣的設計有慢速之嫌，但是因為中文典籍查詢在預設上是針對某一典籍進行查詢處理，資料量甚少，使用此一設計在速度上已能輕鬆應付，沒有問題。甚至除了一般輸入文字搜尋外，也提供了利用邏輯運算元的進階搜尋方式：AND、OR跟NOT搜尋方法。例如查詢「張飛+關羽」則搜尋文章中張飛和關羽「同時」出現的段落；查詢「張飛–劉備」則搜尋文章中出現張飛但沒出現劉備的段落；而當我們查詢「關羽 赤兔馬」則表示搜尋文章中出現關羽或是出現赤兔馬的段落。

在技術選擇上，由於我們系統大部分時間都在執行文字上的處理，而 Perl 語言在字串的處理能力上非常強大，正好符合我們的要求，故選擇了 Perl 做為主要程式的撰寫語言。除此之外、我們認為一個系統除了要具有功能性，還需具有一定的介面親和力，以及不嫌繁雜的介面，如此

才能讓使用者在搜尋時不感到手續繁複，因此我們選擇了 Dojo 這個 JavaScript 的 tookit，主要採用其 UI 的部份來達到基本的美觀，並利用 JavaScript 來盡量簡化操作的過程。

本文使用標籤雲系統 Tag Cloud 或稱作 Word Cloud(圖 4)，他會根據搜尋結果單元所標誌的標籤進行統計，並將每個標籤上加權重目的是在最後呈現時產成大小不一，讓使用者看過就一眼就明白搜尋結果集中各標籤的權重。在本研究剛開始時我們討論過，使用標籤雲來表現右分頁的統計字詞可能會使的畫面變得更美觀，且使用者會以字型大小，清楚的明白字詞重要性，但是最終的結論是標籤雲佔據了一大塊矩形面積、多筆資料顯示雜亂，當使用者想得知每搜尋結果的數據上，是相當不方便的，因此我們採用了條列式加上統計條的方式，並依照統計條的大小來做排序，左邊分頁以顏色明顯標示出搜尋的關鍵詞，方便於使用者閱讀，而右分頁會出現搜尋結果的段數及其他名詞出現在該段的次數，例如在(圖 4)為針對紅樓夢一書「寶玉」關鍵詞的搜尋結果，在右邊分頁的分析可見在有寶玉一詞出現的段數中，賈母和襲人是出現最多次的，針對寶玉做研究使用者就可得知，要研究寶玉可能還須對賈母和襲人一詞加以研究，節省反覆觀看文章的時間跟便利性。

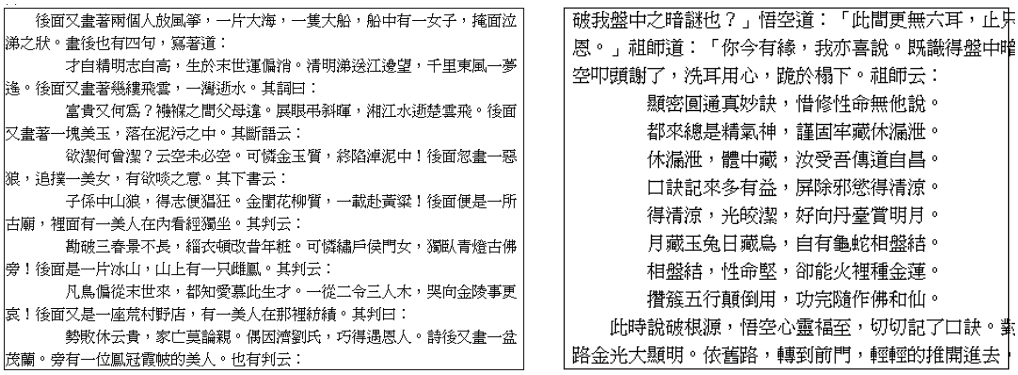
(圖 4) 標籤雲及統計條呈現

4. 中文典籍與分類詞庫

為了製作一個中文典籍領域的搜尋引擎，我們選用一個無版權爭議的開放文學網站[33]，他標榜提供無版權或放棄版權的文章，以宣揚中國文化之國粹的方式，讓大家可以在線上閱讀或下載傳播。我們處理在其中 423 部典籍當作為目標，也在全國博碩士論文資訊網[27]找了相關典籍的研究論文。所有列表整理於[23]。

建立中文典籍搜尋引擎主要考量搜尋的單位。中文典籍的搜尋應該以段落為單位，主要原因是段落通常是一個情境的描述，類似電影或電視劇中的場景切換，具有獨立可閱讀的意涵，另一個原因是以章回為單位則範圍太大，例如一部《紅樓夢》只有 120 回，搜尋「寶玉」將回傳 116 回的資料、那跟叫使用者看完整部典籍沒有不同；而在《施公案》100 回中搜尋「施公」也傳回 91 回的資料。還有一個原因是程式進行分析統計時，只需要對那些小段落進行分析統，可以加快反應的速度。所以 432 部中文典籍的分段工程及搜尋引擎的實作，將是這部分的重點。

從開放文學網中取得的資料中，根據不同的狀況進行分段工程，例如:許多書籍內文中存在有些段落過短的問題，無法在其中取得任何具有價值的結果(如圖 5)。還有文章內容可能存在兩人或多人對話的段落，其內容可能是某人說一句話就一個段落，若利用機器分段可能造成如上述問題的結果(如圖 5)。本研究亦有線上閱讀的功能，在過長的段落裡再給予分段可以讓讀者免看一大串才找到想要的結果。



後面又畫著兩個人放風箏，一片大海，一隻大船，船中有一女子，掩面泣涕之狀。畫後也有四句，寫著道：

才自精明志自高，生於末世運偏消。清明涕送江邊望，千里東風一夢遙。後面又畫著幾縷飛雲，一灣逝水。其詞曰：

富貴又何爲？襁褓之間父母違。展眼弔斜暉，湘江水逝楚雲飛。後面又畫著一塊美玉，落在泥污之中。其斷語云：

欲潔何曾潔？云空未必空。可憐金玉質，終陷淖泥中！後面恕畫一惡狼，追撲一美女，有欲啖之意。其下書云：

子係中山狼，得志便猖狂。金閨花柳質，一載赴黃粱！後面便是一所古廟，裡面有一美人在內看經獨坐。其判云：

勘破三春景不長，緇衣頓改昔年粧。可憐繡戶侯門女，獨臥青燈古佛旁！後面是一片冰山，山上有一只雌鳳。其判云：

凡鳥偏從末世來，都知愛慕此生才。一從二令三人木，哭向金陵事更哀！後面又是一座荒村野店，有一美人在那裡紡績。其判曰：

勢敗休云貴，家亡莫論親。偶因濟劉氏，巧得遇恩人。詩後又畫一盆茂蘭。旁有一位鳳冠霞帔的美人。也有判云：

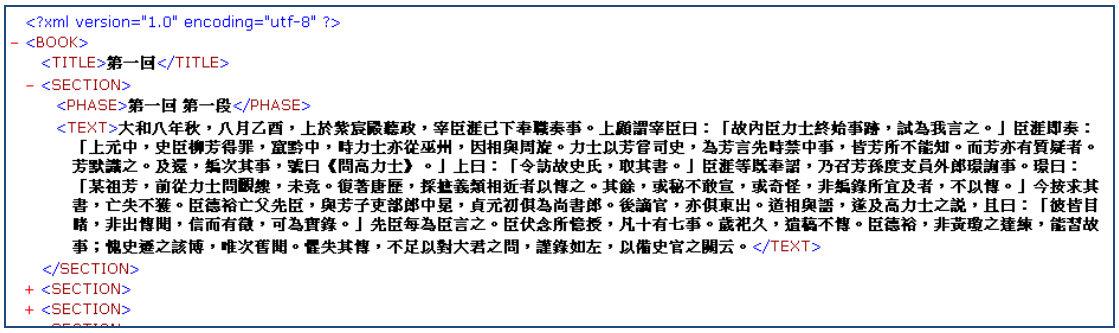
破我盤中之暗謎也？」悟空道：「此間更無六耳，止朱恩。」祖師道：「你今有緣，我亦喜說。既識得盤中暗空叩頭謝了，洗耳用心，跪於榻下。祖師云：

顯密圓通真妙訣，惜修性命無他說。
都來總是精氣神，謹固牢藏休漏泄。
休漏泄，體中藏，汝受吾傳道自昌。
口訣記來多有益，屏除邪慾得清涼。
得清涼，光皎潔，好向丹臺賞明月。
月藏玉兔日藏烏，自有龜蛇相盤結。
相盤結，性命堅，卻能火裡種金蓮。
攢簇五行顛倒用，功完隨作佛和仙。

此時說破根源，悟空心靈福至，切切記了口訣。業路金光大顯明。依舊路，轉到前門，輕輕的推開進去

(圖 5)

所以為了優質化搜尋的結果，我們將收集好的典籍本文進行分段工程並製成 XML 檔如圖 6。其中<BOOK>標籤代表一部典籍、<TITLE>標籤代表一部典籍的第幾回、標題<SECTION>標籤代表第幾回的第幾個部份、<PHASE>標籤代表第幾回第幾段而<TEXT>標籤裡面放第幾回第幾段內文。



```
<?xml version="1.0" encoding="utf-8" ?>
- <BOOK>
    <TITLE>第一回</TITLE>
  - <SECTION>
      <PHASE>第一回 第一段</PHASE>
      <TEXT>大和八年秋，八月乙酉，上於紫宸殿聽政，宰臣漼已下奉職奏事。上顧謂宰臣曰：「故內臣力士終始事踐，試為我言之。」臣漼即奏：「上元中，史臣柳芳得罪，竄黔中，時力士亦從巫州，因相與周旋。力士以芳嘗司史，為芳言先時禁中事，皆芳所不能知。而芳亦有質疑者。芳默識之。及還，編次其事，號曰《問高力士》。」上曰：「令詰故史氏，取其書。」臣漼等既奉詔，乃召芳孫度支員外郎璟詢事。璟曰：「某祖芳，前從力士問舊總，未竟。復著唐歷，採建義類相近者以續之。其餘，或秘不敢宣，或奇怪，非編錄所宜及者，不以聞。」今按求其書，亡失不獲。臣德裕亡父先臣，與芳子吏部郎中冕，貞元初俱為尚書郎。後謫官，亦俱東出。道相與語，遂及高力士之說，且曰：「彼皆目睹，非出傳聞，信而有徵，可為實錄。」先臣每為臣言之。臣伏念所愷悌，凡十有七事。歲祀久，遺藁不傳。臣德裕，非黃瓊之達練，能習故事；愧史遷之該博，唯次舊聞。霍失其傳，不足以對大君之問，謹錄如左，以備史官之闕云。</TEXT>
    </SECTION>
  + <SECTION>
  + <SECTION>
```

(圖 6) 典籍本文檔製成XML檔

在本研究設計之初，也成想過使用現有詞庫來進行後分析呈現，但一般詞庫大多是專業領域的專有術語或是通用分類的詞庫，例如：本國專利技術名詞中英對照詞庫系統[25]、精粹辭庫[34]、國家文化資料庫辭庫[32]等。我們嘗試以西遊記的主角「孫悟空」來測試個詞庫可用性，發現竟無一收藏，這迫使我們必須針對各典籍獨立建立分類詞庫。

我們預計針對每一部典籍決定有意義的詞庫類別，然後將維基百科中該部典籍該類詞庫目前的記載當成樣本，使用人工半自動詞夾子程式進行詞類萃取以擴充詞庫。例如針對封神演義中的人名詞庫，維基百科中記載 55 個人名，使用人工半自動詞夾子萃取程式後，新增了 344 個人名，其增加的幅度不難發現本研究對封神演義的人物研究有其貢獻。根據實驗，使用人工半自動詞夾子萃取程式處理一個分類詞庫平均約需 14 分鐘的機器時間及 35 分鐘的人工時間，平均新增 64

個詞。這表示建立典籍針對性分類詞庫並不是一件難事，如表 3 中所列，目前我們共處理 96 部典籍，139 個分類詞庫，計 21538 個詞。

最後將整理好的詞庫和 XML 交給搜尋引擎去處理，讓搜尋引擎取代過往以人工的方式整理資料，讓使用者可以在我們的搜尋引擎下關鍵字，而搜尋引擎會從您所下的詞對整個典籍做關聯分析或統計，再將結果呈現在我們搜尋引擎上，讓使用者能夠對中文典籍的研究能夠事半功倍，也能因此節省不少的時間，(表 3) 詞庫分類表為經由利用詞夾子程式所增加出來的詞庫。

(表 3)詞庫分類表

典籍	分類	樣本	增加
一枕奇	人物	0	32
二刻拍案驚奇	人物	0	216
八仙得道	人物	0	74
八仙得道	法術	0	20
八仙得道	地名	0	40
八仙得道	武器	0	11
三國演義	人物	1174	0
三國演義	衣物	0	102
三寶太監西洋記	國家	0	28
三寶太監西洋記	人物	0	47
三寶太監西洋記	地名	0	74
于公案	人物	0	106
于公案	地名	0	44
大明奇俠傳	人物	0	79
大明奇俠傳	地名	0	34
女媧石	人物	0	26
子不語	人物	0	320
子不語	地名	0	117
子與續子不語	人物	0	415
子與續子不語	地名	0	23
五色石	人物	0	79
五虎平南	人物	0	84
五虎全集	人物	0	174
五虎征西	人物	0	90
文明小史	人物	0	118
文明小史	地名	0	53
木蘭奇女傳	人物	0	127
毛公案	人物	0	18

典籍	分類	樣本	增加
青箱雜記	人物	0	37
宦海升沉錄	人物	0	122
宦海升沉錄	地名	0	50
封神演義	人物	55	344
封神演義全	人物	0	414
封神演義-改	神祇	0	107
封神演義-改	人物	0	74
後西遊記	人物	0	104
後與西遊記與補	人物	0	426
後與西遊記與補	地名	0	186
後與西遊記與補	武器	0	66
後與西遊記與補	法術	0	54
後與西遊記與補	寶物	0	19
恨海	人物	0	22
施公案	人物	0	422
紅樓全集	人物	0	726
紅樓真夢	人物	0	726
紅樓復夢	人物	0	726
紅樓圓夢	人物	0	726
紅樓夢	人物	22	716
紅樓夢	地名	0	149
紅樓夢	衣服	0	199
紅樓夢	飲食	0	157
紅樓夢補	人物	0	726
紅樓夢影	人物	0	726
郁離子	人物	0	146
飛跎全傳	人物	0	24
徐霞客遊記	地名	40	0

水滸全集	人物	0	354
水滸全傳	人物	0	207
水滸後傳	人物	0	120
水滸傳	人物	108	121
水滸傳	女姓	0	65
包公全集	人物	0	358
包公案-五鼠鬧東京	人物	0	44
包公案-百家公案	人物	0	99
包公案-龍圖公案	人物	0	215
北遊記	人物	0	92
北遊記	地名	0	21
北遊記	武器	0	12
玉照新志	人物	0	17
老殘遊記	人物	0	60
老殘遊記	女姓	0	26
老殘遊記	地名	0	23
老殘遊記二編	人物	0	50
老殘遊記與二篇	人物	0	103
老殘遊記與二篇	地名	0	23
西遊記	人物	5	316
西遊記	神祇	0	178
西遊記	法術	0	54
西遊記	地名	0	172
西遊記	武器	0	19
西遊補	地名	0	18
西遊補	人物	0	74
西遊補	武器	0	16
李娃傳	人物	4	19
奉天錄	人物	0	24
官場現形記	人物	0	151
拍案驚奇全	人物	0	466
東周列國志	國家	0	35
東周列國志	人物	0	154
東周列國志	年號	0	108
東遊記	人物	8	131
東與北遊記	人物	0	209

浮生六記	人物	0	26
鬼谷四友志	國家	0	11
鬼谷四友志	人物	0	76
鬼谷四友志	地名	0	30
聊齋誌異	人物	0	376
聊齋誌異	神怪	0	95
雲仙笑	人物	0	41
順宗實錄	人物	0	15
搜神記	動物	0	65
搜神記	人物	0	253
搜神記	地名	0	96
新石頭記	人物	0	726
楊乃武與小白菜	人物	0	86
楊家全集	人物	0	61
楊家府世代忠勇通俗演義	人物	0	64
楊家將	人物	29	96
痴人說夢記	人物	0	48
蜀碧	人物	3	25
綠野仙蹤	人物	0	60
劉公案全	人物	0	243
劉公案-青龍傳	人物	0	36
劉公案-滿漢門	人物	0	45
劉公案-劉墉傳奇	人物	0	34
劉公案-雙龍傳	地名	0	45
劉公案-雙龍傳	人物	0	43
劉公案-羅鍋軼事	人物	0	107
儒林外史	女姓	0	66
儒林外史	人物	0	32
蕉葉帕	人物	0	20
諧鐸	人物	0	147
諧鐸	地名	0	56
濟公全集	人物	0	271
濟公全傳	人物	0	97
濟公活佛傳奇錄	人物	0	173
魏鄭公諫錄	人物	0	21
孽海花	人物	0	163

東與北遊記	地名	**0**	**10**
東與北遊記	武器	**0**	**12**
東觀奏記	人物	**0**	**17**
林間錄	人物	**0**	**43**
河東記	人物	**0**	**71**
初刻拍案驚奇	人物	**0**	**255**

罌粟花	人物	**0**	**74**
罌粟花	國家	**0**	**20**
罌粟花	地名	**0**	**16**
續子不語	人物	**0**	**101**
續紅樓夢	人物	**0**	**726**

5. 結論

本文為建立中文典及研究環境，選用432部無版權爭議的中文典籍進行數位化加工，建置搜尋引擎，並使用詞夾子萃取各部典籍中中文領域學者有興趣的詞類，目前計共處理96部典籍，139個分類詞庫，計21538個詞，並且陸續增加中，期望此系統能對中文領域學者有實質上的貢獻。

6. 參考文獻

[1]. 王心劍：《西遊記》人物設計上看主旨，西安石油學院學報(社會科學版)，民國 91 年 1 月。

[2]. 王盈方:《紅樓夢》十二釵命運觀之研究，國立台灣師範大學碩士論文，民國 84 年。

[3]. 吳麗卿: 《紅樓夢》的女性認同，東海大學碩士論文，民國 94 年。

[4]. 宋孟貞: 《紅樓夢》與《鏡花緣》的才女意義析論，暨南國際大學碩士論文，民國 88 年。

[5]. 李文瑤:《水滸傳》女性研究，國立彰化師範大學碩士論文，民國 93 年。

[6]. 李貴禎: 論《儒林外史》女性形象塑造，國立屏東教育大學碩士論文，民國 95 年。

[7]. 汪玉玫: 《紅樓夢》中賈府女性人物論，東海大學碩士論文，民國 91 年

[8]. 周正娟: 聊齋誌異婦女形象研究，東海大學碩士論文，民國 83 年。

[9]. 林允:《聊齋誌異》鬼狐仙妖研究，國立彰化師範大學碩士論文，民國 93 年。

[10]. 胡玉珍：西遊記中的精怪與神仙研究，南華大學文學研究碩士論文，民國 92 年 6 月。

[11]. 張尚斌：詞夾子演算法在專有名詞辨識上的應用-以歷史文件為主，臺灣大學碩士論文，民國 94 年。

[12]. 張嘉惠:《聊齋誌異》女妖故事研究國立中山大學碩士論文，民國 90 年。

[13]. 彭毓淇: 丫鬟與小姐之互動關係研究----以《紅樓夢》為主的論述，國立清華大學碩士論文，民國 92 年。

[14]. 黃聿寧:《水滸傳》中的女性及其影響，國立中山大學碩士論文，民國 95 年。

[15]. 楊平平: 父權社會下的女兒國-《紅樓夢》女性研究國立彰化師範大學碩士論文，民國 94 年。

[16]. 賴采蘋: 《搜神記》中的動物類型研究—以動物與人類的關係為中心，國立中正大學碩士論文，民國 92 年。

[17]. 鍾明玉:《紅樓夢》飲食情境研究，國立清華大學碩士論文，民國 81 年。

[18]. 韓秀利：金聖嘆《水滸傳》人物探討，醒吾學報，民國 93 年 1 月。

[19]. 蘇惠玲:《三國演義》中男子服飾的角色刻劃效應—以曹操、關羽、諸葛亮為中心的比較研究，佛光人文社會學院碩士論文，民國 92 年。

[20]. David Nadeau, Satoshi Sekine：A survey of named entity recognition and classification，National Research Council Canada / New York University

[21]. 《紅樓夢系列》：life.fhl.net/Literature/culture/hongindex.htm

[22]. 三國演義網上版 Romance of the Three Kingdoms Online ：3kingdoms.globalsoho.com/

[23]. 中文古籍列表：http://books.arping.idv.tw/books.html

[24]. 中華典籍網路資料中心－紅樓夢網路教學研究資料中心：cls.hs.yzu.edu.tw/HLM/home.htm

[25]. 本國專利技術名詞中英對照詞庫系統：http://paterm.tipo.gov.tw/IPOTechTerm/searchInput.jsp

[26]. 古騰堡計畫首頁：www.gutenberg.org/wiki/Main_Page

[27]. 全國博碩士論文資訊：etds.ncl.edu.tw/theabs/index.jsp

[28]. 紅楼梦网：www.52hongloumeng.com/

[29]. 紅楼梦谭：www.honglm.net/

[30]. 紅樓夢網路教學研究資料中心：cls.hs.yzu.edu.tw/HLM/read/text/TEXT.ASP

[31]. 相片搜尋網站：www.flickr.com/

[32]. 國家文化資料庫辭庫：http://nrch.cca.gov.tw/ccahome/word/

[33]. 開放文學網：open-lit.com/

[34]. 精粹辭庫：http://www.pristine.com.tw/lexicon.php?lang=zh-tw

[35]. Dojo 網頁技術官網：dojotoolkit.org/3

[36]. Tag Cloud： http://en.wikipedia.org/wiki/Tag_Cloud