網路中文文件自動摘要

黃純敏
資訊管理系副教授
國立雲林科技大學
斗六/台灣
huangcm@mis.yuntech.edu.tw

吳郁瑩
資訊管理系碩士班研究生
國立雲林科技大學
斗六/台灣
wuyy@tomail.com.tw

摘要

傳統搜尋引擎自動摘要設計方式多半截取網頁的前幾十個字元，作爲輔助性說明。惟觀其所截取的文句，多無什意義，非但無法提供充足的內文判斷資訊，更甚者，或可能誤導了使用者。本研究剖析網路文件標記特性及中文字詞詞性，研發跨主題的自動摘要系統。經使用者實際測試後，在網頁內容判斷、摘要可讀性，以及選用意願的評估項目，優於目前一般搜尋引擎的自動摘要設計。經交叉比對分析，發現年齡在 20 歲(含)以上者，以及學歷愈高者（研究所以上），有極顯著偏好本研究的自動摘要之傾向。是否透露不同年齡層與學歷程度對摘要有不同的需求，則有賴未來繼續研究。

關鍵詞：自動摘要、中文斷詞、網路超文件、資訊檢索、搜尋引擎

一、前言

傳統上，一般學術性質的期刊，多半會提供簡短的摘要，作爲輔助使用者瞭解論文大意或判斷是否進一步閱讀全文的參考。近年來由於全球資訊網(World Wide Web)的普及，帶動許多上網的人口，也改變了一般人的閱讀及寫作習慣，這使得各式各樣電子型式的資料大量出現，網際網路儼然成爲前所未有的知識大寶庫。爲協助使用者找尋資料，各種搜尋引擎(search engine)也積極扮演著導路的角色。然而，當使用者下達一個搜尋指令時，搜尋引擎動輒回報數百筆，甚至數千筆以上的資訊。五彩繽紛的網頁內文，原多未附含摘要敘述，有些搜尋引擎雖然已有自動摘要之設計，惟其設計方式多半截取網頁的前幾十個字元，作爲輔助性說明，由於網路文件有其特殊的寫法，觀其所截取的文句，多無什意義，非但無法提供充足的內文判斷資訊，更甚者，或可能誤導了使用者。

文件自動摘要雖已是自然語言處理（Natural Language Processing, NLP）的重要標的之一。然而對於所擷取語句的可讀性、前後文句的連慣性，卻一直難有重大突破。對於網路文件自動摘要的研究，並未見著墨。即是目前風行的搜尋引擎，也僅著重於搜尋機制的改良，對於搜尋結果的自動摘要呈現，多僅視爲可有可無的附屬功能。使用者對於查詢結果仍需逐筆連結進入各網頁瀏覽。在頻寬有限的網路環境裡，超連結使用之頻繁，對於網路傳輸之負擔不啻雪上加霜。因此引發我們開發可讀性網路文件摘要系統的動機，希藉此配合

搜尋引擎的檢索結果，自動展現足以表達網頁內文的簡短摘要，一則節省使用者逐筆進入網頁瀏覽的時間花費，更可減少大量不必要的資訊傳遞，提升網路傳輸效率。對二十一世紀資訊社會的發展，應有正面的助益。

二、研究主題

基於資訊化時代對於網路文件閱讀習慣日益形成之需求，本研究希望藉由傳統自動摘要製作技術，研究其實施於中文及網路超文件的可行性。主要的研究議題有三項：

（一）西文自動摘要的方法於中文文件之適用性。

（二）傳統文件自動摘要方法於網路超文件之適用性。

（三）網路超文件自動摘要之實用性。

三、文獻探討

3.1 自動摘要定義

自動摘要是指將萃取組合文章內文重要字句的過程予以自動化之謂。摘要的目的在於產生一個言簡意垓的文件描述，它應比文件標題更具敘述性，但又短的可讓人一眼就明瞭。因此，一個好的摘要應該能夠有效反應原文所要傳述的重要意旨。至於摘要的長度Lancaster(1991)則認爲沒有明確的限制。因爲其中牽涉到的因素頗多，包括：原始文件的長度、原始文件所表達主題的複雜度及多樣性、原始文件對不同組織不同個人的重要性、原始文件獲取的容易度、成本以及目的等等。依此，自動摘要之精神，除節省時間人力外，其結果應仍不脫『言簡』『意垓』之效。

3.2 自動摘要相關研究

語言文字是人類社會中表達意念最主要的工具。對一篇文件而言，其撰寫內容往往是針對某一特定主題，所以文中會存在許多足以代表該主題的關鍵詞彙，這些關鍵詞彙可視爲該文件的重要資訊。因此，關於自動摘要的研究，也多是基於上述理念，進行萃取原始文件中重要句子的過程(Luhn, 1958; Edmundson, 1964)。一般認爲評估句子的重要性可考慮：字彙在文中所出現的頻率、關鍵詞彙出現的位置、與上下文的關係等(Edmundson, 1969; Rush et al., 1971)。是以多年來學者的研究，多著重於向量統計或語意分析的技巧，藉以萃取文中具有代表性的句子（Paice, 1990; Salton, 1983, 1989, 1996）。近年來因爲全球資訊網的興起，資訊檢索技術也已逐漸被應用到探討超連結(hyperlink)自動產生的適用性(Allan, 1996; Salton, 1997)。

3.3 中文資料處理

中文文字不若西文文字有明顯的空白可以將句子中的各個詞彙(term)區分開，因此，長久以來中文關鍵詞擷取一直是資訊檢索領域發展的瓶頸。而中文斷詞最大的問題就是斷詞組合的岐義性(ambiguity)、複合詞研究以及未知詞問題(Chiang, et.al.1992)。國內進行中文字詞分析，以陳克健、黃居仁爲首的中央研究院中文詞知識小組成效最著，其研究成果包括中文詞庫[八萬目詞]、平衡語料庫、中文語料庫、以及各種技術專書等。

已進行的中文字詞分析，可歸納爲三類：詞庫比對法、文法剖析法及統計法(Chen, et al., 1993; Ho, et al., 1993)。詞庫比對法主要利用現有詞庫，比對輸入的文件，擷取出文件中出現在詞庫的字詞。此種作法若有現成詞庫則操作簡單，更可依據詞性作爲關鍵詞篩選參考，惟對於新生詞彙則無法辨識。文法剖析法是利用自然語言處理技術及過濾技術，篩選出文件的關鍵詞彙組。惟中文語法變化過多且標準制定不易，使得此種作法處理甚爲困難。至於統計法需透過大量文件的分析，取得足夠的統計參數（詞頻、門檻值）後，再擷取滿足參數的語彙。此種作法可有效擷取新字詞，惟所產生之詞彙甚多，並無法滿足需詞性對照分析者。

3.4 網路超文件處理

網際網路上所流通的文件，其上所使用的語言爲超文件標記語言(Hyper Text Markup Language, HTML)。它是一種標示網路文件格式的標準語言。基本上 HTML 的文件不過是一般普通的文字檔，再加上一些標記，用以展現有關字體字形的變化、圖片的設置或是一些超連結。當瀏覽器經由網路接收到 HTML 文件後，不但會將文字與圖檔資料顯現，同時也會依照標籤，將內容以適當的方式呈現。全球資訊網的魅力在於，所有的網頁都是遵守共同的 HTML 標準，使得多采多姿的網頁能在各種平台呈現一致的特色。當全球資訊網躍居爲資訊傳遞的主要舞台時，網路文件也自然成指數等級的增長。近年來應用資訊檢索的技術，建置了爲數可觀的搜尋引擎，網路文件處理成爲新興的研究課題(Sonnenreich and Macinta, 1997)。搜尋引擎雖然積極扮演著導路的角色，然而，當使用者下達一個搜尋指令時，搜尋引擎動輒回報數百筆，甚至數千筆以上的資訊。五彩繽紛的網頁內文，原多未附含摘要敘述，有些搜尋引擎雖然已有自動摘要之設計，惟其設計方式多半截取網頁的前幾十個字元，作爲輔助性說明，由於網路文件有其特殊的寫法，觀其所截取的文句，多無什意義，非但無法提供充足的內文判斷資訊，更甚者，或可能誤導了使用者。若分析 HTML 結構，其中不乏與本文主題相關者如：<TITLE>, <META>, <H?>等，如以將這些標籤作爲判斷句子重要與否的線索，依標籤的重要程度，給予不同的比重，應是不錯的嘗試。

3.5 摘要評估

摘要的評估是一件困難且主觀的工作。對於自動摘要的評估，學者多從系統研發成本與成果效益雙方面進行分析。在成果效益上，多半針對自動摘要的可讀性，要求使用者提供意見。學者的研究指出，以使用者直覺式(intuitive) 的評估方式雖然簡單，但其不一致、主觀與非量化是其缺點(Brandow, et al.1995)。Edmundson(1969), Salton, et al.(1997)所進行的自動摘要評估，使用者的反應也都是評估的重要指標。以使用者進行評估，無法避免的必然會參雜人爲主觀因素於其中，因使用者背景及需求的不同，所做出來的評估，也未必能真正的正確。然而，至今似乎仍無法有一個正確而又客觀的自動摘要評估法。

四、研究架構

自動摘要的相關研究在西方已行之有年，國內在這幾年才開始投入這方面的研究，而網路超文件自動摘要的研究則屬一較新的範疇。圖一(見附錄)為本研究系統完整架構圖。系統是在 Microsoft Windows NT 4.0 Server 平台環境下，使用 PC Pentium 等級的機器，配備 AMD K6-2-233 的 CPU、64MB 記憶體，以及 10.1GB 的硬碟，採用 Visual C++ 6.0 程式語言發展。研究中用來產生自動摘要的樣本，係由人工隨意選取網路上不限類別之超文件，共計收錄有醫療、旅遊、圖書館資訊、同志議題、生命教育、女權主義、音樂、電腦、教育改革等十多個類別，字數在一千餘字到萬餘字不等的超文件，共計 300 篇。摘要呈現的方式，配合搜尋引擎的查詢結果一併呈現。

在圖一中包含兩個子系統：搜尋引擎子系統與自動摘要子系統。自動摘要子系統為本論文研究標的，希望藉由自動摘要的呈現，增進使用者在搜尋引擎的查詢效益。圖二為自動摘要子系統的處理流程圖。在自動摘要子系統中，由於網路超文件格式不同於一般的文件資料，加上中文文字不若西文文字有明顯的空白區隔，所以在實際進行自動摘要處理流程之前，需先經過文件標籤剖析及斷詞處理。

4.1 超文件標籤分析子系統

超文件標籤分析子系統的目的為去除超文件中多餘的標籤及符號。本系統初期僅以中文字為實作對象，因此英文字亦在排除之列。由於超文件標籤格式種類繁多且複雜，需輔以標籤線索資料檔，以記載標籤所具有的特殊意義，用來當作辨識重要句子之線索。除此，對於文件中重要句子出現的位置亦予記錄，例如：出現在第幾段的第幾句。在最後重組句子時，上述記錄均用得到。圖三為此子系統的架構圖。

4.2 中文斷詞子系統

中文文件在字與字之間，不像西方文字有明顯的空白可以區分，故在中文文件的資訊處理上，斷詞是一道不可避免的程序。圖四為斷詞子系統架構圖。

本研究所採用的詞庫為，中央研究院所建構的八萬目詞庫。該詞庫共收錄有 78410 目詞。一個句子裡，動詞與名詞通常是句子的核心；在自動摘要文獻探討中，亦不乏採用名詞與動詞當作重要詞彙的例子(Barzilay，1997)。因此在本研究中，僅將句子中的名詞與動詞，視為與內文最相關之重要詞彙。但有些不重要的名詞亦予以惕除，如：定詞 (一些，諸多，許多….)、量詞 (一幫人，一堆雪，一筆…)、方位詞 (以外，以上，當中，方面…)、代名詞 (我們，妳，汝輩，吾人…)、姓氏 (吳，張，諸葛…)等。故在進行斷詞之前，需先進行詞庫的過濾，作為斷詞子系統比對時的參考。經過此一處理程序，詞庫原有 78410 個詞，篩選後剩餘 46243 個詞。

過濾詞後，接著建立詞庫雜湊表(Hash Table)以加快詞庫比對速度。若數個詞彙有著相同的首字，便使用鏈結串列(Link List)儲存，以減少記憶體的浪費。在斷詞方面，本研究採用詞庫比對結合長詞優先法來進行斷詞作業，如比對不到則視爲未知詞，不加以處理與記錄。在斷詞的過程中，仍需計算文件中每個詞彙出現的次數，以及記載詞彙所出現的位置，以作爲自動摘要子系統計算句子重要性權值之依據。

4.3 自動摘要子系統

由於網路上的文件數量龐雜且增加快速，系統運作方式應考慮到即時性，本研究乃採統計方式配合上述超文件標籤線索檔，作爲自動摘要產生的方法。圖五爲自動摘要子系統架構圖。

在自動摘要子系統中，依序可分爲六個步驟進行。

1. 計算重要詞彙得分：

評估準則有四

(1)頻率關鍵詞法

名詞與動詞是一個句子的核心，因此，在這個部份，文件中每一個名詞與動詞在本研究中皆視爲重要詞彙，而詞彙的重要程度，則視該詞彙在文件中所發生次數多寡。

(2)標題關鍵詞法

網路超文件，呈現標題的方式，可能藉由 <TITLE>、<H?>，以及 <FONT SIZE> 等三種方式。<TITLE> 是使用在超文件的最開頭，主要在說明文件的主題。本研究假設每一篇超文件都能有一個良好定義的主題，因此給予在主題中的重要詞彙，權重爲 5。<H?>標籤有六種等級，從 <H1> ～ <H6> ，<H?> 標籤使用在文件內部，藉以區分文件中的大小標題，由於 <H5> ～ <H6>字體大小並沒有特殊之處，也較少有人使用來當做標題，本研究不予處理。本研究給予所有 <H1>～<H4>標籤相等的權重 3。有些超文件使用 <FONT SIZE> 來呈現其主題，其範圍由 1～7，預設值爲 3，也就是一般呈現的字大小是 3，所以本研究大膽假設 <FONT SIZE> 之值大於 3 者爲具有標題的作用，給予權重 3；而<FONT SIZE> 小於等於 3 者，表示比較不重要的資訊，無需做特殊處理。

(3)位置法

一般超文件中，多半是以<P>標籤來區分句子段落。根據學者評估，Mead 資料中心的自動摘要系統 Searchable Lead，只是簡單的摘錄文件中的前 60、150 或 250 個英文詞彙，便達到了 90% 以上的可接受度 (Brandow, et al.，1995)。故在本研究中給予文件的第一段 10 的權重。

(4)標籤線索法

超文件提供了某些特殊標籤，用以呈現其他重要的資訊。<META>標籤能記錄超文件一些額外的資訊，例如：作者所給予文件的關鍵字，故本研究對於出現在<META>標籤中的詞彙給予權重 5。其他相關標籤線索如：<B>、<I>、<EM>、<STRONG>、<BLINK>、<BIG>等所加強表示的詞彙，多半是作者認爲比較重要的詞彙，故給予其權重 2。

2. 計算句子的得分

$$SCORE_{S_{ij}} = \sum_{k=1}^{n} TP_k + PW_{S_{ij}} + \sum_{l=1}^{m} T_l W_t$$

經由第一個步驟的分析計算後，句子的得分可以很輕易的藉由句子中重要詞彙權重的加總而來。句子得分計算公式如下：

上述 S_{ij} 表示文件中第 i 個句子的第 j 個子句，TP_k 代表句子中第 k 個詞彙的重要性分數，n 是指 S_{ij} 子句中重要詞彙的總數，PW_{Sij} 表示第 S_{ij} 個子句的位置權重，T_iW_t 爲詞彙 T_i 的標籤 t 權重 (W_t)，m 表示 Sij 子句中加權詞彙的總數，最後 $SCORE_{Sij}$ 即爲 S_{ij} 子句的總得分。

3. 根據得分數將句子排序

第三個步驟則是將文件中所有句子，依得分高低降冪排序。

4. 根據擷取原則摘錄句子

由文獻得知，摘要的字數長度並無一定的標準。本研究採資訊科學大辭典中對附錄及簡訊性質之資料摘要字數的建議，決定摘錄的總長度爲 125 個字(250 字元)左右。爲顧及句子的完整性，所摘錄的句子將以完整句子爲擷取依據。

5. 按文件順序排列句子

最後一個步驟，是將摘錄出的句子，按照文件原本的順序組合，使成爲一篇可讀性的摘要文件。並將製作出來的摘要與原來的超文件資料之間，建立鏈結關係，便可提供給搜尋引擎，輔助查詢結果的呈現。

6. 摘要結果呈現

完成了上述自動摘要的製作後，所產生的摘要便可與搜尋引擎結合，作爲搜尋引擎查詢結果回報時的提示訊息。以下列示一些本研究自動摘要與傳統搜尋引擎自動摘要，針對同一篇文章所產生的不同摘要結果。

文件主題：	藥品在人體內的旅行
本研究自動摘要：	經口服的藥品進入全身血液循環以前，會先到肝臟旅行，有些藥就在此地被肝臟的酵素破壞了一大半，藥學上稱爲「肝臟首渡效應」，會使藥品的療效打折扣。研究藥品在人體內的旅行過程的學問叫做藥品動力學，藥品動力學可以幫助我們探討藥品在體內的行徑與人體處理藥品的經過，包括吸收、分佈、代謝、排出等。
傳統搜尋引擎自動摘要：	[生活用藥常識]藥品在人體內的旅行本文作者：和信(原孫逸仙)醫院藥劑科主任/陳昭姿　我們吃下去的食物，.....

文件主題：	甘蔗
本研究自動摘要：	甘蔗爲禾本科植物，甘蔗 SaccharumsinensisRoxB，拉丁學名「Saccharum」是指糖或甜的意思，「Sinensis」即中國，指甘蔗產於中國。稓含作干蔗，謂其莖如竹竿也，這就是甘蔗的名源。甘蔗渣中含有對於小鼠艾氏癌和腫瘤-180 有抑制作用的多糖類等藥理作用。
傳統搜尋引擎自動摘要：	[蔬果養生]甘蔗本文作者：中國醫藥學院中國藥學研究所教授-邱年永 &.....

文件主題：	工商時報新聞
本研究自動摘要：	【記者梁玉立台北報導】爲加速金融改革，統合金融、證券、保險事業的監理事權，行政院長蕭萬長今天將宣佈一重大訊息，將在行政院之下，成立獨立的「金融監理委員會」或「金融總署」，該單位主要將由現行央行金檢局、財政部金融局第六組、財政部證期會等相關單位組成，將直屬於行政院，至於該單位首長是由財政部長兼任，或是另外有專職首長，則將另作討論。
傳統搜尋引擎自動摘要：	中時電子報｜ 中國時報｜ 工商時報｜ 中時晚報｜ 新聞專輯｜ 新聞檢索｜ 即時新聞｜ 新聞攝影 工商時報 焦.....

五、研究成果與未來發展

本研究以網際網路中文超文件爲摘要主體，經過超文件剖析系統、中文斷詞系統的分析處理，最後運用統計方式計算，擷取出文件的摘要字數。爲了驗證本研究自動摘要的適用性，採取了線上問卷的評估方式，針對兩種不同摘要（一般搜尋引擎自動摘要，本研究自動摘要），評估網頁內容容易判斷程度、可讀性、字數適當性，以及願意選擇之摘要等項目。線上問卷採登載各校電子佈告系統(BBS) 方式，由受測者主動填寫問卷。經過十天的開放時間，共計有 194 位受測者上線填寫問卷，其中一份爲無效問卷。問卷資料分三部份做分析：一爲受測者基本資料、二爲受測者選填各項評比項目的百分比，最後爲各基本資料與摘要評估項目的交叉分析。測驗使用者滿意度的評估方式。評估結果，在摘要判斷網頁內容的容易程度，及摘要可讀性上，獲得 60%以上的認同；在最後摘要的選擇上獲得 54%的認同，高於選擇傳統搜尋引擎摘要的 29%。這樣的結果，說明本研究在系統適用性上實優於目前一般搜尋引擎。惟經交叉比對基本資料與摘要評估項目發現，年齡與學歷對摘要各項評估中，有顯著的差異 (顯著值 $P < 0.05$)。尤其年齡在 20 歲(含)以上者，以及學歷愈高者（研究所以上），在所有評估選項中，都極顯著偏好本研究的自動摘要。是否透露不同年齡層與學歷程度對摘要有不同的需求，則有賴未來繼續研究。

總括而言，本研究有下列幾項優點：1. 較容易判斷網頁內容，2. 可讀性較佳，3. 自動摘要產生之速度快（文章字數在二千字以內，產生速度約 2 秒），4. 節省製作成本，5. 可應用於各種不同類別文件。惟在發展上爲判斷詞性，仍受限於現有詞庫，以致新詞仍無法納入，如此一來，極有可能忽略了文件中有用的資訊。此外，本研究純粹以統計方式擷取句子，在句子意義考量上，難免有所缺失，若能在統計方法之外，輔以人工智慧的方法，或改用其他相關演算法，或許能夠找出文章中真正的重要句子，更是未來值得探討的的方向。

參考文獻

[1] Allan, J. (1996). Automatic Hypertext Link Typing. *Hypertext '96, The Seventh ACM Conference on Hypertext* (pp. 42-52). New York: Association for Computing Machinery.

[2] Barzilay, Regina & Elhadad, Michael. (1997). Using Lexical Chains for Text Summarization. available at http://www.cs.bgu.ac.il/summarization-test.

[3] Brandow, R., Mitze, K. & Rau, L. F.，(1995). Automatic Condensation of Electronic Publications by Sentence Selection，*Information Processing & Management* 31(5)，pp.675-685.

[4] Chen, et al., (1993). Some Distributional Properties of Mandarin Chinese – a Study based on the Academia Sinica Corpus. In *Proceedings of the First Pacific Asia Conference on Formal & Computational Linguistics.* , pp. 81-95.

[5] Chiang, et al., (1992). Statistical Models for Word Segmentation and Unknown Word Resolution. In *Proceedings of COLING V 92*, pp. 123-146.

[6] Edmundson, H. P. (1964). Problems in Automatic Abstracting. *Communications of the ACM* .7(4), pp.259-263.

[7] Edmundson, H.P. (1969). New Method in Automatic Extracting. *Journal of the Association for Computing Machinery.* 16, pp.264-289.

[8] Ho, et al. (1993). Using Syntactic Markers and Semantic Frame Knowledge Representation in Automated Chinese Text Abstraction. In *Proceedings of the First Pacific Asia Conference on Formal & Computational Linguistics.* , pp. 122-131.

[9] Lancaster, F. W. (1991). *Indexing And Abstracting In Theory And Practice.* Ann Arbor: Gushing-Malloy Inc.

[10] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development.* 2(2), pp.159-165.

[11] Paice, C.D. (1990). Constructing Literature Abstracts by Computer：Techniques and Prospects., *Information Processing & Management* 26(1),171-186.

[12] Rush, J.E., Salvador, R. & Zamora, A. (1971). Automatic abstracting and indexing. II. Production of indicative

abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science.* 22(4), pp.260-274.

[13] Salton, G & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*，New York：McGraw-Hill, inc.

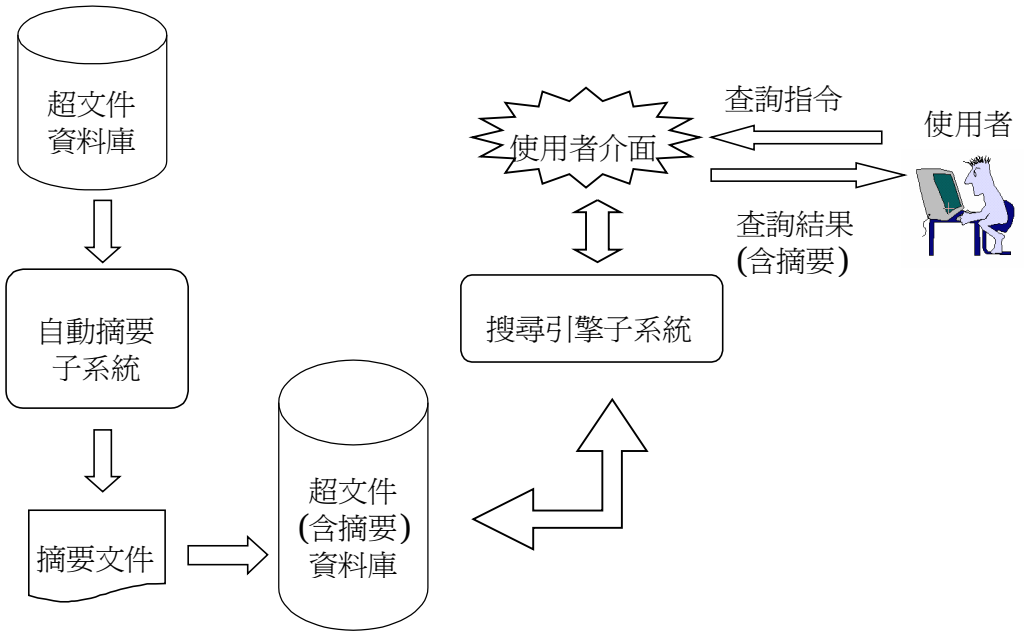
[14] Salton, G. (1989). *Automatic Text Processing-the Transformation, Analysis and Retrieval of Information by Computer.* Reading, MA: Addison-Wesley Publishing Co.

[15] Salton, G., Allan, J., Singhal, A. (1996). Automatic Text Decomposition and Structuring. *Information Processing & Management*，32(2),127-138.

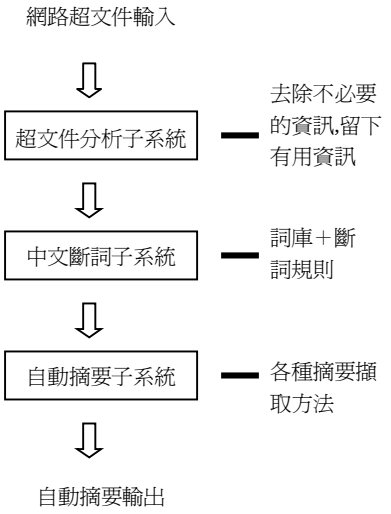
[16] Salton, G. et.al. (1997). Automatic Text Structuring and Summarization. *Information Processing & Management*，33(2),193-207.

[17] Sonnenreich, Wes and Macinta, Tim (1997). *Web Developer.com Guide to Search Engines.* New York: John Wiley & Sons, Inc.

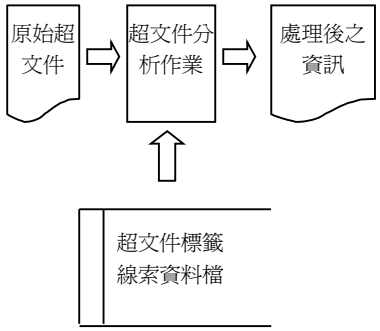
附錄:



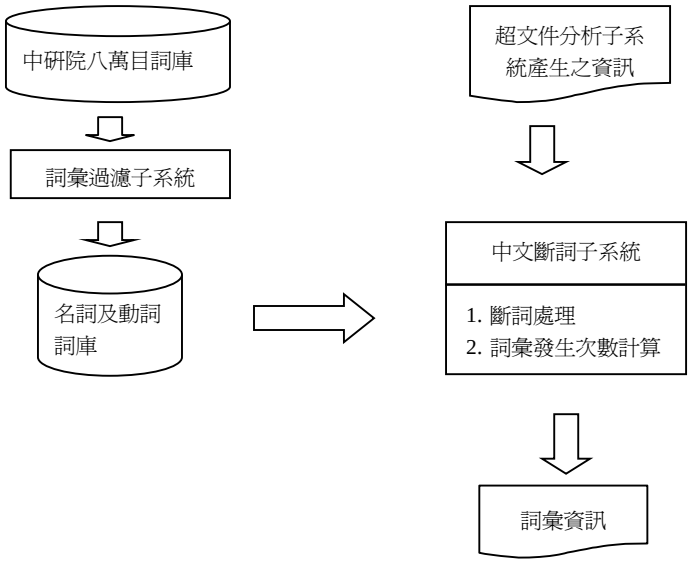
圖一：系統完整架構圖



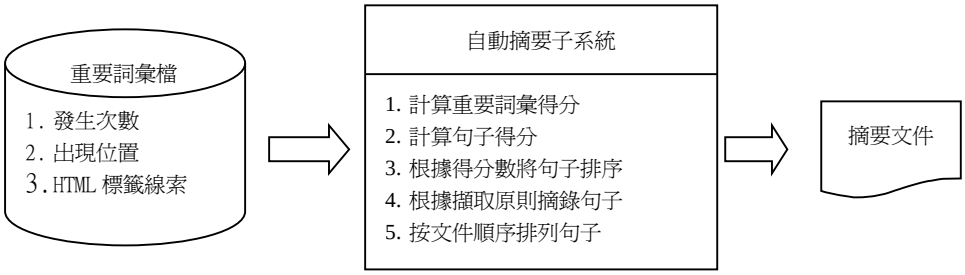
圖二：自動摘要架構圖



圖三：超文件標籤分析子系統架構圖



圖四：中文斷詞子系統架構圖



圖五：自動摘要子系統架構圖