網際網路服務的過去現在與未來

陳光華 教授
臺灣大學圖書資訊學系

壹、網際網路服務

網際網路是以 TCP / IP 的通訊協定所運作的網路，Internet的演進歷史是從1969年的 ARPANT 開始，1981 年的 BITNET (Because Its Time Network)，1983 的 MILNET，ARPANET uses TCP/IP，1986 的 NSFNET。1987 年臺灣連上了 BITNET，1991年12月經由 TANet 連上Internet。由1996年Internet全球分佈狀況來看，沒有連線的地區都是些較爲特殊的國家如緬甸、北韓等，其主要原因是經濟較落後國家，或是極權統治國家即使有能力也不願開放網際網路。

在沒有Internet時代，圖書館工作同仁與從事圖書館學研究的專家學者，早就提供資訊服務，早期的資訊服務系統或資訊檢索系統是在獨立而封閉的環境下運作。今天透過網路連接的世界，使得資訊傳遞非常迅速，雖然環境變得非常複雜且改變非常快速，但協助讀者取得有用、適用資訊的目標卻沒有改變。

從工程的觀點看資訊服務，會牽涉到以下幾個相關的技術，(1) Keyword searching；(2) Information retrieval (Document retrieval)；(3) Information filtering，使用者若對某些資訊感興趣，可經由智慧代理人或智慧型秘書過濾資訊；(4) Information extraction，從文件中將使用者想要的資訊直接檢索出來； (5) Information summarization，爲資訊的摘要； (6) Information understanding；(7) Question and answering。網際網路資訊服務的類型大致可分爲：原生性服務、加值性服務、訊息性服務、知識性服務。

貳、原生性服務

網際網路的基本功能，例如遠端登錄(Telnet)、電子郵件 (e-mail)、新聞討論群 (Usenet)、Gopher、全球資訊網 (WWW) 等，即是原生性服務。經臺灣蕃薯藤所做的調查， 1998 年臺灣網際網路的使用以 WWW 69.2% 高居首位，第二名是 e-mail

；1999 年 WWW、e-mail仍維持相當高的使用率，其他的服務使用率就減低。

原生性服務以 WWW最爲重要，因爲在WWW 的機制之下可建立很多加值性服務。Tim Berners-Lee 爲 WWW伺服系統的創始者，它最特別之處在於串連(Hyperlink)的功能，以及可以處理文字、圖像、影像、聲音等不同媒體資源。最重要的里程碑則是1993年推出的 Mosaic瀏覽器，短短幾個月時間讓 WWW在網際網路上有驚人的成長率，Mosaic瀏覽器即爲現今的Netscape。WWW Server 的成長狀況，1993 年全世界只有 50 Web servers，如今網路上 Web servers 量難以估計，保守的估計每年的成長率爲 3000% (資料來源爲 http://www/carloga.com/bgttwww.html)。就蕃薯藤1998年及1999年臺灣地區瀏覽器的佔有率調查來看，IE還是佔有較大比率。

參、加值性服務

WWW開發了許多加值性服務，不管是商業市場或學術界，最受注目的服務爲搜索引擎，是目前市場上大家所追求的利基，可提供使用者取得特定事件的相關資訊；另一類則是主題指引，使用者可以取得相關主題的資訊，同時有階層式的架構；其他的服務如搜尋 e-mail 、 BigFoot 、 Fourll等找人的服務，均建構在WWW之上的服務。雖然搜索引擎與主題指引提供相當大的幫助，但是檢索出來的文件相當多，且無法立即判斷是否相關，使用者必須連結文件，閱讀之後才能知道是否適用，而文件的往來佔用頻寬，使得網路流量大增，使用者沒有享受到網路帶來的方便，反而產生更多的負擔。

肆、訊息性服務

在 NII、GII、NGI等口號震天價響的新資訊時代，享用更好的服務，並非是過分的要求，訊息性的服務於焉產生，其最主要的服務爲資訊擷取及自動摘要。

一、資訊擷取

資訊擷取不提供一堆可能相關的文件，而是希望提供足以回答問題的文件。其方法是透過某種預定的樣板(template)，由文件集合擷取適切的資訊。進行檢索時，資訊擷取提供一個類似表格的樣板，使用者將適當的資料填妥，透過樣板上網尋找文件。樣板是所謂屬性名與屬性值配對的集合，與詮釋資料 (metadata) 的格式具某種程度的相關性，但兩者從不同的領域出發，資訊擷取是屬於 Computer Science 領域，詮釋資料則是由 Library Science 開始，繼而有Computer Science 加入討論。

1. 資訊擷取的特殊性

樣板是針對某特定的事件做特定的樣版處理，各個特定領域的文件會使用不同樣板，是資訊擷取的特殊性。

2.樣板

資訊擷取是依據使用者的資訊需求自訂樣板（User-defined templates）。從詮釋資料研究的角度來看，機讀編目格式屬權威機構制訂的樣板，當我們爲描述組織某一特定的資訊或文件時，會設想在這特定領域的使用者會用那些檢索點，需要那些屬性管理，於是權威機構制訂某些特定樣板來管理描述這些文件。然而詮釋資料研究面臨一些問題，就是必須由人來做。爲了網際網路上的文件處於有序的狀況，資訊擷取則嘗試用自動化來處理網際網路中的文件。

3. 資訊擷取自動化

將樣板視爲 Metadata，以資訊擷取的技術自動填 Dublin Core 的欄位，如此可期待解決時間及人力的問題，使檢索服務變得更具功效。若以自動化的技術來處理，當樣板愈複雜其自動化就困難，當樣板愈簡單則自動化品質將變差，故樣板複雜度與自動化程度是呈反比。

4. 資訊擷取系統的功能模組

文件由書面語構成，故需應用自然語言處理技術，包括詞幹處理、斷詞處理、語彙分析、語法分析，語意分析、照應詞處理、領域知識等功能模組。

(1)詞幹處理模組：主要是針對印歐語系。以英文爲例，將詞彙中代表眞正意義的主幹找出，如：beautifully 、 beauty、Beautiful都表示 beaut，其相關的技術包括工程演算法及Two-level morphology。

(2)斷詞處理模組：著重於東方語系，如中、日、韓文。斷詞處理概念是認爲語義的基本單位是個詞而不是字，故要做文件的分析，就必須先做斷詞，目前的中文檢索系統，都沒有做斷詞處理是因爲難度太高。一般而言，在不考慮專有名詞下，其正確率達 95%，若一併考慮專有名詞則正確率降爲 70-60% ，所以不做斷詞，要如何反應檢索者眞正的意含？斷詞的作法是採用詞頻方式，比對文件中詞彙出現的頻率，較高者就是可能的詞彙。電腦沒有辦法判斷人名、地名、組織名，專有名詞一旦判斷錯誤後面就會跟著錯，人名可用一些方法來辨識，現在的標準作法是利用「中華民國納稅資料庫」人名資料去分析，正確率還不錯，難度最高的是組織名稱。目前從事中文斷詞研究的單位有：中央研究院詞庫小組、新竹科學工業園區的智原科技、還有清華大學張居正教授及臺大的陳進興教授都在發展類似的模組。

(3)語彙分析模組：是目前發展成熟的自然語言處理技術。資訊檢索時，使用者搜尋的多數爲名詞，所以只要針對文件中的名詞做索引即可，自動化詞類判定工作正確率達95%以上。

(4)語法分析模組：即分析句子的語法。語

言的歧異性非常高，美國曾做過研究統計，一個由十個字所組成的簡單句子可畫出一百個語法樹，這種技術實際應用有相當困難，只能在有限範圍內使用，所以有人提出妥協的方案：部分剖析。部分剖析並不嘗試畫出剖析樹，只是將句子切割，以判斷名詞片語或動詞片語，片語詞組中有一詞首，詞首是名詞，這詞組就是名詞片語，在訂定詞類後利用部分剖析的技術及語言知識的輔助，就可判斷何者爲主詞、動詞、受詞。

做資訊擷取是非常耗時費事，至目前爲止還有很多努力的空間。資訊擷取未來可能的研究課題：1. 動態樣板轉換爲靜態詮釋資料的研究，即將資訊擷取及詮釋資料兩者合而爲一；2. 發展快速正確的剖析演算法，即畫出語法樹；3. 言談分析，分析文件段落討論的單一主題，如此可檢索文件的段落主題；4. 透過資訊擷取的樣板檢索直接產生摘要。

二、自動摘要服務

搜尋引擎檢索產生許多Link，若每個Link下的敘述都是正確的摘要，那麼只要看摘要就可知是否爲所需，這是自動摘要非常重要的理念。目前使用自動摘要的系統，有國外製作的 MS Word 及 InXight (InXight.com)，國內製作的 Trans - EZ (Trans-EZ.com)，每天有固定新聞摘要，同一新聞事件不同報紙的摘要。

1. 摘要的功能

(1) 宣示功能 (Announcement) ：宣示原始文件的存在性。

(2) 篩檢功能 (Screening) ：判定原始文件的相關性。

(3) 取代功能(Substitution) ：取代原始文件。

(4) 回溯功能 (Retrospection) : 查詢原始文件。

2.摘要的類型

(1)指示性摘要

摘要是文件的精緻版，它以較少的文字表述原始文件所欲傳達的訊息，圖書館學與資訊科學大辭典認爲「研究報告及專論，摘要宜少於 250 字，附錄及簡訊性質之資料，以 100 字爲佳，至於社論或讀者來函只需幾要個句子即可，長篇論著如技術報告、學位論文，其摘要以一篇以內並且以 500 字爲限」。指示性摘要具有宣示功能、篩檢功能及回溯功能；資料性摘要有取代功能及回溯功能；評論性摘要有回溯功能；摘錄則要有宣示功能、篩檢功能、取代功能及回溯功能。

(2)自動摘要

自動摘要是以自動化的程序製作原始文件的精緻版，自動摘要有兩種作法，一是由文件中挑選適當的段落或句子構成摘要，亦即製作所謂的摘錄；

一是分析原始文件，抽取文件的概念表意，再進行摘要的產生，此法非常難以自動化來製作，目前現成系統大都是摘錄。

三、自動摘要模型

本人採取的作法是製作「摘錄」型摘要，直接由文件擷取重要的句子製作成摘要。模型的建構是基於以下的假設：一般組織完善、意念完整的文件，其名詞與名詞以及名詞與動詞的關係相當密切，且名詞與動詞共存於述語參數結構，即在一個句子中，將動詞視爲述語，參數則爲其他名詞，而名詞間的關係是建構於言談層次。詞彙的統計值是爲統計這兩種不同詞彙間關係的重要性。其特性包括詞彙的重要性、詞彙的重複性、詞彙的共現性以及詞彙的距離。詞彙的重要性，並非所有的詞彙一樣重要，將文件中的冠詞、副詞以及介系詞等詞彙刪除，仍然知道這份文件的梗概，這說明了上述詞彙不十分重要，名詞與動詞則十分重要，以 IDF代表詞彙對文件的重要程度，IDF(w)=log((P-O(w))/O(w))。詞彙的重複性，作者爲強調某一概念致使有些詞彙不斷出現，這些詞彙有機會成爲文章討論的主題，透過它作摘要統計是恰當的。詞彙的共現性，意念一致的文件資料，作者使用的詞彙必然趨向某一個語意範疇，從統計的觀點來看，該語意範疇的詞彙一起出現的機率比較大。詞彙的距離，詞彙的位置也很重要，基於文件是有生命的文字組合的觀點，相關的詞彙其出現的距離必定不會太長，一旦相隔太遠，彼此之間的相乘效果就大打折扣，引入距離的因素，比較能夠忠實反應寫作的行爲，距離計算：

$$SNN(ni)=\sum_{j}\frac{IDF(ni)\times IDF(nj)\times f(ni,nj)}{f(ni)\times f(nj)\times D(ni,nj)}$$

$$SNV(ni)=\sum_{j}\frac{IDF(ni)\times IDF(vj)\times f(ni,vj)}{f(vi)\times f(vj)\times D(ni,vj)}$$

$$CS(n)=pn\times SNN(n)+pv\times SNV(n)$$

ni 代表名詞，SSN(ni)爲處理ni這名 詞跟其他名詞的關係， IDF (ni) 就是 ni這詞彙的重要性，IDF (nj) 則是 nj 這詞彙的重要性， f (ni,nj) 是兩個詞彙一起出現的頻率，除以個別名詞出現的頻率 f (ni)、 f (nj) 及 ni、nj 之間的距離 D (ni,nj)，Σ把特定名詞及所有名詞的關係加總，運算結果視爲 ni 這名詞的 SNN。SNV(ni) 爲處理 ni 這名詞跟其他動詞之間的關係，最後把SSN和SNV這兩個分數相加，pn、pv 爲權重，如果認爲一樣重要，則各0.5，如果認爲不一樣重要，可用消去內差法計算，需重覆計算直到 pn、 pv 穩定爲止，算出結果爲 pn0.8、 pv0.2 左右。以下爲消去內差法的公式：

$$SN=\sum_i \frac{Pn\times SNN(ni)}{Pn\times SNN(ni)+Pv\times SNV(ni)}$$

$$SV=\sum_i \frac{Pv\times SNV(ni)}{Pn\times SNN(ni)+Pv\times SNV(ni)}$$

$$Pn=\frac{Sv}{Sv+Sv} \qquad Pv=\frac{Sv}{Sn+Sv}$$

Ni 代表名詞，SSN(ni)爲處理 ni這名詞一旦求 得每一個名詞的聯結強度，便能夠進而得到每一個句子的重要性。假設某一個句子Si 有 m個不同的名詞，該句子被摘錄強度(Extraction Strenght，簡稱 ES)，可用下列數學程式度量：

$$ES(si)=\sum_{j=1}^{m} CS(nij)= /m$$

一個句子是否被摘錄，要算其摘錄強度，將文章所有句子的摘錄強度算出，在依據強度的高低來選擇句子。若要製作定長摘要與最佳摘要，則可以設定一個門檻值 (Threshold)，刪除摘錄強度值小於門檻值的句子，即構成最佳摘要，從最佳摘要再刪除部份的句子，使摘錄的句子數小於原文的10% ，即構成定長摘要。摘要句子出現的順序，應該跟文章出現的順序一樣，因其之間沒有連接詞，讀起來支離破碎，然而製作自動摘要主要爲提供資訊；但這種作法有一些不好，因句子的位置事實上扮演重要的角色，如第一段的第一個句子及最後一段的第一個句子，且線索詞彙(Cue Word)具有舉足輕重的份量，如增益詞 (Bonus Word) 爲重要、顯著，或損益詞 (Stigma Word) 爲不可能、幾乎不，後面的字便很重要，這些句子更需摘錄，因此最後將模型修正爲：

$$ESC(si)=W_1 \times \sum_{j=1}^{m} CS(nij)= /m+W_2 \times POS(si) +W_3 \times CS(si)$$

第一部份用上面的模型算，第二部份爲本身句子出現的位置，第三部份爲句子前面、後面或本身是否有線索詞彙，這三個因素的權重 w1、w2、w3 要自行考量，可得到修正的模型。資訊擷取及自動摘要還有很大努力空間，有待進一步的實驗與修正，以適應網路上各種不同類型的文件，在這領域現有的技術均有其限制，然而最大的問題還是時間，在電腦的軟硬體不斷的進步下，我們期待未來數位化處理方式能簡化，但不會降低精確率。

伍、知識性的服務

網站系統已由資訊檢索進步到提供更有價值的資訊，目前知識性的服務有幾個研究正在進行，大多數爲 Q&A。 Q&A 有兩種型式，一種是直接回答問題，一種是 Answer extract，直接從文件中把答案抽出，這答案可能是一個句子，這些研究自動化的成效目前並不夠好，有部分的答案

動化的成效目前並不夠好，有部分的答案是以人力來做的，相關網站如下，可加以參考。

www.expertcentral.com

www.askjeeves.com

www.inforocket.com

陸、結論

網際網路加速了資訊的流通，縮短資訊形成知識所需要的時間，但網際網路膨脹過於快速，資訊累積太快造成雜訊過多，干擾了資訊的需求及知識的形成，目前雖已提供加值性服務，但仍然不夠，應該更進一步提供訊息性和知識性服務，但後兩者目前仍在初期階段，有待進一步研究開發。

就訊息性資訊服務而言，透過文件摘要可直接、快速地判讀文件的相關性，降低網路的流量；資訊擷取是依使用者所定義的樣板表達資訊需求，由系統直接回覆使用者的資訊需求，在某個角度跟圖書館從事的詮釋資料研究是非常接近，只是詮釋資料的格式是靜態的，訂定好就不太容易變動，資訊擷取的樣板變動性快，隨著使用者資訊需求改變。知識性資訊服務，問題答應（Q&A）是目前努力的目標，自動處理知識短期內沒有可靠的解決方案，因知識非常複雜而精緻，目前仍需依靠人的提供服務。

圖書館學研究的利基在於知識工程及知識管理，今年的中國圖書資訊教育學會舉辦了知識管理這方面的研討會，臺灣大學圖書資訊學系目前也集合了資訊工程學系、資訊管理學系、工商管理學系老師，開設專題研討的課程以討論知識管理，也希望未來成立相關的知識管理學程，都是我們樂見的發展方向。

本文爲書香學苑演講記錄，由蔡崇慧小姐記錄，並經主講者審目同意刊登。