一網看盡五十年—聯合知識庫的建置與應用

聯合線上公司資訊部經理
何 銘 傑

五十年來，臺灣社會跨越了戰後的困乏，經歷了無數經濟及政治環境的變革，聯合報系始終堅持著「正派辦報」的精神，不但忠實地提供讀者「知的權利」，也留下了寶貴的歷史紀錄。這些珍貴新聞資產正是臺灣民眾過去五十年生活的共同記憶，它不只是印刷文字，而是歷史的軌跡與驗證。

過去這些珍貴資料除了做成合訂本之外，也曾製作微縮影及縮印本，近年也有廠商與報系合作將報紙掃描成影像檔，並輸入標題文字製成光碟版。除了在聯合報系之外，只有少數大型的圖書館或研究單位才有。資料或有不全，使用也十分不便，充其量只能稱為資料，不能為社會大眾廣泛應用。

聯合報系線上新聞資料庫－聯合知識庫（http://udndata.com） 的推出。是結合歷史與科技的跨世紀使命，聯合報系過去半世紀的舊報紙，將要在新世紀展現它的新生命，創造新的「數位典藏」價值。

「數位典藏」不僅可以永久保存，更將影響知識的累進過程。目前國內有故宮博物院、國立歷史博物館、國立自然科學博物館、國家圖書館、國立臺灣大學、臺灣省文獻會及中央研究院等七個單位，正在進行數位典藏工作，國科會並通過成立「數位典藏國家型計畫」，今年（2001）共編列一億一千多萬元預算執行數位典藏。

但是這並不包括華文世界新聞資料庫的典藏工作，聯合報系決定以回饋社會的態度發展聯合知識庫。以聯合知識庫的規模，不僅目前國內其他新聞出版事業沒有如此打算，放眼世界，也未曾發現有類似規模的計畫；聯合報系執行這項知識工程，係從「社會價值」的角度出發，希望可以對華人社會的每一分子都有所貢獻，這與聯合報系過去對全球華人辦報的想法是一致的。

聯合報系在跨越新世紀的同時，藉由科技的協助，以無比的決心讓舊報紙展現它的新生命，不再只是躺在大樓的空調房間裡的一大堆泛黃的紙張，歷史也藉由新科技的呈現而鮮活起來。

當讀者可以很快速地搜尋過去五十年的新聞，閱讀新聞的態度及使用資料的方式也可能有所轉變。當「臥虎藏龍」得到奧斯卡金像獎十項提名，相信有很多人想要了解李安過去的點點滴滴，過去除了到圖書館一張張地翻尋舊報紙之外，並沒有其他工具可用。除此之外，讀者也可以透過聯合知識庫清楚地了解每一個政治人物在不同時期對一些重大政策的態度是否始終如一。

聯合知識庫不僅只是提供新聞資料的檢索而已，更衍生數位典藏歷史的社會責任。藉由數位典藏的科技，活化聯合報系五十年的新聞資料，完整而翔實地呈現臺灣近代珍貴史料，傳遞文化變遷的軌跡，成為新世紀知識的庫藏中心，打造知識社會的基礎工程。

聯合知識庫從 2001 年 2 月 19 日推出以來，廣受社會的支持與肯定，在短短不到三個月的時間就有三萬使用者申請加入會員，也有相當高的比例申請加入付費會員。在目前正值網際網路業艱困的時期，無疑像是寒冬的一股暖流；黎明前的曙光。鼓舞了同業，願意以更積極的做法發展網際網路的內容產業。

在此我們也十分樂意與大家分享聯合報系資料數位化及建置聯合知識庫的過程與心得。

「聯合知識庫」的建置

聯合報系近二十年來推動報紙作業流程的電腦化，也建置了無數的系統，不過只能稱為是生產報紙的系統，所有的流程都是為報紙的生產而規劃、設計，所有的作業都只到報紙生產完成就結束了，早些年也試著將每天生產過程的電子檔留下來。技術上是可以逐筆轉換成網站所需的資料庫，不過會有些資料與實際見報的內容不符，且成本不低，就作罷。

在 1999 年 9 月配合聯合新聞網 udnnews.com 的推出，聯合報系經由一年的努力，成功的將每天各報所有見報的新聞一字不差的逐篇存入資料庫中。經近兩年的收集也有數十萬筆可觀的資料。不過在此之前的資料只有報紙合訂本可用。在林口廠有一個溫溼度控制的倉庫，存滿了這50年的報紙合訂本。估算有一百三十萬塊版，一千餘萬則新聞，七十餘億字。

要把整個倉庫泛黃的舊報紙轉換成文字電子檔，確實是一件相當艱鉅的工程。在歐美、日本等有悠久歷史的知名報社也未曾有如此大規模處理過去紙本資料的計劃，當然在國內也是一項空前的創舉。

多年來不斷嘗試市場上的各種技術後，覺得原先主要的技術瓶頸均有解決的方案，應有相當高的成功機率可以完成此一工程。就在1999年初提案建議社方發展此一計劃，也立即獲得報系的全力支持，在同年四月底就正式成立籌備小組，積極推動各項籌備工作。

主要的關鍵技術

聯合報系之所以能夠完成此一資料數位化的計畫，並順利推出聯合知識庫，主要當然是社方的遠見及全力支持的魄力。除此之外，當然就是我們擁有的一些突破性的技術。我們覺得能夠完成此一計劃有三項主要的關鍵技術，分別介紹如後：

一、多核心的OCR及高效率的校對技術

由於有七十餘億字，要全部重新輸入預估需要3至4億新台幣，而且一時之間也無法找到如此龐大的人力。送到大陸處理也有諸多技術上的困難。多年來我們一直在收集大陸及臺灣各種光學辨識（OCR）的技術，最後終於在全景公司的協助下，成功發展出一套符合需要聯合報系需要的光學辨識（OCR）處理系統。

這套處理系統率先採用多核心的光學辨識技術，讓多個 OCR 核心同時辨認，將系統認為錯誤機率較高的文字影像另外標示，並透過稱為「集字校對」的作業，將 OCR 認為同一個字的影像全部集中批次校對。經由這樣的處理之後，平均每人每天可以處理 20 萬字，而且文字的錯誤率可以低於千分之二。以如此的作業效率及品質均已遠優於重新打字輸入。

也是因為有了這個突破性的做法，將原本預估需要 3 至 4 億新台幣的作業，縮減至幾千萬就可以完成。

二、高效率的搜尋引擎

一般的搜尋引擎在幾萬筆及少人使用時，會有不錯的表現。資料超過數十萬筆或是多人同時使用時，馬上出現瓶頸。以聯合知識庫的資料量及經常數百人同時在線上檢索的使用下，不是一般搜尋引擎可以勝任的。

在搜尋引擎的選用上，也是經過長時間的評估及測試，最後採用了中華訊息公司的高效率搜尋引擎 Q base。不論在大資料量的檢索、建庫的效率，多人同時使用的情況下，均有不錯的表現。

三、人工智慧技術的引進

「聯合知識庫」為了讓使用者可以有最簡單的方式找到想要的資訊，特地引進美國亞利桑納大學人工智慧實驗室與「知識運算公司」（KCC）的知識管理系統，開發中文相關詞的人工智慧分類，除了提高使用效率，也可以激發使用者對於資訊的不同思考。

主持此一實驗室是一位十分傑出的旅美華人陳炘鈞教授。在美國，人工智慧的知識管理系統已被很多單位應用，例如，情報系統、警政系統、證券期貨、醫學界、大型公司的內部管理等，當分析資料庫的需求是「準確、快速、簡單」時，需要使用人工智慧。

人工智慧的知識管理系統應用在中文新聞資料庫上，聯合知識庫是首開先例。由於聯合報系五十年來累積的新聞資料量非常龐大，除非使用者可以自行掌握日期、作者或其他關鍵字等資料，縮小範圍繼續檢索，否則反而會造成更大的困擾。因此，想要協助使用者查詢到真正想看的新聞資料，需要加上一些輔助工具，例如人工智慧系統，簡單地說，聯合知識庫的「智慧查詢」是「一種會思考的全文檢索」。

智慧查詢不但幫助使用者有效率地查詢，更可能讓使用者產生新的想法，因為在兩百個相關詞之中，經常可以發現人腦所無法記憶的相關詞彙，這種資料庫尋寶過程的意外發現，經常能促成多面向的新思考。

此一技術運用在中文環境時並無障礙，重點在於文化背景差異，在分析詞彙時需採用不同方法，而聯合報系累積的豐富資料正是最好的對象，因為記者素質整齊，語法用詞接近，透過人工智慧系統呈現的結果也會更好。

「聯合知識庫」的主要功能

「從資訊邁向知識」是聯合報系發展「聯合知識庫」的主要目的之一。因此，在系統功能的設計上，不只有一般的全文檢索，而是以「知識管理」為主軸設計了四項主要工具：熱門字串、我的剪報、專卷查詢、自動剪報系統。讓使用者可以用最快速、最簡單的方式，找到真正想要的資訊，並成為個人化的知識管理工具。

一、全文檢索

目前國內相關資料庫或入口網站都具有基本的搜尋引擎功能，雖然搜尋的速度有別，但基本上都是基礎的全文檢索，聯合知識庫當然也有類以服務。除此之外，在搜尋字串的旁邊有「熱門字串」功能，系統將自動統計每週最熱門的一百個關鍵字提供使用參考。

二、智慧查詢

但事實上，一般的全文檢索不但不能滿足使用需求，甚至可能造成更大的困擾。例如鍵入「陳水扁」三字，使用一般的搜尋引擎檢索過去的資料庫，結果可能出現數萬筆的資料，反而更無所適從。因此，聯合知識庫引進美國亞利桑納大學人工智慧實驗室發展的工具，設計成「智慧查詢」功能，成為「會思考的全文檢索」，可以導引使用者的多面向的思考，將龐雜的概略性資料去蕪存菁，留下使用者真正需要的資料。

三、專卷查詢

聯合報系資訊中心經二十年來收集整理，目前數量已有一萬餘卷的紙本專卷資料庫。不僅是報系編輯部十分重要的參考來源，多年來有許多單位也透過各種關係，希望能夠來使用此一臺灣收集最完整，資料最豐富的資料庫。籌備小組也依據過去多年製作專卷的經驗，發展一套網路版的「專卷製作系統」。專業的資料編輯就依人物、事件、主題來設定題目製作專卷，以專業的搜尋技巧及人的判斷，提供比一般檢索更精準的資料。方便使用者以更方便的方式取得更精確、完整的資料。

網站上也特別提供「專卷查詢」的首頁，使用者可以透過「卷名檢索」或是「分類瀏覽」的方式來尋找所需的專卷。另外也有「最愛專卷」的個人化服務及「熱門專卷」排行及「推薦專卷」的服務，方便使用者以輕鬆的方式找到所需的專卷。

四、我的剪報

除了找資料，聯合知識庫也提供個人化的資料管理系統，長久以來，人們習慣用剪刀與膠水做剪報，現在則讓電腦負責「剪報」。對於個人而言，可以使用「我的剪報」，設定關鍵詞之後，將新的資訊和舊的資料逐日累積，創造屬於個人的知識管理系統。由於這是個人化的知識管理平台，「我的剪報」也提供「匯入外部資料」的功能，使用者可以把非聯合知識庫的資料納入個人剪報夾中。

五、自動剪報系統

由於聯合知識庫每天新增有一千五百餘則新聞，回溯數千則新聞。相信每個人每天看到的資料只是全部資料的一小部份。所以也特別為企業會員規劃了「自動剪報系統」的功能，使用者可以預先設立匯入資料的條件，每日新增的新聞如有符合條件，就會自動匯入指定的檔案中，也可以透過email 自動寄給所有的訂閱者，協助使用者消化每天新增的龐大資料。

會員的服務

聯合知識庫的基本服務是免費的，使用者只要上網，不填寫任何資料，仍然可以檢索最近一個月的新聞標題，並免費查詢最近一週的新聞全文；當使用者填寫資料登記為會員後，就可以檢索最近一年的新聞標題和免費查詢最近一個月的新聞全文。上述兩種免費服務都可以使用「智慧查詢」的服務，包含「進階查詢」、「相關詞」和「相關分類」，並了解「相關專卷」的卷名。

進一步的服務則基於「資訊有價」的基本理念，採取「付費制」，分為學生會員、菁英會員以及企業會員三種，都可以查詢閱讀所有日期的新聞標題和全文，也可以使用所有智慧查詢的功能。

由於需求不同，會員的費用、可使用的功能與可查詢的資料筆數額度也不相同。基於回饋社會的理念，學生會員的會費只要六百元，一般使用者的菁英會員會費為兩千元，二者都可以使用「專卷查詢」與「我的剪報」等功能，但可閱讀的全文資料筆數不同。企業會員的會費為兩萬元，除了上述功能，更有獨享的「自動剪報系統」功能。聯合知識庫推出之後，因應各大圖書館及研究機構的需求，也特別推出一些不限使用點數的專案。

「聯合知識庫」的運用

過去對於大型資料庫的印象都是十分稀少的資源，只有在國家級的圖書館、研究機構或是大型企業才能用得起。相信大部分的大學畢業生不曾使用過資料庫或是微縮影片。相信是因為資料庫高昂的費用及使用的不便，讓大家覺得似乎只有專業的研究人員才會用到資料庫。

相信聯合知識庫的推出，有助於將此一原屬「貴族」的資源「平民化」；「全民化」。從幾十年前筆者讀小學到現在筆者的小孩念小學，都還在比賽查字典。相信很快國民小學除了比賽查字典之外，還可以比賽查資料庫。

兩個月來筆者念小學三年級的女兒幾次要我幫她找老師交代的功課，有關於春天的文章及照片，還有東港燒王船、頭城搶孤、鹽水蜂炮…等等臺灣民俗的源起及相關資料，前後幾個題目都在聯合知識庫找到非常豐富完整的圖文資料，女兒很滿意之外，隔天當然也得到老師的讚賞與鼓勵，還問在哪裡才可以找到這麼豐富的資料。

相信大家都能夠同意，訓練我們小朋友提昇他們運用資料的能力，可能比要他們背資料重要得多。再者現在硬碟很便宜，其實不需要將我們大腦當硬碟用，只要我們能夠需要資料的時候，能夠很快取得，也有一些方便的工具可以迅速將一大推的「資料」整理成「資訊」進而轉換成自己的「知識」，也就足夠了。

過去限於科技無法達到此一理想，現在經聯合報系的努力，推出聯合知識庫，並透過無遠弗屆的網際網路，確實改變大家使用資訊的方式。相信可以藉此幫助許多使用者及機關提昇他們決策的品質。

聯合知識庫推出以來，除了有許多媒體同業及圖書館加入企業會員，我們也發現有許多警察局等警政單位也陸續申請成為我們的會員。從聯合晚報的報導中得知，警方開始在所經辦各類刑案中，尋找早期類似作案手法案情或未破案線索，協助刑案的偵辦腳步。警方也表示，偵辦刑案中最頭痛的就是無法找尋以往類似犯罪手法的資訊，或相關可供參考線索，如今有聯合知識庫的協助，對警方偵辦案件時尋找早期相關資訊有極大助益。

桃園縣警局局長侯友宜在2月26日的聯合晚報也指出，警方今後辦案，除了從上網尋找類似犯案手法案件或資訊外，各轄區也要完成包括縱火犯、強制性侵害嫌犯等嚴重危及地方治安類型人犯資料，這些資料除了可以上網連成一氣外，還可以供辦案快速釐清線索，讓所有辦案腳步結合，對辦案將有相當助益。

我們也十分高興看到聯合知識庫除了可以協助學術研究、小朋友寫功課之外，還能幫助警方辦案，為打擊犯罪盡點心力。

「聯合知識庫」近期的計劃

一、提昇資料的深度及廣度

（一）在資料的深度方面

原本只是計劃以印務部每日例行生產的空檔來處理這些回溯的資料，不過社方為加速整個作業的進度，另外每年撥一千餘萬的預算，以外包人力的方式加速整個生產的進度。希望能夠在 2001 年 9 月 16 日聯合報50週年慶前，能夠處理完成聯合報自 1988 年報禁解除以來共計13年的全國版資料。全部預計四年將聯合、經濟、民生、聯晚及星報五家報紙自創刊以來的資料，全部放入資料庫中。

（二）在資料的廣度方面

雖然報系擁有如此豐富的資料，不過對於一個資料使用者而言只有聯合報系的資料還是不完整的。希望藉由聯合知識庫的投入能夠獲得其他媒體或是出版商的認同與支持，陸續邀請臺灣、香港及大陸其他正派經營的媒體一起加入聯合知識庫的平台。真正落實成為「華文世界中最完整新聞知識的查詢中心」。

二、提供更完整的網站功能及服務

雖然聯合知識庫已經在第一階段提供智慧的「全文檢索」、「專卷查詢」、「我的簡報」及「自動剪報系統」四項服務及工具。不過距離理想的知識管理還有相當的距離。未來也將繼續不斷的更新各種人工智慧資料加值技術及各種貼心的服務，能夠將「資料」轉化成「資訊」在提昇成為「知識」，為「知識管理」提供新的註解與典範。

只是單純的新聞資料並沒有太大的意義，但如果可以經過組織整理，輔以人工智慧的技術，使這些內容能有意義地呈現，如此的資料才能產生知識。從新聞資料庫開始，利用智慧型的關鍵字檢索，應是很好的工具。目前全世界建置資料庫的主流方向為「Metadata」，中文可解釋為「詮釋資料」，基本理念是「從資料發展資料」（data about data）。在聯合知識庫即將在下半年的推出的新版的功能中，就可以讓使用者在所找到的資料中，自動帶出一些專有名詞的解釋，或是一些知名新聞人物或是公司行號的詳細背景資料，可以讓使用者非常方便的瞭解有關該篇報導的更深入的資料。

聯合知識庫網址：http://udndata.com