大數據分析在檔案管理的應用潛力(一)

最近幾年，大數據分析與人工智慧的結合在各種領域的應用都有令人驚豔的成果，例如人臉辨識、汽車自動駕駛、市場研究、高齡族生理資訊監控與危險訊號警示、信用卡防詐欺、全球暖化監控與預警。大數據分析的前提當然是要有大量的數據，這些數據通常是經過數位化的數字、文字，影像、甚至聲音檔案。透過特徵 (Features)的萃取與模式 (Patterns)的建立與測試，找出數據之間的關聯性，藉以達到解釋、預測、控制的目的。

這種「經過數位化的數字、文字，影像、甚至聲音檔案」恰巧是檔案管理領域中數量最多，應用潛力無限的資源。許多機關在檔案數位化的過程中，將歷年來數十萬，甚至高達百萬份的公文掃描，並且透過回溯建檔，依照檔案管理局的規範，建立了影像檔案管理系統。這些資源正是大數據分析可資運用的「寶藏」。

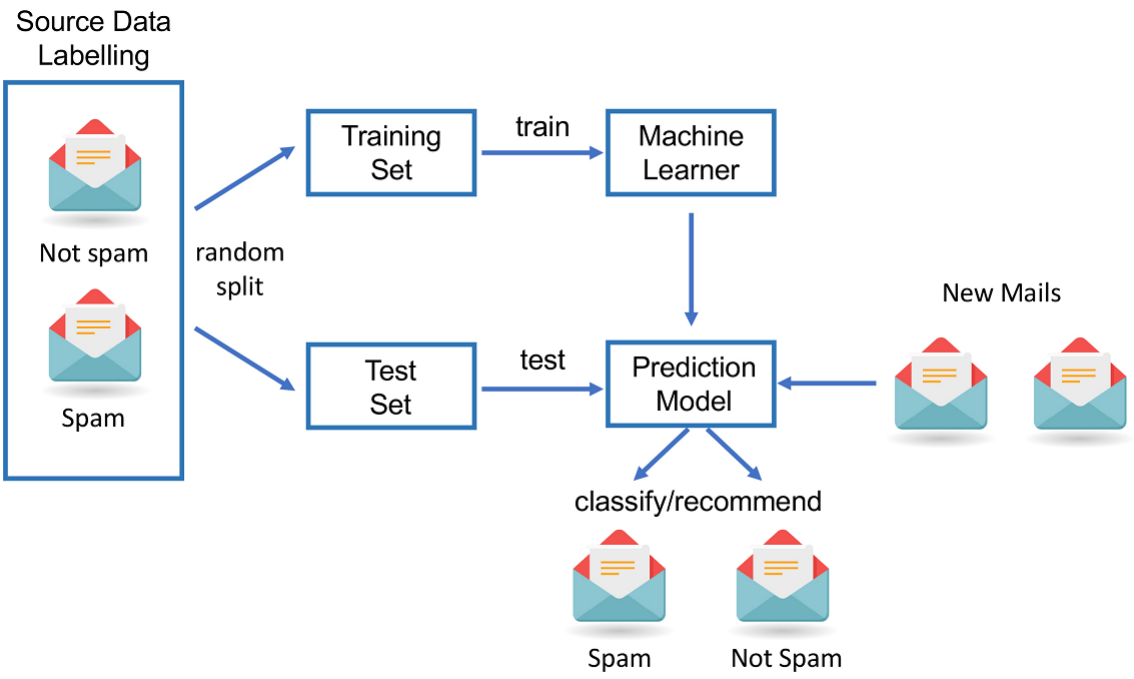
有了寶藏並不意味著就能「發財」！還需要經過探勘、採礦、萃取提煉的程序，而這些程序的每一個步驟都必須投入大量的人力、物力、財力與時間，而且可能要承受失敗的風險。同樣地，大數據分析與應用也必須投入龐大的資源才能獲得有用的成果。前面所提的許多成功案例都有一個共同特徵，就是其背後有龐大的商機，足以讓投資者願意投入龐大的資源進行研究發展。不幸的是，檔案管理領域講求的不是「商機」，而是「文化資產的保存」，因此無法吸引投資者的青睞。或許正是這個原因，大數據分析在檔案管理領域的應用研究始終處於落後的狀況，使得其擁有的「寶藏」始終深埋在地底。

觀之最近幾年國際上與檔案管理相關的研討會中，都有徵求大數據分析有關的議程，可是發表的成果卻相當少。比較具有參考價值的，據筆者的看法，只有 IASA 2019 研討會中由英國薩里大學 (University of Surrey)的 Mark Plumbley 教授專題演講「AI for Sound: A Future Technology for Sound Archives」，以及 ICA 2019 Adelaide，由檔案管理局委託維運的電子檔案長期保存實驗室所發表的「The Application of Machine Learning to Archives」。

本系列文章的主要目的是以淺顯的方式介紹大數據研究在檔案管理的潛在應用，以及一些相關的技術背景介紹。期望能夠引發檔管專業人員的興趣與參與討論，為我國在這一方面的研究能夠奠定足以持續耕耘的基礎。因為考量目標讀者的專業與技術背景以及篇幅的限制，文內所提的技術與方法，在學術上而言，並不是很嚴謹的，特此聲明。

機器學習 (Machine Learning)是最常應用在大數據研究的方法之一。機器學習的方法大致上可以區分為兩大類：監督學習 (Supervised Learning)與非監督學習(Non-Supervised Learning)。監督學習最常應用在分類 (Classification)的問題上，例如垃圾郵件的分類。假設我們有一些電子郵件，由「專家」將這些電子郵件先進行分類，標記為「垃圾郵件」與「非垃圾郵件」。然後利用機器學習的演算法(後面會介紹)找出這兩類郵件的「特徵」，也就是找出有哪些特徵可以用來區別垃圾郵件與非垃圾郵件。有了這些特徵，我們就可以用來過濾或分類垃圾郵件了。

在實際的應用上，我們通常會先把所有的電子郵件標記(Labelling)為垃圾/非垃圾郵件兩組，如圖一最左邊所示的 Spam 和 Not Spam，然後隨機分類為訓練組(Training Set)與測試組(Test Set)，再用機器學習方法(Machine Learner)針對訓練組資料進行訓練，找出預測模式 (Prediction Model)之後(圖一上方的程序)，再用這個預測模式針對測試組資料進行分類。因為測試組資料事實上是已經被標記的資料，所以我們可以拿這些專家標記的結果和預測的結果進行比對，藉以評估我們得出來的模式正確率有多高。如果正確率達到滿意程度我們就可以利用這個預測模式來判斷新進的電子郵件是否為垃圾郵件，如圖一右邊所示。



圖一：機器學習分類問題示意圖，以垃圾郵件判別為例

圖一所介紹的方式只有兩種分類結果：垃圾郵件與非垃圾郵件，我們可以將這種方法應用到「檔案永久保存價值鑑定」。也就是說，我們事先將檔案分類為「永久保存」與「非永久保存」兩種，再利用圖一的程序，找出分類的預測模式，如果達到滿意的正確率，就可以協助進行檔案管理中歷史檔案鑑定作業。

在實際的研究中，這一項分析的應用價值不高，原因是兩個分類中資料量比例差距過大，也就是說，被歸類為「非永久保存」的數量遠比歸類為「永久保存」的數量高，例如兩者的比例為 99:1，造成預測模式幾乎一律將測試組的資料歸類為「非永久保存」，即使這樣，這個模式的準確率仍高達 99%，因為 99%的資料都是「非永久保存」，所以這種研究結果是無法被接受的。研究團隊目前還在研究如何解決這種比例差距過大的資料問題。

事實上，分類問題也適用於多種分類的問題上，也就是說在將原始資料標記時，可能的標記選項超過兩種。例如如果我們有一組電影劇情的文字介紹，我們要根據這些劇情介紹將電影進行分類為「劇情片」、「紀錄片」、「動作片」、「驚悚片」...。雖然標籤分類的選項數量不一樣，但是機器學習的方法與程序大致相同。在檔案管理的應用上可以用來做「分文建議」或「分類號建議」。在分文建議作業中，我們可以利用圖一所示的方法建立預測模式，建議一份新的來文分配到哪一個單位/承辦人承辦較恰當。在這種情境下，所有的單位/承辦人都是可能的分類。在分類號建議作業中，當有一份公文要歸檔時，我們可以利用公文的主旨或說明，判斷這一份公文應該歸類到哪個分類號。在實際的實驗中，我們採用三個機關的歸檔資料，運用前述的方法測試結果，正確率介於 65% – 82%之間。我們進一步分析發覺，

正確率較低的機關，其歸檔分類標準本來就不夠一致性，換句話說，如果一個的機關歸檔資料中，分類號的建立規則缺乏一致性，用這些資料所建立的預測模式準確度也較低。

機器學習中最常使用的技術之一是稱為 TF-IDF (Term Frequency - Inverse Document Frequency)。TF-IDF 是一種統計方法，用以評估一字詞(Term)是否夠資格成為特徵。簡單來說，如果某一個字詞大量重複出現，就表示它可能是一個特徵。例如，在垃圾郵件中常常會出現「促銷」、「機會不多」、「大拍賣」、「機會難得」等字詞，所以這些字詞可能就是垃圾郵件的特徵。相對地，如果一份公文的主旨出現「總統」、「法案」、「頒布」、「派令」等字詞，那麼這一份公文屬於歷史檔案的機率就很高。

以分文建議為例，在實際的運算中，首先我們要蒐集已經完成分文的【公文主旨】以及該公文分文到哪一個【單位承辦(標記)】的資料。接下來要進行「斷詞(segmentation)」作業，也就是將所有的主旨裡面的字詞辨識出來。例如表一公文一的主旨「函轉行政院機密檔案管理辦法乙份(如附件)，請查照。」範例可以被斷出「函轉」、「行政院」、「機密」、「檔案管理」、「辦法」、「乙份」、「如附件」、「請查照」等詞句。主旨「函轉行政院來函頒布行政院修正文書處理手冊，並自中華民國九十年二月十三日生效，請查照」可以被斷出「函轉」、「行政院」、「頒布」、「修正」、「文書處理」、「手冊」、「中華民國」、「生效」、「請查照」等字詞。斷詞作業可以透過事先建立好的詞庫，用電腦程式自動執行，斷詞的精準度直接影響後續建模與預測的精確度，因此辭庫的內容非常重要。

表一：範例公文與斷詞

公文編號	主旨(斷詞後)
公文一	【函轉】【行政院】【機密】【檔案管理】【辦法】乙份(【如附件】)，【請查照】。
公文二	【函轉】【行政院】【來函】【頒布】【行政院】【修正】【文書處理手冊】，並自【中華民國】九十年二月十三日【生效】，【請查照】。
公文三	【檢送】【本部】【第】九九六次【部務會報】【重要】【決議】暨【主席指（裁）示事項】一份【如附件】，【請查照】【辦理】。

斷詞完成後要進行的是特徵萃取 (Feature extraction)，常用的方法就是 TF-IDF。計算 TF 字詞頻次就是統計某一字詞在某一份公文出現的次數。例如「行政院」在前面的例子第一份公文出現一次，在第二份公文中出現兩次，「請查照」在所有範例公文中各出現一次。當所有的字詞在所有的公文中出現的頻次都統計完畢後，我們可以得到類似表二所示的矩陣(省略部份字詞)。

表二：字詞頻次 (Term Frequency)表

	函轉	檢送	行政院	檔案管理	文書處理手冊	部務會議	請查照
公文一	1	0	1	1	0	0	1
公文二	1	0	2	0	1	0	1
公文三	0	1	0	0	0	1	1
TF	2	1	3	1	1	1	3

就直覺來說，某一字詞在某一分類中出現的頻次越高，越有機會成為該分類的特徵。例如分文給秘書處的公文主旨內出現【文書處理手冊】或者【檔案管理】的頻次很高，所以這兩個字詞可以當作分文給秘書處的特徵。換句話說，要是某一份公文的主旨出現【文書處理手冊】或者【檔案管理】，這份公文就很有可能應該分文給秘書處(當然還要搭配其他特徵)。可是這種直覺做法有例外，例如【函轉】或者【請查照】這兩個字詞出現在其他單位公文主旨內的頻次也很高，但是卻不能當作是該分類的特徵，因為這些字詞太通用了。避免過於通用的字詞被當作特徵，就要用 IDF 來修正。

如果某一字詞同時出現在多數公文內，表示這個字詞通用性高，相對地，其特徵性就降低。所以我們在計算 IDF 時先計算某一字詞在各不同公文中出現的頻次(Document Frequency, DF)，再用 DF 來修正 TF。表三是前面範例的 DF 列表。注意，表二的【行政院】一欄 TF 是 3，因為在三分公文中總共出現過三次，可是在表三中 DF 是 2，因為只有兩份公文中出現過【行政院】這個字詞。

表三：文件頻次 (Document Frequency)表

	函轉	檢送	行政院	檔案管理	文書處理手冊	部務會議	請查照
公文一	1	0	1	1	0	0	1
公文二	1	0	2	0	1	0	1
公文三	0	1	0	0	0	1	1
DF	2	1	2	1	1	1	3

表二的 TF 和表三的 DF 再套入公式：

$$w_{i,j}=tf_{i,j}\times\log\left(\frac{N}{df_i}\right)$$

上面的公式 N 表示所有公文件數，$tf_{i,j}$表示矩陣（i, j）元素的 TF 值，df_i表示矩陣第 i 個字詞的 DF 值，log 是自然對數，$w_{i,j}$是第 i 個字詞在第 j 份公文裡的 TF-IDF 值。以上面範例的【請查照】為例，這個字詞的 TF-IDF 計算如下：

【請查照】在公文二中的 TF，$tf_{2,7}=1$，總公文數 $N=3$，該字詞的 DF（表三)是 $df_7=3$，套入公式，得到 $w_{2,7}=1\times log(\frac{3}{3})=1\times log(1)=1\times 0=0$，所以該字詞在公文二成為特徵的權重是 0，因為【請查照】這個字詞再公文中太普遍，所以不能當作特徵。全部權數計算出來後，得到的矩陣稱為特徵矩陣(Feature Matrix)。這個特徵矩陣後續還可以用來進一步進行深度學習(Deep Learning)，目前最常用的深度學習方法是 TextCNN (Convolutional Neural Network for Text)，一種將類神經網路(Neural Network)用在分析文字文件的方法。深度學習後得到的修正後的特徵矩陣就可以用來協助進行單位分文作業，例如從特徵矩陣中發覺，主旨中同時出現【行政院】、【文書處理手冊】，或者【檔案管理】則 92%的公文是分文到秘書處承辦的，我們就可以得到一條分文的規則，如果來文主旨中出現行政院】、【文書處理手冊】，或者【檔案管理】，則建議分為到秘書處。

在實際運算中，TF-IDF 與 TextCNN 的計算，以及規則的萃取，驗證以及應用都可以使用諸如 Scikit, Tensorflow, Kera 等開放原始碼 (Open Source)套件執行。在下一季的國外科技新知中，我們將介紹另外一種機器學習方法群聚分析 (Cluster Analysis)以及這種方法在檔案管理中主題項、併案、相關文件作業的應用潛力。