「全文資料庫」專輯
卷首語

謝清俊

每一個人都看過不少書，然而您可曾想過：一本書究竟帶來多少種不同的訊息？

打開一本書，你會發現印在書上的，除了正文以外，還有引得，它包括正文目錄、圖表目錄、索引....等等；有提供背景資料的，包括序、跋、摘要、作者介紹....等等；有輔助說明的訊息，像是公式、註解、校勘、圖表說明、圖、表....等等；書目資料，如出版者、發行、出版年代、版次....等等。除此之外，還有兩類很重要的訊息，那就是：此書的結構和排印的訊息，雖然他們不是刻意表明的，但確實呈現在讀者的眼前。

事實上，上述的清單還不完備，一本書還有些屬性是不曾印在書上。試看看圖書館中的書目卡片，或是計算機中書目檔案的結構，你將會訝然發現一本書的相關屬性，竟然有那麼多！別急，還沒有算完這筆賬呢！試想，每本書不是遺世獨存的，這本書中有指著其他書的訊息，其他的書也許有參閱上書之處，這些不止出現在參考書目和備註之中，也有些潛伏在正文、鍵語和索引之內。

如果由表達的媒體來看，書中有文章的字串、有圖、有影像、有表格。光是處理這些不同的表達法，就須用到自然語文處理、計算機圖學、影像處理、圖形識別、資料庫等一大堆先進的技術。若是要把書念出來，機器還得有良好的語音合成本事。

當你看了這些，也約莫了解全文處理面臨的環境。上述的資訊都可能是全文處理需要應付的，或者是在處理過程中相關連的。全文處理技術的發展，爲的是要使電腦能處理我們人使用的文獻，而不再局限於表格或欄位化的資料。它是屬於自然語文處理中的一支，對象是處理在結構上大於句子的各種自然語文的文獻。

全文的應用很廣，從幫忙作文的文句處理（word processing），到排版的自動化；從辦公室的公文自動化，到各種資料庫網路；從個人的電子檔案（不僅是目前表格形式的）到全電子化的圖書館，甚至到人工智慧型的人機界面、知識庫系統....等等。本專輯所談的全文資料庫，是全文處理中很重要的一環，它是所有全文資料庫存、管理、維護、檢索、列印....等等功能的基礎工程，也可作知識庫、智慧型工作站之資料基地。

全文資料庫的發閤甚早，大量推出是1983年以後的事。發展的地區以美國爲首要。在本專輯的第一篇文章中，將介紹美國全文資料庫的演進和發展，作爲本專輯的背景資料。在第二篇文章中，將詳爲讀者解說全文資料庫與現有的資料庫在結構上之異同，以期對全文資料之典藏與管理有正確之體認。第三篇文章爲各位介紹全文資料庫中很重要的一項必要技術——檢索方法，這是一個目前十分熱門的研究方向。最後一篇文章是一個個案介紹，我們將以中研院史語所爲首，所發展的史籍全文資料庫爲案例，介紹一個中文的全文資料庫，以期與前三篇互相印證。除了第一篇文章外，其餘三篇均以中文作爲主要舉例的語言，這是作者們奉獻給讀者的一番心意。

目前，在日本、大陸亦有計畫在研究中文的全文資料庫，均有相當的成果。在國內，已有幾個實用的系統，瀕臨實用的階段，其中包括：立法院的法規資料庫、新聞資料庫、質詢與答覆資料庫等等，規模甚爲完備；行政院法規會亦有計畫將全國法律以全文資料庫作線上之服務，此系統已曾在去年展示。此外，電傳視訊中亦有全文資訊，只可惜檢索功能尚須加强。由是觀之，國內對全文處理所需日爲殷切，舉凡法院判例、專利、出版、印刷、公文等等，皆需此技術以資自動化；再如地政與户政之自動化亦與全文處理有關。像這樣一個影響深遠的技術，在本專輯中，甚難涵蓋周延。我們略去了全文資料庫的查詢語言設計部分、語句分析和語意處理部分、文獻報表產生的部分、標誌語言（mark-up language）的設計與標準化部分、欄位化的資料庫（傳統的資料庫）和全文資料庫的整合部分、人工智慧應用於資料庫設計的部分、工具書配合人工智慧自動化的部分，以及其他的各種應用之敘述。這些部分都是目前正在研究的方向。

配合著硬體的進展，尤其是計算機的容量和計算能力的提升，計算機處理資料的能力，已擴大到可以處理大批文獻的地步。如果有一天，計算機眞能把一本書中的訊息處理得得心應手，那麼，像「霹靂遊俠」中的霹靂車，或許會有眞正實現的一天。

最後，希望本專輯能提供些你想要的訊息。