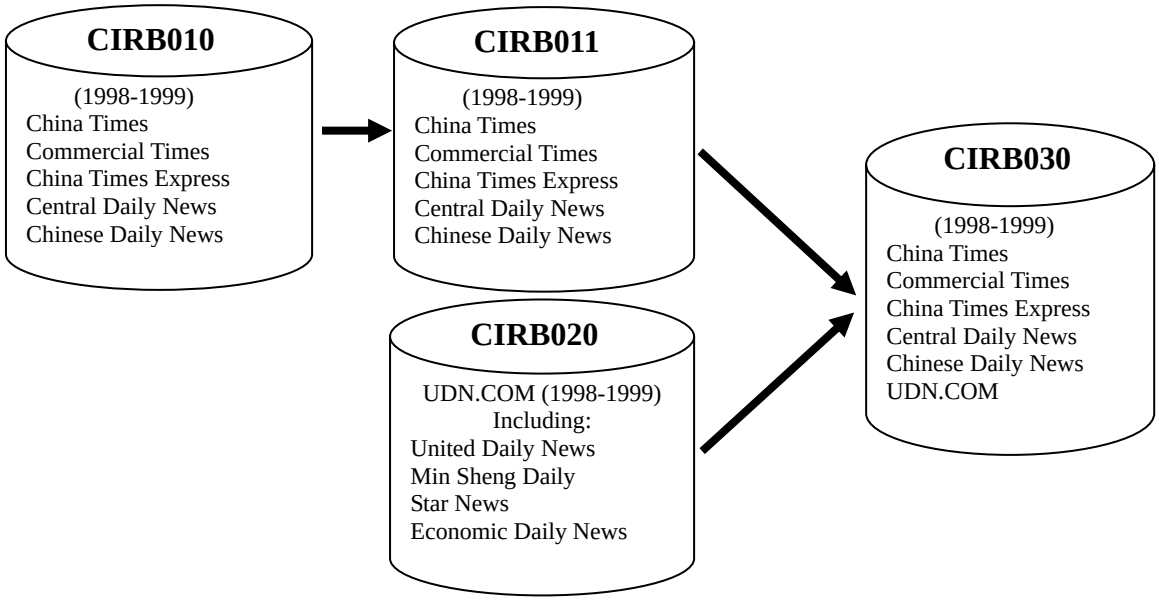
CIRB030 (Chinese Information Retrieval Bench, version 3.0)簡易說明

陳光華
國立台灣大學圖書資訊學系
khchen@ntu.edu.tw

陳信希
國立台灣大學資訊工程學系
hh_chen@csie.ntu.edu.tw

資訊檢索測試集是用於評估資訊檢索系統的績效。測試集對於資訊檢索系統開發過程的評量與資訊檢索系統的評量而言，都是一種極為重要且有效的工具。此次發行的 CIRB030 (Chinese Information Retrieval Benchmark, version 3.0)便是用於評估中文資訊檢索系統的測試集。一般而言，測試集包含三個部分：文件集，問題集，相關判斷（答案集）。本文簡要說明 CIRB030 資訊檢索測試集。CIRB030 和用於 NTCIR3 Workshop 的 CIRB011 與 CIRB020 有些許不同。基本上，CIRB030 的文件集是由 CIRB011 與 CIRB020 的文件集合組而成，但修改了部分文件亂碼的問題，以及部分文件 HEADLINE 與內文不符的問題。而 CIRB011 和先前發行的 CIRB010 僅有標記上的不同；CIRB011 與 CIRB020 的標記則完全相同。這一次我們決定直接發行 CIRB030 而跳過 CIRB020，原因就是想整合 CIRB011 的文件集與 CIRB020 的文件集。因此，您無須擔心如何取得 CIRB020 的問題，它不會在台灣單獨發行，除非您曾參與 NTCIR 資訊檢索評估會議。為了更清楚地說明 CIRB 版本的演變情形，請您參考下圖。



另外必須要注意的是，CIRB030 的新聞文件已經整合為 7 個文件檔案，它們是 cdn1998-1999 (中央日報)，chd1998-1999 (中華日報), ctc1998-1999 (工商時報)，cte1998-1999 (中時晚報)，cts1998-1999 (中國時報)，udn1998 (聯合報系 1998)，與 udn1999 (聯合報系 1999)；而這些新聞文件在 CIRB011 與 CIRB020 時是各自獨立的，計有 381,375 個文件檔案。

至於問題集，我們最初共製作 50 個問題，然而我們在 NTCIR 資訊檢索評估會議後，刪除了 8 個較不適用的問題，因此最後僅有 42 個問題。

相關判斷（答案集）仍有兩組答案，一組為嚴謹相關（Rigid Relevance），也就是非常相關與相關視為相關；一組為鬆散相關（Relaxed Relevant），也就是非常相關、相關、部分相關皆視為相關。