

中英文字檔案區域調整資料壓縮方法

A Locally Adaptive Data Compression Scheme for Chinese-English Text Files

張克章* 徐熊健** 朱賢武+

*國防管理學院資訊管理學系

E-mail:chang@rs360.ndmc.edu.tw

**銘傳管理學院資訊管理學系

+國防管理學院資源管理研究所

摘要

隨著資訊科技快速發展與普及，人類仰賴資訊網路進行國際間資料交換的需求與日俱增，因此適合應用於網路傳輸的資料壓縮方法成爲一項紓解資訊網路擁塞問題必需的方法。由於中英文混合檔案資料在國內外一般的應用相當普遍。本文提出雙區域調整串列對照編碼法(Double Adaptive List Correspondence, DALC)方法，屬於單一回合過程且適合應用於網路傳輸的文字檔案壓縮方法，基本的構想係建立在二維串列與參考區域原則，改進以往二維串列區域調整編碼法無法處理中英文混合資料壓縮的缺失。經由本文運用move-to-front原則及提出的group-move-to-front原則，在資料壓縮過程中建置兩個串列對照輔助結構並配合前導位置整數表示法，有效結合字母導向調整方式，完成了本文之資料壓縮方法。爲測試本文中提出壓縮方法的優點與可行性，本文針對同時出現中英文混合資料與較可能使用的常用字爲著眼考量，乃從時報資訊立即新聞稿中擷取測試資料，包括三種全英文、全中文及中英文混合等三種檔案資料。經過測試並比較壓縮率發現本文中所提出編碼方法之壓縮效益優於其他方法；同時本文方法亦具有容易製作的優點，將可實際應用於資訊網路。

1 緒論

隨著資訊技術發展一日千里，人類仰賴國際間資訊網路通訊的需求與日俱增，無可避免地面臨了各類龐大資訊充斥於網路上所造成擁塞現象。資料壓縮(data compression)的重要性不言而喻，尤其快速且適合應用於網路傳輸的資料壓縮工具更是不可或缺。如同其他東方語系國家一樣，普遍存在著中文與英文資料混合的一般文章及報導資料，因此就實用價值觀點，資料壓縮的研究必

須實際地探討中英文字混合檔案壓縮技術，不應再侷限探討缺少實用程度的完全英文或中文檔案壓縮方法。現行許多中文檔案壓縮方法是由英文檔案壓縮方法改良而來。設計文字壓縮技術在基本策略上有整體最佳(global optimum)與區域最佳(local optimum)兩種策略。所謂整體最佳策略係針對檔案整體全部作為編碼依據，而區域最佳策略則依據壓縮過程中按目前所讀取的檔案資料狀況作為編碼依據。因此若採行整體最佳策略，進行壓縮作業時，至少須經過二個回合(two passes)以上的處理過程，而區域最佳策略則僅需一個回合(one pass)處理過程。許多文字檔案資料壓縮研究透過上述兩種策略，運用各種例如樹狀結構(tree)、表格(table)、串列(list)以及統計運算(statistics)等資料結構來探討文字檔案資料壓縮方法。有關文字檔案壓縮技術分類如圖1.1表示。就實用觀點，壓縮過程讀取檔案兩個回合，且壓縮(compress)及解壓縮(decompress)皆須負擔前置成本(overhead)，相當不適合應用於網路傳輸。

為設計一個適合應用於網路傳輸的文字檔案資料壓縮方法，必須符合單一回合處理過程、壓縮及解壓縮過程簡單且快速等需求。因此本文選擇區域調整串列結構(locally adaptive list structure)作為中英文字檔案壓縮之探討方向。本文所提出的雙區域調整串列對照編碼法(Double Adaptive List Correspondence, DALC)，係針對原二維串列(two dimensional linked list)區域調整編碼法[14]無法處理英文檔案資料壓縮的瓶頸。本文嘗試結合字母導向(alphabet-oriented)調整方式，採用前導位置整數(prefix positional integer)表示法並配合BIG-5中文碼特性，一併解決了中英文混合檔案資料壓縮、前導位置整數編碼長度及控制串列長度等成本問題，並驗證本文所提方法的實用價值及壓縮效益。為驗證本文所提方法的可行性與優點，本文以三種檔案，包括全英文、全中文及中英文混合等形式，進行實證分析與比較，測試結果證實本文所提方法的可行性。

2 相關文獻探討

以下介紹與本文研究相關之文獻，並指出各方法之特性與缺點。

2.1 霍夫曼多群編碼法之中英文混合檔案壓縮方法

張真誠與蔡文輝[11]於1991年依據霍夫曼多群編碼法(Huffman multigroup coding)完成了中英文混合檔案壓縮方法。中英文混合檔案所指的英文是泛指英

文字母、阿拉伯數字、英文標點符號等（暨非中文字）。壓縮過程中須分成兩回合來進行壓縮。霍夫曼多群編碼法所具有的特性如下：

- (1) 霍夫曼多群編碼法屬於統計式(account)編碼法。
- (2) 壓縮過程是需經過二回合處理，無論編碼或解碼都必須先行建構相同的霍夫曼樹，且建立霍夫曼樹時必須計算字元(byte)出現的次數。
- (3) 出現頻率最高的位元組，其霍夫曼碼最短；而出現頻率愈低者則霍夫曼碼愈長。

而霍夫曼多群編碼法所具有的限制及缺點如下：

- (1) 加入一筆新資料於原始檔案中，則所有霍夫曼樹都必須重新建構。
- (2) 若應用霍夫曼多群編碼法於網路傳輸時，首先必須先傳輸至接收端所需解壓縮時建構與傳送端相同三個唯一H、L及E霍夫曼樹所需的中序順序(inorder sequence)及後序順序(postorder sequence)資料所耗費的前置成本，使得壓縮效益大打折扣。顯而易見此方法不適合應用於網路傳輸。

2.2 藍波-立夫-衛曲編碼法之完全中文檔案壓縮方法

本節將介紹張真誠與蔡文輝[12]陸續於1991年所提出另一個依據藍波-立夫-衛曲(Lempel-Ziv-Welch, LZW)編碼法之完全中文檔案資料壓縮方法。所謂完全中文資料檔案乃指檔案中的資料完全皆為中文字碼，暨中文字乃至於所有的阿拉伯數字、標點符號、英文字母及螢光幕上看不見的換行跳列符號(Carriage Return and Line Feed, CRLF)等皆視同中文字來處理。資料區分為高位元組群、低位元組群和換行跳列群等三個子群(subgroup)。藍波-立夫-衛曲編碼法之完全中文檔案壓縮方法所具有的特性如下：

- (1) 屬於代換式字典(dictionary)編碼法。
- (2) 壓縮與還原過程同樣需要額外建置成本(表格)，但藍波-立夫-衛曲編碼法較之霍夫曼多群編碼法有相當小的壓縮還原所需表格的前置成本，因此較之霍夫曼多群編碼法有相當不錯的壓縮效益。
- (3) 依據藍波-立夫-衛曲編碼法之完全中文檔案資料壓縮方法所處理的檔案性質必須限制所有的阿拉伯數字、標點符號、英文符號及空白符號等均需完全佔用兩個位元組。

2.3 單一串列區域調整英文檔案壓縮方法

Bentley等人[6]於1986年所提出英文檔案資料壓縮的方法係屬於將固定長度(fixed length)轉換成不固定長度(variable length)輸出的檔案資料壓縮方法[3]。所處理的檔案資料被視為空白(space)隔開的諸多文字串所組成的組合。這些被空白隔開的文字串(string)，在本文中皆視為節點(node)。主要是利用單一串列(single list)來儲存壓縮處理過的文字串，並且對於曾經出現過的文字串，當其再次出現時，便以一個整數來取代，藉以縮短輸出碼長度。Bentley所提出的區域調整資料壓縮方法有下列限制及缺點：

- (1) 串列長度不宜無限制地增長。
- (2) 舊字串出現的頻率與出現的時間間距(interval)機會影響壓縮效果。
- (3) 在實際應用上可能文字串間相隔好幾個空白，超過一個以上的空白部份也當做一個空白文字串來處理，造成還原失真(distortion)。
- (4) 將整數及實數等數字形態資料亦視同文字串來處理，由於這一類資料的重複出現次數低，影響壓縮效益。

3 雙區域調整串列對照編碼法(Double Adaptive List Correspondence, DALC)

本節提出改進陳信宏[14]於1995年所提出二維串列(two dimensional linked list)區域調整完全中文壓縮方法。所謂二維串列基本上可視為座標分別有直、橫兩個基準線分別經由指標(pointer)自啓始節點由上而下、由左至右地單一方向串連(link)。在本節定位為高位元組串列。主要將中文BIG-5碼之高位元組(high byte)與低位元組(low byte)予以分別開來存放，因此所有高位元組所存放的串列就稱之為高位元組串列(high byte list)。而以高位元組串列上每個高位元組節點下各自領頭(lead)串連一組串列，其每個節點則用來存放文字中具有相同高位元組之低位元組，其二維串列結構詳如圖3.1。高、低位元組串列之新舊節點係依(move-to-front, MTF)原則排列。其實move-to-front原則相當於作業系統(operating system, OS)中記憶體管理策略(memory management strategy)中最近時間內置換最少被使用(least recently used replacement, LRU)的頁置換(page replacement)策略運用模式。可使出現頻率高的文字大部份可落在串列的前半部，而出現頻率低的文字便逐漸被推移至後半部，以期使高出現頻率的文字能獲得較小的位置整數。但是move-to-front原則較著重

於舊字元重複出現的時間間距(interval)，暨重複出現的時間間距愈長，則輸出前導位置整數碼長度就愈長。

因此本文為期明白展示本文所提出彌補move-to-front原則缺乏對於最常出現的字元有最小編碼長度的掌握的限制，以及無法處理中英文混合檔案壓縮的不足與限制。另行建置一組與實施move-to-front原則串列一樣擁有相同內容且同步調整的串列。但此新建置的串列的字元排序調整係依照在壓縮過程中所有文字所累計的次數，由大至小，而同次數的群集中元素的排列，仍依照move-to-front原則，稱之為(group-move-to-front, GMTF)原則區域調整串列。藉以保證高次數出現率的文字能獲得最小的前導位置整數，二組調整串列在壓縮過程中相互各自同步調整，並自其中選擇最小位置整數碼的串列作為輸出。如此一直在壓縮過程中充份相互發揮“截長補短”作用以有效縮減位置整數碼的長度，這就是本文在這一節中所提出的雙區域調整對照串列(Double Adaptive List Correspondence, DALC)架構原則。

因此，本文嘗試在中文字二維串列及英數字串列同時各建置另一組擁有相同內容的調整串列。而此串列裡各個字元的排列係分別依照壓縮過程中各中文字之高位元組及英數字ASC II位元組被使用過累計的次數由大到小，並隨著資料內容的不斷地處理中而調整改變。換言之，存在著二組中文字二維串列及二組英數字串列。其中所不同的是，新建置的一組是以GMTF原則，係按照各字元出現次數總計由大到小，但同次數的字元元素群中則仍按move-to-front原則排列的串列。二組串列的內容相同，同步調整，目的是藉由二組串列進行比較哪一組串列被使用字元所得前導位置整數碼較小，便以此串列的整數碼輸出，藉此“截長補短”的作用，有效控制及縮短整數碼長度的成本。

至於如何縮短及控制串列長度成本問題。由於現行最普遍使用的BIG-5中文碼共包含13053個中文字，其中有5401個用字，而次常用則有7652個字，而5401個常用字的高位元組範圍落在A4至C6之間[2]。為避免常用中文字之高位元組串列長度不受其他次常用或罕用中文字混合其間而增長，進而分散串列長度成本。嘗試另建置一組新中文字二維串列，用來儲放其他高位元組不屬於A4至C6範圍之間的中文字。

綜合上述，本文提出DALC架構及BIG-5中文編碼原則，原英數字及中文字二維MTF串列各增加了一組GMTF輔助串列，於是基本上產生了下列八個串列調整條件的狀況，其分類如下：

- (1) 載入新字元於英數字MTF及GMTF串列。
- (2) 字元已存在於英數字MTF串列中。
- (3) 字元已存在於英數字GMTF串列中。
- (4) 載入字元於中文字二維MTF及GMTF串列。
- (5) 已存在於常用中文字MTF串列中。
- (6) 已存在於常用中文字GMTF串列中。
- (7) 已存在於不常用中文字MTF串列中。
- (8) 已存在於不常用中文字GMTF串列中。

本文使用三個位元自"000"至"111"來分別表示上述八個串列調整條件狀況作為固定長度辨識碼。在壓縮過程中除了辨識(pattern)前面佔用三個位元外，由二組MTF及GMTF串列相互比較而取其最小的值輸出。因此最後串列調整條件的表示及其輸出結果如下：

H 表示高位元組；
 L 表示ASC II值或低位元組；
 prefix()表示前導位置整數編碼函數；
 d_H 表示高位元組的位置整數；
 d_L 表示低位元組的位置整數；

- (1) 載入新字元於英數字MTF及GMTF串列。
 000 L
- (2) 字元已存在於英數字MTF串列中。
 001 prefix(d_L)
- (3) 字元已存在於英數字GMTF串列中。
 010 prefix(d_L)
- (4) 載入於常用或不常用中文字二維MTF及GMTF串列。
 011 $H L$ 或
 011 prefix(d_H+1) L
- (5) 已存在於常用中文字MTF串列中。
 100 prefix(d_H) prefix(d_L)
- (6) 已存在於常用中文字GMTF串列中。
 101 prefix(d_H) prefix(d_L)
- (7) 已存在於不常用中文字MTF串列中。

- 110 prefix(d_H) prefix(d_L)
(8)已存在於不常用中文字GMTF串列中。
111 prefix(d_H) prefix(d_L)

本方法之流程圖如圖3.2所示。詳細說明本流程執行之過程及各串列內容的變化，請參考文獻[4]。

4 實證測試與分析比較

為證實本文提出的雙區域調整串列對照編碼法的壓縮效益，並就霍夫曼中英文多群編碼法與二維串列區域調整壓縮方法兩種方法作比較，分別就全英文、全中文以及最後的中英文混合檔案資料，以彰顯本文所提出的壓縮方法的優點與可行性。

4.1 全英文檔案資料壓縮效益比較

在本節處理英文資料部份，所執行模擬比較的方法，除了二維串列區域調整編碼法及雙區域調整串列對照編碼法外，另包括在文獻探討中提到的霍夫曼中英文多群編碼法。選擇其作為比較的原因，由於霍夫曼中英文多群編碼法具備可處理英文資料壓縮能力，可與之比較以表現本文提出的壓縮方法在英文壓縮的效益。所使用的測試資料為本文檔案(text file)，測試資料取自微軟視窗(Microsoft Windows)軟體中各項英文註解檔案，資料檔案大小範圍由10K 位元組至74K 位元組分成五組，效益分析與比較是以位元組(byte)為單位。以其編碼後所需記憶空間大小作為依據，模擬結果數據如表4.1。而圖4.1為測試資料各方法所獲得的壓縮效益。圖4.2則為各方法所展現的壓縮率(compression ratio, CR)。所謂壓縮率定義係壓縮後所減少之位元數除以原始檔之位元數； $\text{壓縮率} = (\text{原始檔案位元數} - \text{壓縮後檔案位元數}) / \text{原始檔案位元數}$ 。壓縮率愈高，表示省略的資料量愈多，因此壓縮效益越佳。

由表4.1之數據可以說明二維串列區域調整壓縮方法在處理英文資料壓縮效益顯然不佳，證實單純使用字母導向調整方式之英數字串列，僅應用MTF原則在英文資料壓縮的做法並不是一個可行的構想。本文所提出的雙區域調整串列對照編碼法在英文壓縮效益表現平均有50%的壓縮率。主要原因在於壓縮過程中，由於MTF及GMTF 兩組串列相互發揮輔助作用，確保高出現率的字元隨著壓縮過

程中一直控制在英數字GMTF串列的前端位置，經由掌握高出現率的字母及數字有較短小的前導位置整數編碼長度。證實本文提出DALC編碼方法也可經由字母導向調整方式而應用在英文檔案資料壓縮。

4.2 全中文檔案資料壓縮效益比較

在本文中所處理的中文碼係採用為最為普遍應用的 BIG-5碼，參與執行模擬比較的方法如同在4.1節所使用的三種方法。所測試的性質係採完全中文檔案資料，資料來源取自民國八十二年二月十五日「時報資訊」立即新聞稿，並將其舉凡英文及數字等不屬BIG-5內碼資料皆予以剔除，再從中選擇五組作為測試。檔案範圍由130K至920K位元組，模擬結果如表4.2。各方法之間壓縮結果比較圖 (histograms) 如圖4.3。壓縮率比較圖如圖4.4。由壓縮率比較圖所顯示壓縮率的表現，最佳為雙區域調整串列對照編碼法，次為霍夫曼多群編碼法，再次為二維串列區域調整編碼法。由圖4.4中顯示，雙區域調整串列對照編碼法僅將二維串列區域調整邊碼法提升壓縮率約略10%，不若在英文資料壓縮率的差異程度。足可說明中文資料檔案之造字結構及分佈形態與英文資料檔案不同。中文是由固定雙位元組所組成，而英文則由固定的26個英文字母所組成。因此設計文字資料壓縮方法，必須考量其性質差異之處。

4.3 中英文混合檔案資料壓縮效益比較

最後在本節將探討中英文混合檔案資料壓縮效益的表現，在4.1及4.2節係個別針對全英文及全中文資料檔案在各種壓縮方法之比較。現就以中英文混合檔案作為測試資料，資料來源除4.2節之時報新聞稿外，另外加上4.1節英文資料，同樣選擇五組資料作為測試。檔案範圍自140K至990K位元組，同樣參考4.2節所使用壓縮效益比較標準。各方法在中英混合檔案壓縮模擬結果如表4.3。壓縮結果檔案大小比較圖如圖4.5。壓縮率比較圖如圖4.6。由壓縮率比較圖顯示，雙區域調整串列對照與霍夫曼多群編碼法同時提昇2%壓縮率，相對地二維串列區域調整編碼法則減少3%壓縮率，這說明二維串列區域調整編碼法在英文資料方面壓縮的表現影響到中英文混合檔案壓縮效益，壓縮率比較圖仍顯示雙區域調整串列對照編碼法的壓縮率最佳，其餘依次為霍夫曼多群編碼法、二維串列區域調整編碼法。

綜合以上分別就英文、中文及中英文混合檔案部份測試結果，證實雙區域調整串列對照編碼法由於增加GMTF原則串列分別掌握英文字元與中文BIG-5碼高位元組出現次數由高至低的排序，正與霍夫曼多群編碼法之中英文混合檔案壓縮方法中最高出現率字元擁有最短(小)碼長度原則相同，並且又搭配了最近時間內最久未使用的被置換之MTF原則，在壓縮過程中充份發揮相互「截長補短」作用。不僅提昇陳信宏所提出二維串列區域調整編碼法之壓縮效益外，並解決了無法處理中英文混合檔案壓縮的問題，增進實際應用的價值。

5 結論與未來發展方向

文字檔案資料壓縮目的在於減少文字資料中重複資訊，不論在儲存空間或網路傳輸成本皆有其探討的重要性。網際網路與人類生活緊密結合，更應突破純粹英文或中文資料檔案壓縮方法所欠缺實用價值或不利網路傳輸等瓶頸。本文選擇區域調整串列結構作為中英文字混合資料檔案壓縮方面探討對象，原因在於區域調整串列結構有結構簡單、單一回合處理、無其它壓縮方法中壓縮還原所需額外前置成本、壓縮還原過程簡單快速，且應用網路傳輸時，傳送與接收端可立即同步進行壓縮及還原等諸多優點。不同於以往諸多壓縮方法乃先將傳輸資料予以壓縮後，再將壓縮還原所需前置成本先送至接收端後，再將壓縮檔案予以解壓還原等多回合過程所耗費的成本，而大大降低其優越性與實用價值。

本文針對上述目標，修正陳信宏所提出完全中文資料檔案壓縮方法中二維串列區域調整結構，在第四章分別就英文、中文及中英文混合部份實證分析其壓縮效益。由測試數據可以清楚了解本文提出的DALC編碼方法經由同時掌握壓縮過程中「最常被使用」及「最近時間內最常被使用」兩項原則，分別在英文、中文及中英文混合資料壓縮測試結果都獲得相當程度的效果。尤其在英文資料壓縮部份，更提供了另一種不同於Bentley原先所使用單字導向調整做法，證實字母導向調整做法經由本文提出DALC編碼方法，同樣可以獲致壓縮效益的另一種探討方向。

未來因應中文標準碼之頒行，屆時在中英文混合檔案資料壓縮領域同時也提供了一個值得深入研究探討的課題，也是本文繼續探討的方向與目標。

參考文獻

- [1] 施威銘 (民七七), 「C語言實務」, 旗標出版有限公司, 台北。
- [2] 倚天資訊 (民七九), 「倚天中文系統使用手冊」, 倚天資訊股份有限公司, 台北。
- [3] 張真誠 (民八三), 「資料壓縮原理與實務」, 松崗圖書公司, 台北。
- [4] 朱賢武 (民八四), 「中英文字檔案區域調整資料壓縮方法之設計與製作」, 國防管理學院資源管理研究所碩士論文。
- [5] Abrahamson, David M.: "An adaptive dependency source model for data compression," *Communications of the ACM*, Vol. 32, No. 1, pp. 77-83.(1989).
- [6] Bentley, J. L., Sleator, D. D., Tarjan, R. E. and Wei, V. K.: "A Locally Adaptive Data Compression Scheme," *Communications of the ACM*, Vol. 29, No. 4, pp. 320-330.(1986).
- [7] Bell, Timothy C.: "Better OPM/L text compression," *IEEE Trans. Communications*, Vol. COM-34, No. 12, pp. 1176-1182.(1986).
- [8] Bell, T., Witten, I. H., and Cleary, J. G.: "Modeling for text compression," *ACM Computing Surveys*, Vol. 21, No. 4, pp.557-591.(1989).
- [9] Bailey, R. L. and Mukkamala, R.: "Pipelining data compression algorithms," *The Computer Journal*, Vol. 33, No. 4, pp. 308-313.(1990).
- [10] Chang, H. K. C. and Chen, S. H.: "A new locally adaptive data compression scheme using multilist structure," *The Computer Journal*, Vol. 36, No. 6, pp. 570-578.(1993).
- [11] Chang, C. C. and Tsai, W. H.: "A data compression scheme for Chinese-English characters," *Computer Processing of Chinese & Oriental Languages*, Vol. 5, No. 2, pp. 154-182.(1991).
- [12] Chang, C. C. and Tsai, W. H.: "A compression scheme based upon Lempel-Ziv method for chinese texts," *Journal of Computers*, Vol. 3, No. 2, pp. 1-10.(1991).

- [13] Chang, H. K. C. and Chen, S. H.: "Extended predictive data coding scheme for Chinese text Files," *Computer Processing of Chinese & Oriental Languages*, Vol. 7, No. 2, pp. 154-182.(1993).
- [14] Chang, H. K. C. and Chen, S. H.: " A locally adaptive coding scheme for Chinese text files," *Journal of Information Science and Engineering*, Vol. 11, No. 1, pp. 51-71.(1995).

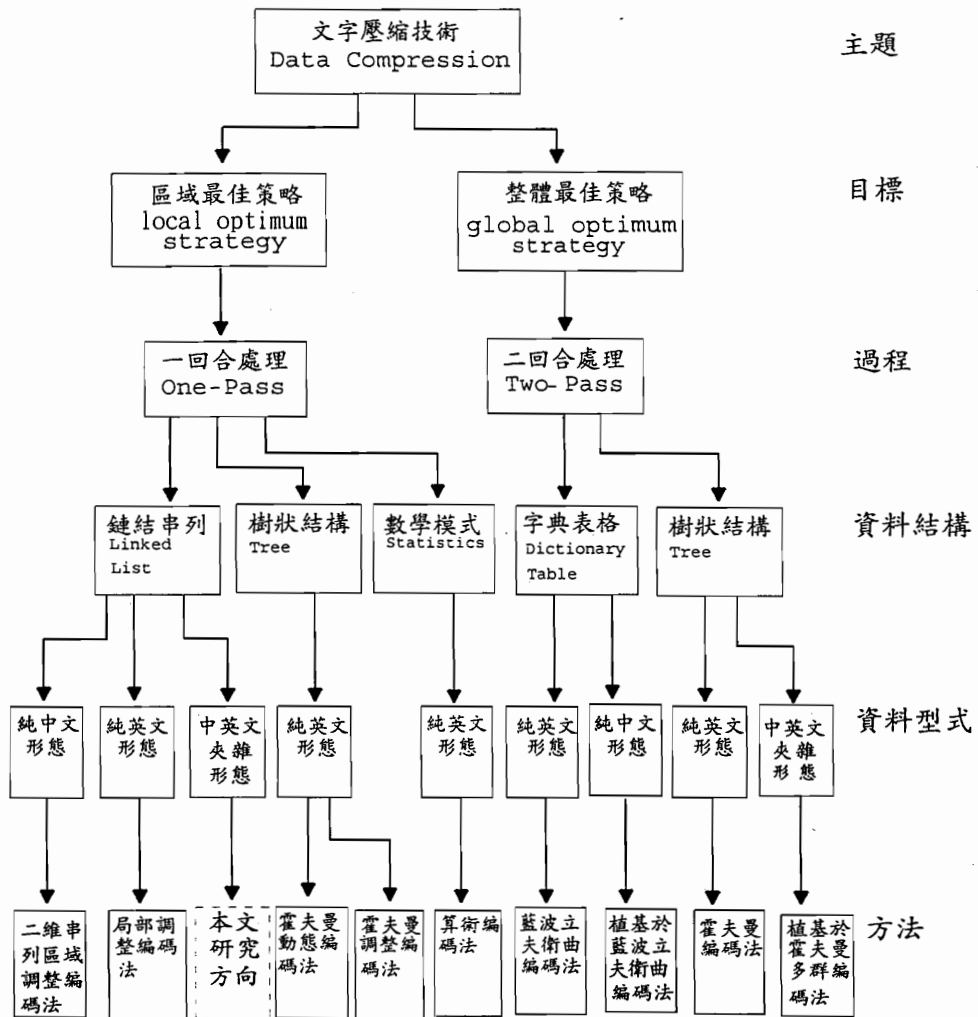


圖1.1 文字資料壓縮分類圖

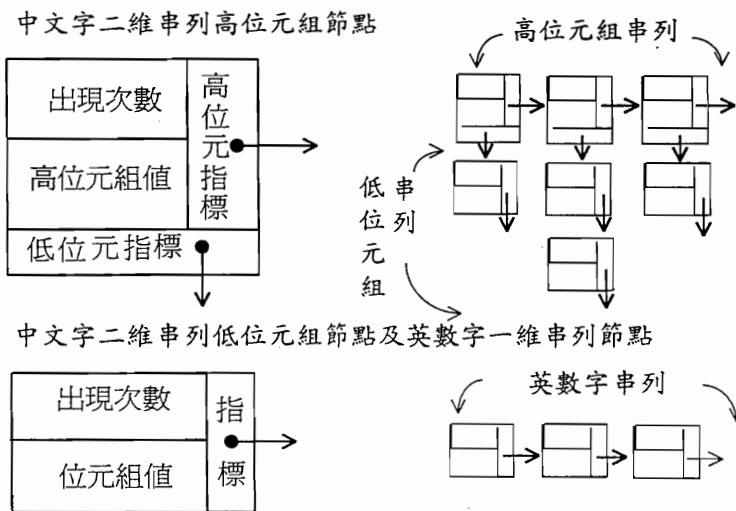
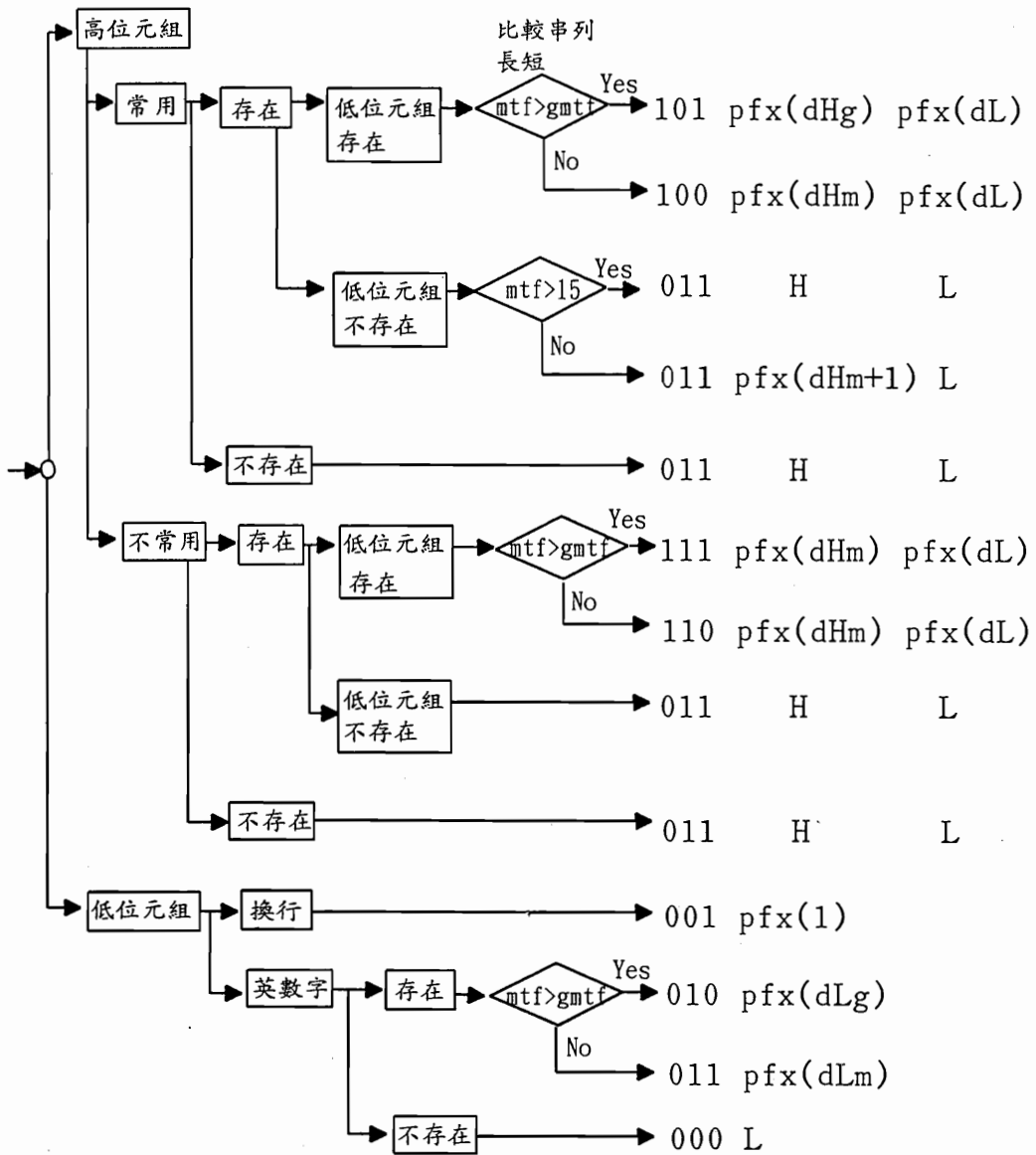


圖3.1 中英文字串列及節點結構



Note:

- pfx():prefix code
- dHg:integer position of high byte for gmtf list.
- dLg:integer position of low byte for gmtf list.
- dHm:integer position of high byte for mtf list.
- dLm:integer position of low byte for mtf list.
- H: high byte
- L: low byte

圖3.2 本方法之流程圖

表4.1 全英文檔案各方法壓縮結果比較表

測試檔案長度 (bytes)	11,614	20,950	32,947	43,431	76,379
二維串列區域調整 整編碼法	14,066	25,364	39,901	52,592	92,490
霍夫曼中英文多 群編碼法	6,866	12,131	19,110	25,626	43,536
雙區域調整串列	5,796	10,412	16,347	21,531	37,828

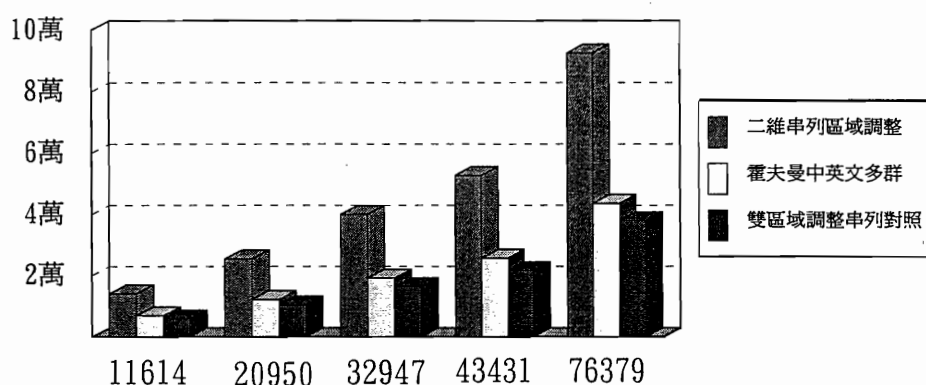


圖4.1 全英文檔案各方法壓縮結果比較圖

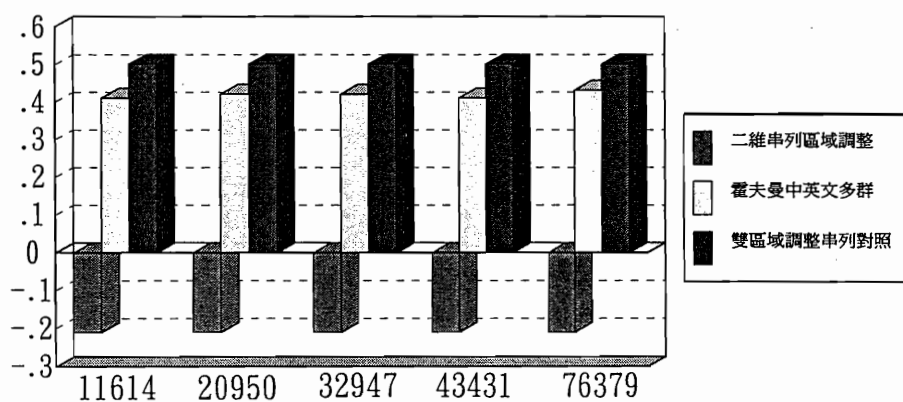


圖4.2 全英文檔案各方法壓縮率比較圖

表4.2 全中文檔案各方法壓縮結果比較表

測試檔案長度 (bytes)	134,885	269,771	404,656	539,541	944,196
二維串列區域調整編碼法	108,309	216,805	325,300	433,759	759,279
霍夫曼中英文多群編碼法	101,165	199,641	295,344	393,871	689,262
雙區域調整串列對照編碼法	95,773	191,548	287,319	388,466	679,781

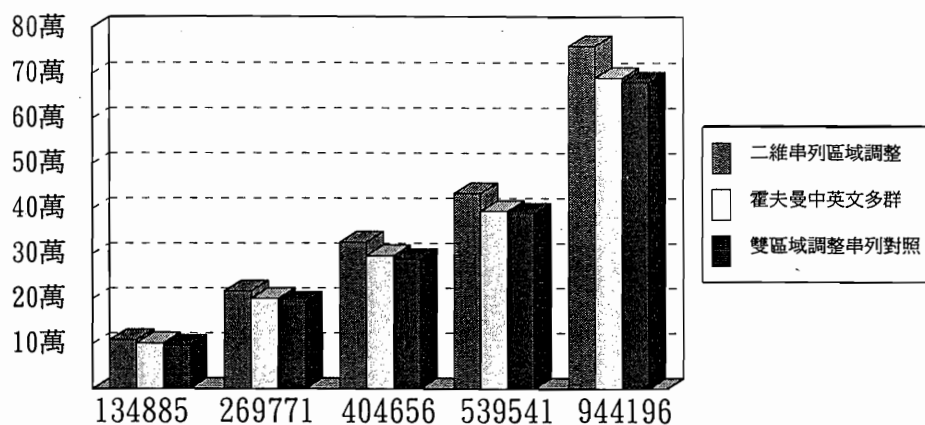


圖4.3 全中文檔案各方法壓縮結果比較圖

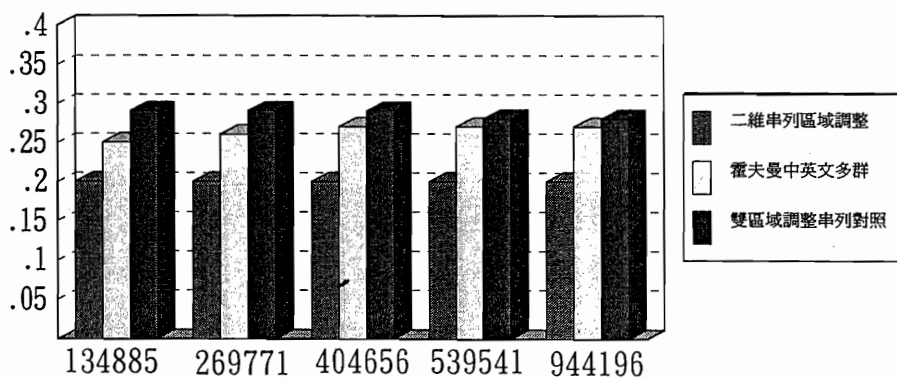


圖4.4 全中文檔案各方法壓縮率比較圖

表4.3 中英文混合檔案各方法壓縮結果比較表

測試檔案長度 (bytes)	146,499	290,721	437,603	582,972	1,020,575
二維串列區域 調整編碼法	121,596	241,323	367,599	489,687	857,322
霍夫曼中英文 多群編碼法	106,950	209,323	315,161	419,808	724,616
雙區域調整串	101,088	200,599	301,940	402,251	714,404

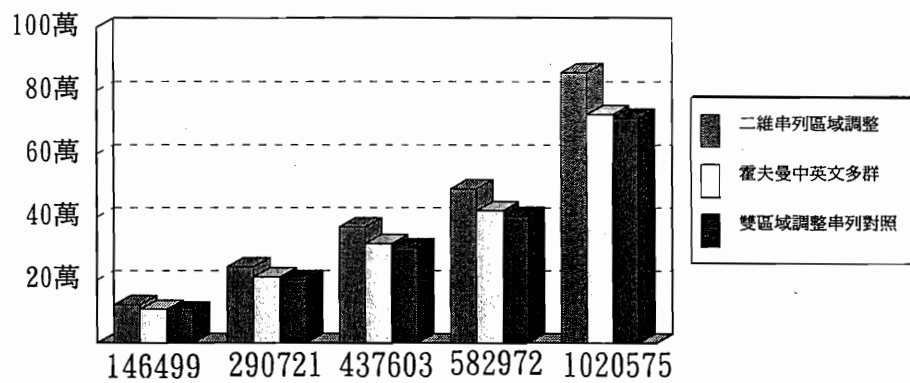


圖4.5 中英文混合檔案各方法壓縮結果比較圖

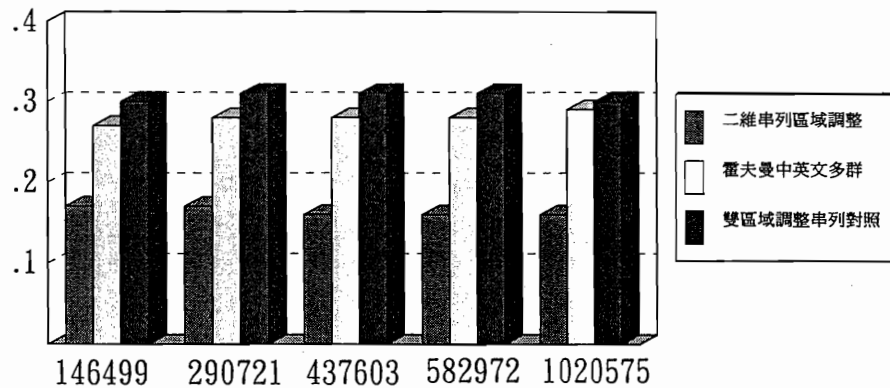


圖4.6 中英文混合檔案各方法壓縮率比較圖