從計算機處理食貨志的經驗

談計算機處理史籍的問題

作者：謝清俊

中華民國台灣省
中央研究院計算中心

中華民國七十五年二月

一、前　　言

史籍自動化計畫是一個長期的學術研究計畫．其目的在探索計算機應用於文、史工作的可行性．在此計畫之前，國內以計算機處理文、史資料並不多見．所以，在台灣，這計畫算是一種新的嚐試．

由於文、史資料多為文獻形式，而非格式化的紀錄，是故此計畫在本質上將面臨一個全文(Full-Text)的環境，而全文的表示法(Represen--tation)、檢索的方式、人機的介面，等等都將是此計畫的面臨問題．再者，文、史方面的研究一向少列後備經費，所以如何開發低成本的系統，譬如在 PC 上，也是此計畫考慮的要點．

此計畫始於 1984 年 7 月，至今已一年有半．在第一年中，我們做了個很好的示範－－可行性的探索．詳細的情形請參閱"史籍自動化食貨志輸入電腦第一年總報告"．目前第二年已進行過半．在第二年中的主要目的是：

1．完整地建立廿四史中食貨志部份的全文查詢系統並對第一期計畫作些改進．
2．探求語文處理能力，尤其是斷詞、斷句、構詞分析等對史籍處理的協助．
3．探求計算機輔助之索引建立方法，史學研究工具之自動化，以及可利用之新技術對此系統之影響．並以此作下年度計畫之基礎．

由於第二期計畫尚在進行中，以上各點之工作並未告一段落，所以本報告將以第一期計畫為基礎，並將第二期計畫中改善之各點向各位報告，以期達到互相砌磋的目的．本文中之各項資料結構及檔案結構，以及系統架構等等，以後均可能會改變，這也是由於尚未定案之原故．

以下就是此計畫年度的工作概述，以及與第一期中不同的改進之處的報告．

二、字集與資料登錄

此計畫的第一年是在 IBM 5550 計算機上做的，到第二年又將此系統轉移至 3B/2 和天龍 570 的組合系統上．IBM 5550 和天龍 570 有很類似的漢字集合，均有約一萬三千字左右，但是此字的集合對食貨志而言是不夠的．在第一期中，我們造了 250 個字，到了第二期又加多了 41 個字．總共有 291 字不在字集之中，約佔原字集的 2.24%．比例是相當高的．因此，時下商用的中文資訊處理系統，其字彙不足以處理史籍，已獲得有力之佐證．關於各書之字數請閱表一．所造之新字請閱表二．

在此計畫中，我們訓練了一批資料輸入人員，他們都是專科程度的，由完全不會學起，所以輸入及校對速度不會很專業化．平均的速率，包括輸入與校對，約每分鐘 12.2 字．其中約 2/3 的時間是輸入，1/3 是校對及修正．校對採三人七校制．可是在執行上，發現五校已相當好了，不需校到七次之多．在末期的平均輸入速度已超過每分鐘 25 字，校對及修正的速度在每分鐘 40 字以上．也就是每分鐘的有效輸入約在 15.4 字，或每小時約 920 字．

除正文外，計畫中還建立了一些機讀式的對照表．包括人名字號表、年代對照表，地名表、職官表等，以及一些專業名詞表等等．這些表格的輸入速率約同於正文，可是錯誤率較低，大致說在校對及修正時，三次就已足夠．

表一：食貨誌各書字數表

書　　名	字　　數
漢　書（上、下）	３６４１１
史　記	１４３４３
舊　唐　書（上、下）	２４８０７
唐　書（一至五）	３１３８０
隋　書	１２０５３
魏　書	１１６１５
晉　書	１１０８５
舊　五　代　史	８８１０
遼　史（上、下）	４９５６
金　史（一至五）	５２３２０
宋　史（上、下）	２１４９９４
元　史（一至五）	６８５３３
明　史（一至六）	６８５１３
共　　計	５５９８２０

表二．一 史籍自動化第一期所造新字

00 區

01	茭	32	杞	63	沂
02	豣	33	盜	64	滕
03	[illegible]	34	勑	65	賁
04	洴	35	攢	66	[illegible]
05	卹	36	輕	67	夏
06	隮	37	[illegible]	68	偪
07	殽	38	[illegible]	69	[illegible]
08	[illegible]	39	叓	70	莫
09	彻	40	荼	71	阜
10	鄹	41	尅	72	[illegible]
11	甄	42	敕	73	[illegible]
12	羅	43	檝	74	[illegible]
13	罹	44	叀	75	廚
14	滎	45	[illegible]	76	[illegible]
15	涯	46	栗	77	耨
16	异	47	渠	78	亂
17	鉄	48	粟	79	[illegible]
18	批	49	矧	80	[illegible]
19	豊	50	櫓	81	[illegible]
20	趍	51	緊	82	[illegible]
21	竈	52	鹺	83	監
22	俊	53	污	84	呪
23	叶	54	亘	85	[illegible]
24	譆	55	磿	86	鉀
25	沴	56	[illegible]	87	脉
26	[illegible]	57	迴	88	坂
27	汜	58	[illegible]	89	邑
28	决	59	[illegible]	90	[illegible]
29	賒	60	郝	91	緡
30	牀	61	穎	92	詩
31	[illegible]	62	須	93	雙
				94	[illegible]

01 區

01	[illegible]	32	葱	63	毅
02	鐀	33	稆	64	梳
03	鄜	34	[illegible]	65	元
04	閻	35	裸	66	[illegible]
05	殺	36	毀	67	[illegible]
06	[illegible]	37	[illegible]	68	効
07	棧	38	[illegible]	69	纊
08	堯	39	鷹	70	[illegible]
09	穽	40	賀	71	[illegible]
10	[illegible]	41	[illegible]	72	[illegible]
11	搊	42	[illegible]	73	[illegible]
12	[illegible]	43	德	74	[illegible]
13	晉	44	微	75	[illegible]
14	[illegible]	45	染	76	[illegible]
15	却	46	錄	77	冗
16	恒	47	[illegible]	78	[illegible]
17	沉	48	珏	79	福
18	沿	49	迹	80	垍
19	脚	50	[illegible]	81	[illegible]
20	諫	51	齋	82	[illegible]
21	祗	52	[illegible]	83	[illegible]
22	强	53	綫	84	[illegible]
23	晧	54	鈎	85	[illegible]
24	[illegible]	55	泣	86	涉
25	[illegible]	56	[illegible]	87	邠
26	猪	57	瑤	88	[illegible]
27	雞	58	匀	89	弧
28	[illegible]	59	[illegible]	90	腥
29	糇	60	叙	91	[illegible]
30	[illegible]	61	[illegible]	92	[illegible]
31	頸	62	[illegible]	93	[illegible]
				94	[illegible]

02 區

01	韶	32	硝
02	昆	33	[illegible]
03	煊	34	[illegible]
04	[illegible]	35	吕
05	[illegible]	36	[illegible]
06	[illegible]	37	[illegible]
07	[illegible]	38	[illegible]
08	鬪	39	[illegible]
09	冲	40	敔
10	[illegible]	41	[illegible]
11	[illegible]	42	[illegible]
12	[illegible]	43	[illegible]
13	謁	44	[illegible]
14	淵	45	敘
15	[illegible]	46	[illegible]
16	吴	47	幺
17	[illegible]	48	[illegible]
18	[illegible]	49	[illegible]
19	埒	50	鉛
20	[illegible]	51	勢
21	蔕	52	刃
22	[illegible]	53	匡
23	[illegible]	54	[illegible]
24	[illegible]	55	斥
25	哥	56	[illegible]
26	[illegible]	57	[illegible]
27	[illegible]	58	[illegible]
28	[illegible]	59	[illegible]
29	[illegible]	60	切
30	[illegible]	61	[illegible]
31	[illegible]	62	[illegible]

表二.二 史籍自動化第二期所造新字

1. 醱	21. 腭
2. 蠹	22. 趴
3. 鏁	23. 䲛
4. 羨	24. 开
5. 碙	25. 壇
6. 鷔	26. 橅
7. 廢	27. 垾
8. 栿	28. 懇
9. 蕒	29. 憯
10. 乣	30. 嵒
11. 賈	31. 迉
12. 弆	32. 弛
13. 逥	33. 峦
14. 隸	34. 离
15. 蓊	35. 哆
16. 簏	36. 蹊
17. 欸	37. 廏
18. 肥	38. 飱
19. 稉	39. 慎
20. 碣	40. 鮰
	41. 栗

三、 資料結構

在第一期中，全文資料是以行、頁、段為單位分割儲存的．以書名將相關的頁段串連成為整個全文檔案．這樣做的好處是定長式的檔案結構，較易處理．可是缺點很多，第一，它與版本關係太密切，其次要分割句子就比較麻煩．譬如，要顯示一句或一段時，經常跨行、跨頁，在處理上必須增加磁碟機讀取的次數，而且必須做全文掃描以取到句子或段．第一期所使用的資料結構並不好，其詳細的描述請參閱史籍自動化食貨志輸入電腦第一年總報告中系統設計部份之第四頁．

在第二期中，我們將資料結構做了一個相當大的修改，主要的觀念是：盡量將之模組化、結構化、通用化．這樣做的目的是探索一個通用的結構，希望它能適用於許多史籍．究竟這麼做的優劣如何，還有待進一步的測量和比較．

在本期中，主文(全文)的資料結構是一個樹狀結構．此樹是依正文的文章架構而設計的，請參閱圖一．圖中，終點是文章之段．譬如：在食貨志系統中，其源根(Root Node)是食貨志，而第一階是表一中的各書，而各書又分上、下或冊數，而每冊下再分段．是故在此食貨志系統中，它是一個五階的樹．

在此樹中，每一階之節點(Node)均有一個控制域(Node Control Block 簡稱 NCB)以描述其共通的結構部份，NCB 之資料結構如表三．

在此樹中與某特定史籍資料之關係則以一個描述域(Node Descriptor) 敘述之，再由此描述域連到正文和使用者的卡系(Card System)去，以達成連繫使用者的特定應用系統的目的，請參閱圖二之結構．ND之資料結構如表四．

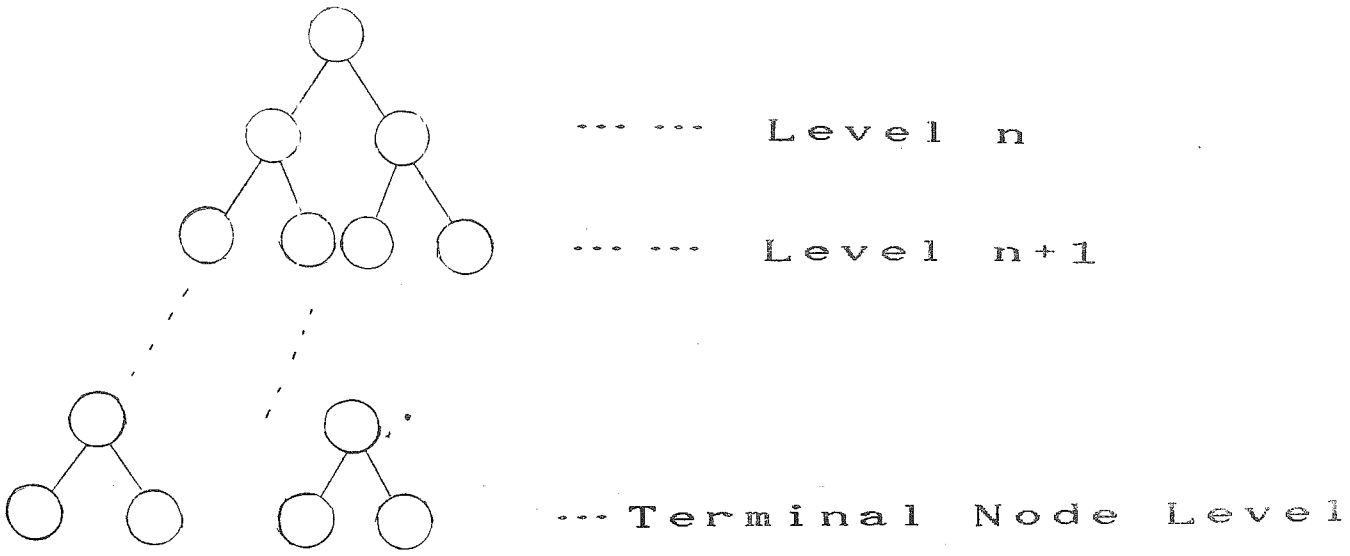
本文檔是直接存取檔，以段為單位，為本系統中最小的全文單位．若

是使用者只要查到句子，那麼，就必須掃描過此段全文，方可摘出相關的句子．在鼎文版的食貨志系統中，句子是依表五的區別符號(Delimiters)而偵知的．

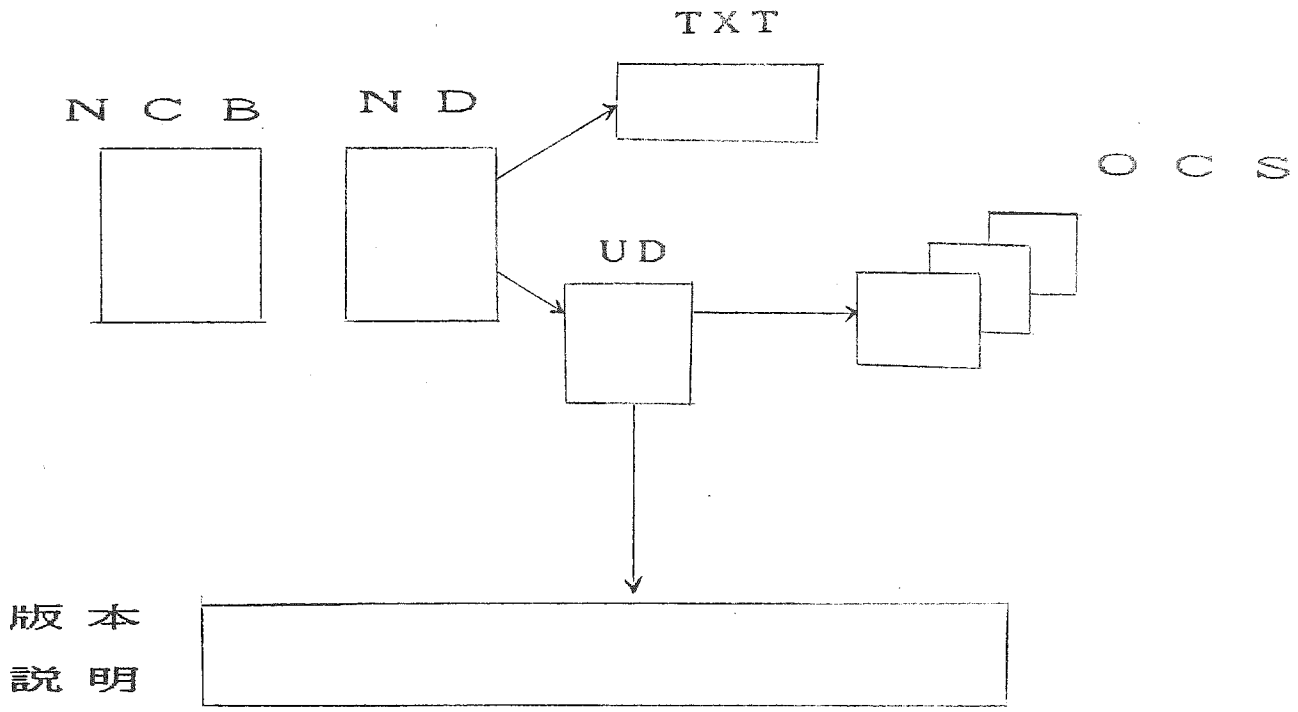
關於頁和行的部份，則另以一頁及行所構成的樹指向本文檔．若是使用者不需參照原版本之頁行，則此部份可不需用到．此部份的描述是僅與版本相關的．

關於卡系部份，容再做一說明．在卡系中，使用者可以建立他的心得卡片；包括他私人建立的索引，鍵語(Key Words)，眉題摘要，以及他的心得．換句話說，當它閱讀史籍時可隨時建立閱讀的心得卡系．此卡系建好後可以透過一些工具程式，如文字處理、統計、編排列印以及建立相關的鍵字表和索引檔等等程式，作為產生文獻之基礎．此部份的工作，已超越了查詢的範圍，且尚未發展完整，故在此僅從略．

關於鍵語之例檔(Inverted Files)則與一般索引例檔相似，在此不詳敘述．



圖一、資料結構系統示意圖



圖二、資料結構中節點之架構

表三：ＮＣＢ之資料結構

ＮＣＢ(Node Control Block)的結構：
描述 Node 之間的邏輯結構，如指標的連結、層級等等．

欄　位	位元組數	說　　明
1	8	Node ID.
2	1	Node Level，表示在那一層．
3	1	下一層 Node 的數目．
4	1	出生序．
5	8	If Node Level=1 可指向 Super Node ID. If Node Level=8 可指向 Under Node ID

表四：ND之資料結構

ND(Node Descriptor)的結構：
描述 Node 本身的固定資料如名稱、本文檔名稱等等.

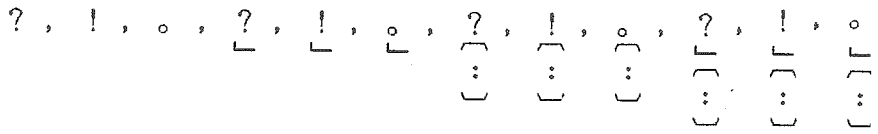
欄位	位元組數	說明
1	128	Node 的全名.
2	16	Node 全文檔名.
3	2	保留
4	1	指向使用者描述檔(UD).

UD之資料結構

UD(User Descriptor)的結構：
描述使用者所需的資料

欄位	位元組數	說明
1	1	在版本1那個 Record.
2	1	在版本2那個 Record.
3	16	第一個使用者全名.
4	1	指向第一個使用者卡系.
5	16	第二個使用者全名.
6	1	指向第二個使用者卡系.
7	1	Overflow Record(如版本超過一種，使用者超過二人)

表五：偵測句子的符號組合



四、建檔工作

在第一期的工作中，花費了許多歷史學者的時間和精力來標註所有的索引．這樣做當然沒有充份利用計算機的能力．然而這麼做有其必要．它可以提供一組完全正確的檢索，這對以後的研究工作大有助益．若是我們想用計算機產生此索引，其正確性之測量必須要一組完全正確的檢索作為比較之準繩．

在第二期中，所有的索引檔是以計算機輔助的方式來建立的．首先，一組鍵語表先建立起來，然後再用全文掃描的方式，輔以斷詞及少許語文處理的技巧建立索引例檔．此索引例檔再以人工去查核，一則可以測出由機器做索引之正確程度，二來可以分析缺點所在，作為以後研究之對象．到目前為止，此工作仍在進行中，具體之結果要等本期計畫完成後才知分曉．

五、語文處理能力與本計畫之關係

如上節所述，在本計畫中涉及到些語文處理的問題．目前主要用到的語文處理技術有斷詞和斷句，以及構詞的分析等．

關於斷詞部份，就已發表的研究而言，已經可以付諸實用了(不是百分之百的正確，但配以一良好之編纂系統已經可省不少人力)．要言之，鍵語首先組成樹狀結構，然後在掃描時，找出所有可能之組合，再經設定之規則除去其蕪雜的部份．

致於斷句部份，則很單純，照表五之符號以常用之文法架構(Syntax Directed Structure)處理就可以了．

較複雜的部份是構詞的分析．此計畫中構詞的分析集中於名詞片語部份，而且已是名詞片語中的一部份，若包含子句、動詞片語等成份的片語則不在吾等處理之內．換言之，所處理之名詞片語多半為名詞之複合詞及數量詞之複合詞．此部份之詞彙結構僅限於用文法(Syntax)可表示且只需知道文法架構就可分析出來結果的．如：第五章，第四十二卷，萬曆五年，秦王政三十七年．公元一九００年等等．

構詞分析應用在二處．其一是在欄位結構化的資料中對某欄作分析．其目的是保留該欄的彈性，以期適應更多樣的文獻．其二是應用在年代轉換的服務程式之中．

在此系統中，我們有意將許多正史的工具書逐一地變成機讀格式(Machine Readable Form)．在此期中，主要的實驗對象是年代轉換．在年代轉換的程式中，前後約涵蓋三千五百年[中國歷史紀年表，1983]．在此期間，我們將各時代、朝代以及其所含之政權之帝王、帝號、年號，對之西曆等等建為機讀格式，並提供一組查詢語言以利與其他系統結合．在此語言中，對各年代的表示法則必需以構詞分析去了解．

此部份的研究，亦將在本期計畫結束時，始有定論．然而，我們已經肯定的是：語文處理能力和建立機讀式的研究工具和資料是絕對相關的．要建立些專家系統(Expert Systems)或智慧系統型式的輔助工具並非一蹴可及．然而簡單如一些只需文法架構就可分析清楚的構詞處理，已對史籍的全文處理有很大的幫助了．

六、系統之構想

由於此研究的目的是利用計算機協助文史的研究工作，我們研究的題目並不局限於全文檢索．此研究的最後目標可能是一個供文史研究人員使用的工作站(Work-Station)，可是目前還有許多基本的問題未能解決，還有很長一段路要走．然而在此過程中卻充滿研究的挑戰和樂趣．

在這樣的環境下，我們不敢說用由上而上的方式(Top-Down)作整體的規劃和設計，我們只能認清一些基本問題並逐一去克服，以希望能逐漸改變技術環境，使之成熟到可以著手去設計上文所說的工作站．因此，在目前，我們只有一個系統的構想，或可說是：為把握未來方向的一些體認，作為我們考慮的原則．茲將這些原則分述如下：

(一)研究文、史所用的工具書必須逐個自動化．智慧型的工具(Knowledge-Based Tools)亦很重要，譬如：族譜、職官架構、中西歷法、…… 等等的自動化都是值得研究的題目．

(二)語文處理能力必須加強，以提供更好的人機介面和有效的利用自動化的工具書．

(三)輔助研究工作的工具程式，譬如：文字處理、電子卡系、統計程式、私人全文檔案等等均待開發．友善的查詢語言(User Friendly Query Language)，對使用者很重要．

(四)通用的全文結構(Full-Text Schema)及處理的方法有待研究．又此系統應將公用的全文資料庫與私用的分別處理，以便做到 PC 可處理的程度(Reduce Operating Data Base Size)．

(五)史籍的全文資料比較是靜態的，雖然處理上較方便(Batched Update)，然而檢索的技術還是很重要，值得研究，Free-Text Search 的技術亦然．檢索的自動化(Auto-Indexing)是必然的趨勢．然而亦可利用使用者私人所做的檢索加強系統的檢索能力(檢索之學習模式)．

－ 全文完 －

－ 請指教 －