CA Condensates를 위한 KORSTIC SDI 檢索시스팀

柳 慈 卿*

1. 序 言

데이터 베이스의 確立 및 프로파일의 作成과 함께 컴퓨터 情報檢索에 必要한 또 하나의 要素가 檢索시스팀이다. 檢索시스팀은 데이터 베이스 관리시스팀 (Data Base Management System)으로 그림 1에서와 같이, 프로파일에 나타난 利用者의 要求를 읽고 願하는 情報를 데이터 베이스에서 찾아내어 一定한 形態로 提供한다. 즉 이것은 利用者를 데이터 베이스에 接近시켜 주는 必要不可決한 過程인 것이다.

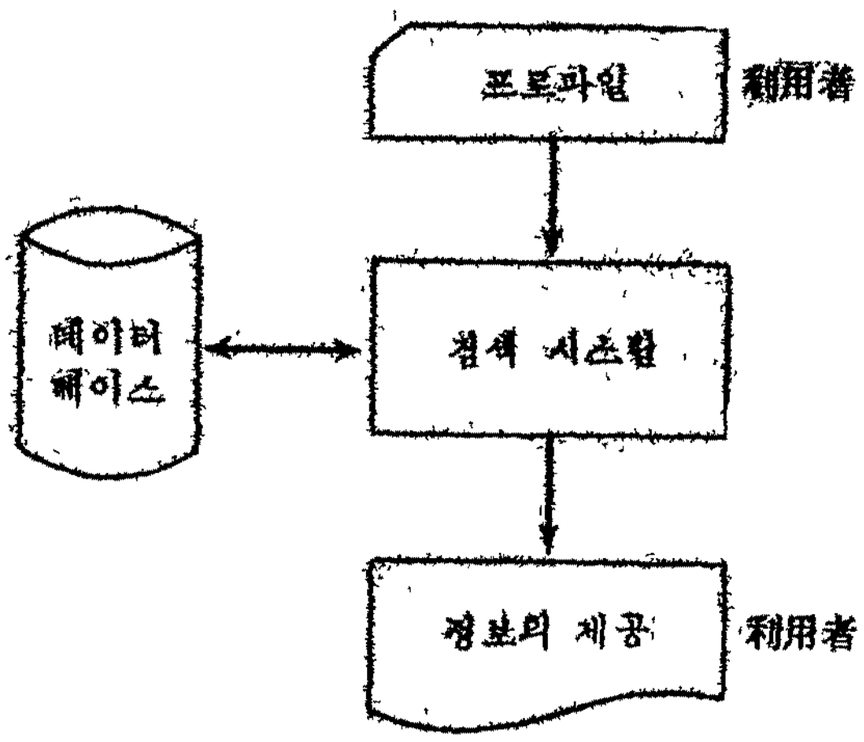

그림 1.

KORSTIC에서는 1975年 1月부터 美國化學會로부터 Chemical Abstract의 書誌的 事項과 抄錄 代身에 重要語(Keyword)가 包含되어 있는 CAC(Chemical Abstracts Condensates)라는 데이터 베이스를 購入하여 分類別로 檢索하여 提供하여 왔다. 그러나 이와 같은 分類에 의한 檢索은 한 分類에 包含되는 主題의 範圍가 넓어, 많은 不必要한 文獻이 함께 檢索되는 不便을 주어 왔다. 7月부터는 分類別뿐 아니라 重要語나 著者名 등, 利用者가 要求하는 事項別로 檢索이 可能하게 된 바, 이 글에서는 이 檢索에 使用될 시스팀에 대하여 說明하고저 한다.

*KORSTIC 電子計算室

2. 시스팀 概要

이 시스팀은 CAC의 書誌的 데이터가 들어 있는 파일과 利用者의 프로파일로 構成된 파일을 比較하여, 各 프로파일別로 그 프로파일의 論理를 만족시켜 주는 레코오드를 찾아내어 提供하는 役割을 한다. 濠洲 CSIRO에 의해 1972년에 開發된 이 시스팀은 磁氣테이프 1卷, 카아드로는 12,000餘枚에 該當하는 尨大한 것으로, 컴퓨터는 CSIRO에서 使用되는 것과 같은 機種인 KIST에 設置된 Control Data의 CYBER-73을 使用한다.

일이 處理되는 段階別로 獨立된 프로그램으로 나누어진 이 시스팀은 모두 9個의 프로그램과 60餘個의 Sub-routine으로 構成되어 있으며, 이 9個의 프로그램중 CDC의 Assembly Language인 COMPASS로 짜여진 한개의 프로그램을 除外하고는 모두 FORTRAN으로 되어 있다. 또한 이 시스팀은 CAC뿐만 아니라 BA PREVIEWS, INSPEC 등 모두 10餘個의 데이터 베이스를 프로그램의 修正없이 處理할 수 있는 互換性이 높은 시스팀이다. 현재 KORSTIC에서 蒐集하고 있는 原子力 關係의 데이터 베이스 INIS도 약간의 프로그램 修正으로 이 시스팀을 통한 SDI서어비스가 可能할 것으로 보인다.

3. 構成프로그램

시스팀을 構成하고 있는 프로그램 및 이들 프로그램의 相互關係는 그림 2에 表示되어 있다.

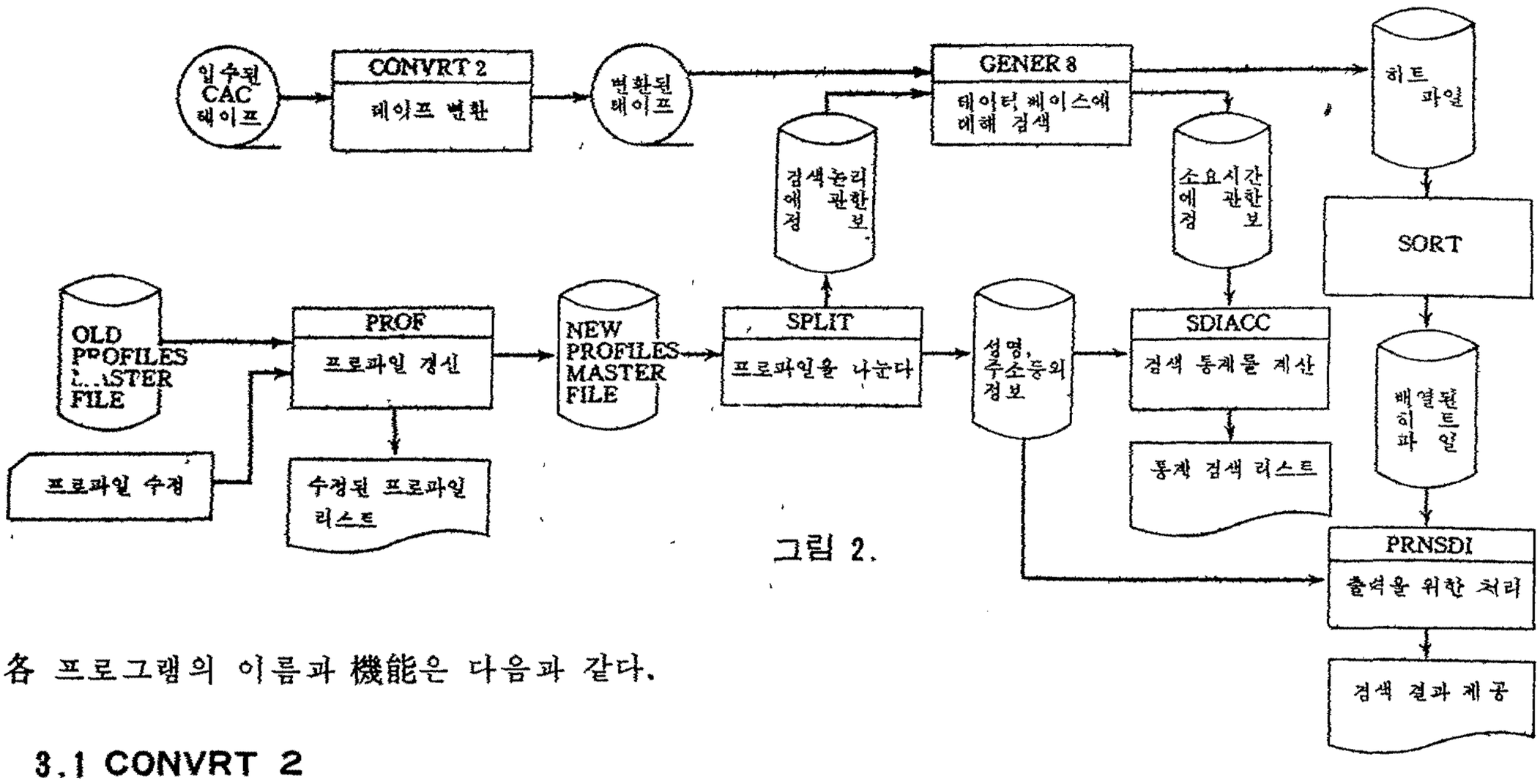

그림 2.

各 프로그램의 이름과 機能은 다음과 같다.

3.1 CONVRT 2

美國化學會에서 購入한 CAC 테이프로부터 레코드를 읽어 Standard Format로 바꾸는 프로그램이다. 購入된 테이프는 IBM 機種에 맞도록 만들어진 것이므로 CDC를 使用할 때에는 많은 問題點이 있다. 우선 IBM에서 使用하는 文字의 內部코오드와 CDC에서 使用하는 文字의 內部코오드가 다르며 IBM에서는 32bit 즉 4 Byte가 1 Word인 반면 CDC에서는 60bit 즉 10Character가 1 Word를 構成한다. 그러므로 이 프로그램에서는 코오드의 變換뿐 아니라 機械檢索의 效率을 높이기 위한 레코드의 變換도 包含하고 있다.

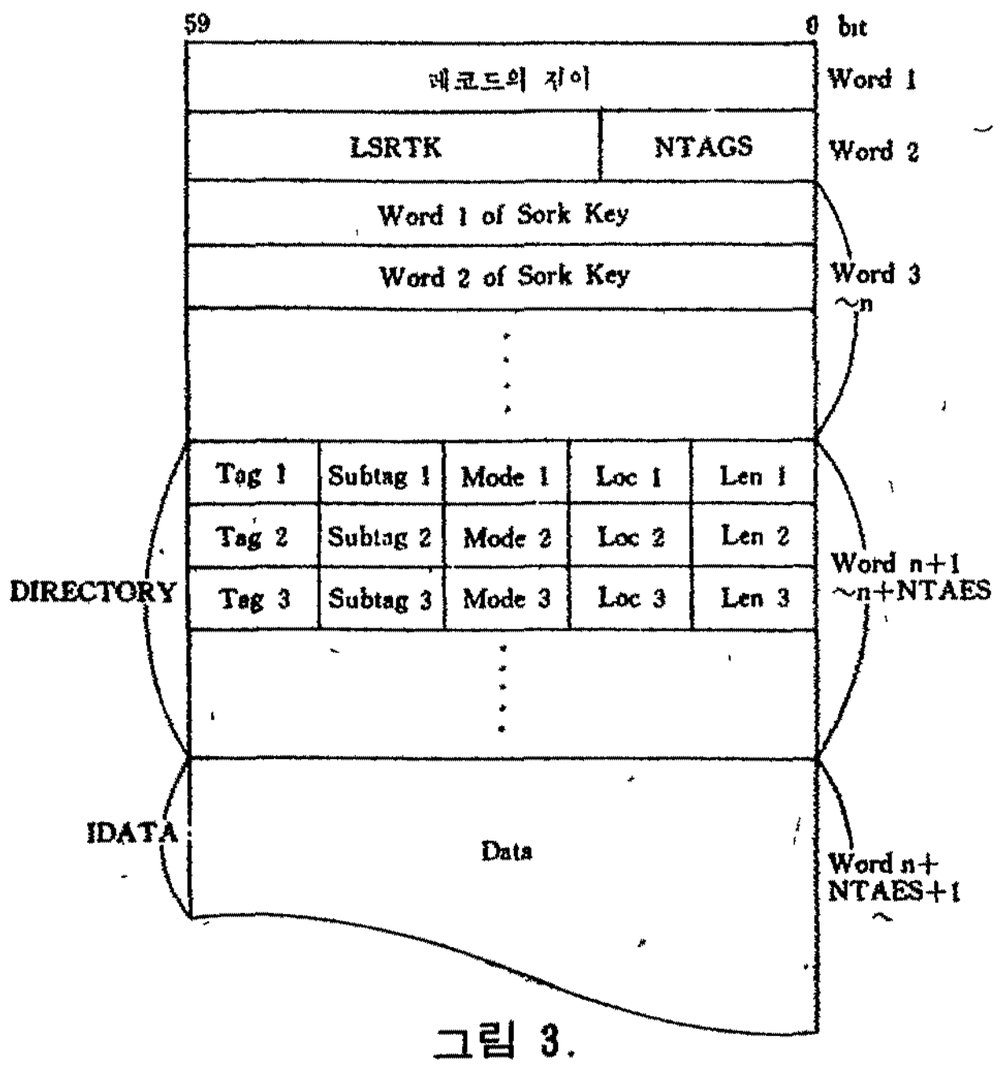

그림 3.

變換된 데이터 베이스 테이프의 레코드는 다음과 같이 構成되어 있다.

各 Word에 包含된 內容은 다음과 같다.

① Word 1 : IDATA의 길이가 Word數로 表示되어 있다.

② Word 2 : bit 0～17은 레코드안에 包含된 tag의 數를 表示하며 bit 18～59는 Sort에 使用될 Sort-Key의 길이를 表示한다.

③ Word 3～n : 實際 Sort-Key를 包含한다. 여기에서 n=3+(LSRTK−1)/10.

④ Word n+1～n+NTAGS : Directory를 構成하고 있다. Directory內의 各 Word는 하나씩의 項目을 나타내고 있으며 各 Word의 內容은 다음과 같다.

○ Bit 59～48 : 項目의 Tag를 包含한다. Tag란 各 項目에 주어진 固有番號를 말한다. 例를 들면 Standard Format에서 著者名은 21, 標名은 31, CA分類番號는 103 등으로 表示되고 있다.

○ Bit 47～36 : 項目의 Subtag를 말한다. Subtag란 tag를 修飾하는 部分으로 著者가 두명인 경우, 첫번 著者의 Subtag는 1이 되고 둘째번 著者의 Subtag는 2가 된다.

○ Bit 35～24 : 項目의 Mode를 表示한다.

現在는 다음의 4가지가 있다.

6 : integer Variable

7 : Character item (BCD Mode)

8 : Floating-point Variables

9 : Arbitrary Bit String

○Bit 23~12 : IDATA內에서 項目의 位置를 나타낸다. 새 項目은 새 Word에서 시작되는 것을 原則으로 한다.

○Bit 11~0 : 項目의 길이를 Character의 數로 表示한다.

⑤ Word n+NATES+1~ : 實際 項目을 包含한다.

例를 들어 어떤 文獻이 4個의 項目을 갖고 있는 경우를 생각해 보자. 즉 2名의 著者(J. Smith와 K. Brown), 標名(SPECTROSCOPIC ANALYSIS), CA Citation Number(CA 082 02 123456)와 C123456이라는 7字의 Sort-Key를 갖고 있다면 Standard Format로 變換된 테이프는 다음과 같다.

				7.
			7	4
C 1 2 3 4 5 6				
21	1	7	1	8
21	2	7	2	8
31	1	7	3	22
11	1	7	6	16
J . S M I T H				
K . B R O W N				
S P E C T R O S C O				
P I C A N A L Y S				
I S				
C A 0 8 2 0 2				
1 2 3 4 5 6				

그림 4.

3.2 PROF

이 프로그램은 SDI시스템에서 프로파일의 마스터 파일을 維持하는데 必要한 것으로 프로파일의 添加, 削除 및 修正에 使用된다. 여기에 入力되는 파일은 이 프로그램 遂行 以前까지 使用되어 오던 마스터 파일(Old Master File)과 更新하는 경우에는 更新하는 프로파일들로 構成된다. 目的에 따라 다음과 같은 Option중에서 選擇하여 Control Card에 Punch하여 주되 =은 1欄에 오도록 한다.

① = UPDATE ; 가장 많이 使用되는 것으로 프로파일의 更新에 使用된다. 修正되는 카아드들은 =UPDATE 카아드 바로 다음부터 시작하고 제일 끝은 =EOD 카아드를 쓴다.

② = SELECT ; 어떤 特定한 프로파일을 選擇하여 리스트하기 위하여 使用된다. 選擇될 프로파일들은 利用者 코오드로 特定시켜 주며 이들의 마지막 카아드로는 =EOD를 쓴다.

③ = STAT ; 이 카아드를 使用하면 프로파일을 收錄한 프로파일 파일의 內容에 대한 統計가 만들어 진다.

④ = LIST ALL ; 프로파일 파일의 完全한 리스트를 만드는데 使用한다.

이 프로그램의 結果로는 更新된 새로운 마스터 파일(New Master File)과 修正된 프로파일의 리스트 그리고 =LIST ALL Option을 준 경우에 限하여는 更新된 마스터 파일의 리스트가 나온다.

3.3 SPLIT

프로파일 파일을 읽어 各 프로파일에서 實際 檢索에 必要한 情報를 골라내는 프로그램으로 檢索프로그램인 GENER 8 以前에 행해져야 한다.

이 프로그램의 出力으로 두개의 새로운 파일이 만들어 지는데 첫째는 檢索에 必要한 情報 즉 *P카아드와 *L카아드 등의 內容을 包含하고 있는 것으로 그 構造는 데이터 베이스 파일과 같다. 둘째는 檢索結果를 리포트할 때 必要한 利用者코오드, *D카아드와 *N카아드의 內容 및 프로파일 形態 등의 情報를 包含한 파일로 이것 역시 構造는 데이터 베이스 파일과 같다.

3.4 GENER 8

實際로 檢索을 行하는 프로그램으로 檢索시스템의 核心이 되는 部分이다. 이 프로그램은 SPLIT에서 만들어진 檢索情報가 담긴 파일과 데

이터 베이스 파일을 읽고 서로 比較하여 매치(Match) 되는 것을 히트파일에 收錄한다.

GENER 8는 크게 Preprocessor와 Main Processor의 두 段階로 나누어 檢索을 행하며 各 段階에서 하는 일은 다음과 같다.

3.4.1 Preprocessor

데이터 베이스의 레코드에 包含된 項目들과 어떤 프로파일의 하나의 파라미터에 나타난 項目들을 比較하였을 때 매치가 없다면 이 레코드는 그 프로파일에 대하여 히트가 될 수 없다는 原理를 利用한 것으로, 檢索에 앞서 될 수 있으면 많은 프로파일을 檢索過程에서 除外시키는 것을 目的으로 한다. 檢索에 가장 많이 使用되는 項目을 Preprocessable Tag라 하며 現在 CAC에서는 標名과 重要語의 두개를 Preprocessable Tag로 使用하고 있다.

檢索을 始作함에 있어 GENER 8는 프로파일을 하나씩 읽고 그 프로파일內에 Preprocessable Tag를 包含한 파라미터가 있는 가를 調査하여, 있는 경우에는 그 파라미터 안에 있는 檢索用語를 읽어 Preprocessor Array에 넣어 준다. 이때 檢索用語의 길이가 8字 以上인 경우는 8字만 使用되고 나머지는 잘린다. Preprocessor Array의 構造는 다음과 같다.

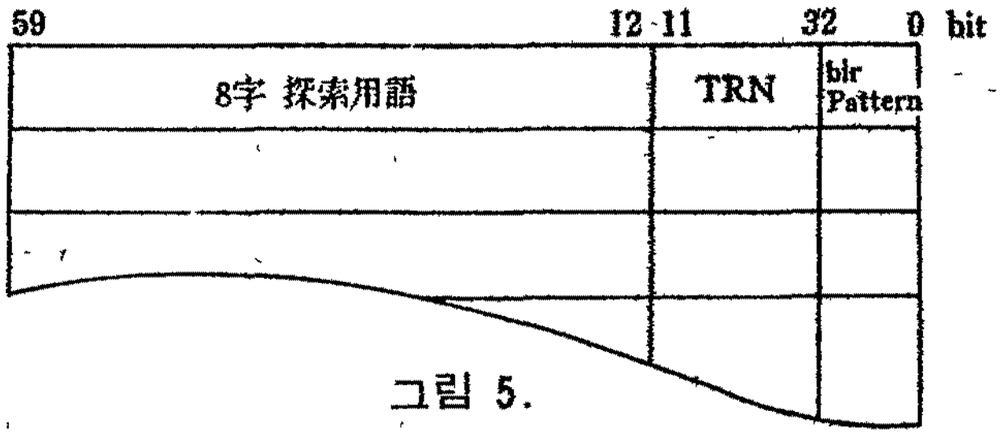

그림 5.

여기에서 TRN이란 SPLIT에 의해 프로파일에 주어진 一連番號로 프로파일ID로 使用되고 있으며 Bit Pattern이란 이 檢索 用語를 使用한 探索種類를 말한다.

Bit Pattern과 이에 따른 探索種類는 다음과 같다.

100 K (Keyword) 探索
010 H (Title) 探索
110 T (Title and Keyword) 探索

모든 프로파일에서 하나씩의 可能한 파라미터가 Preprocessor Array에 들어간 후에 이 Array는 알파벳順으로 排列된다.

다음에는 데이터 베이스에서 하나의 레코드를 읽어 檢索에 必要한 操作을 한다. CAC의경우에는 Substring Array와 Pointer Array라 불리우는 두개의 Array가 만들어 진다. Substring Array란檢索에 使用될 用語의 첫번, 둘째번, ……의 各 10字씩을 택하여 넣은 Array로, 이 Array는 다시 알파벳順으로 排列된다. 만약 檢索用語가 SEMICONDUCTOR인 경우의 Substring Array는 다음과 같다.

SEMICONDUC	CONDUCTOR 0
EMICONDUCT	CTOR 000000
MICONDUCTO	DUCTOR 0000
ICONDUCTOR	EMICONDUC 0
CONDUCTOR0	ICONDUCTOR
ONDUCTOR00	⋮
⋮	
R000000000	UCTOR 00000
	(배열된 array)

이와 같이 檢索用語가 N字로 構成되어 있으면 Substring Array에 N個의 項이 必要하다. 두번째의 Pointer Array란 Substring Array와 같이 N個의 項으로 되어 있으며 이것은 Array 안에서 項의 前後關係를 表示하는데 必要하다. 이 Arary의 各 Word는 크게 두 部分으로 나뉘어져 있다. Bit 59~30의 30 Bits는 排列된 Substring Array의 원래 String안에서의 位置를 말해 주며 나중 30 Bit는 排列되기 以前 Substring의 排列된 Substring Array內에서의 位置를 表示한다. 例를 들어 String이 CEAFDB인경우를 보면 다음과 같은 Substring Array와 Pointer Array를 얻게 된다.

CEAFDB :

배열된 Substring Array	Pointer Array	
	bit 30—59	bit 0—29
AFDB000000	3	3
B000000000	6	5
CEAFDB0000	1	1
DB00000000	5	6
EAFDB00000	2	4
FDB0000000	4	2

또 하나 必要한 Array가 Flag Array로 이 Array에서의 한 Flag는 하나의 프로파일을 代表하고 있다. 이 Flag가 1로 表示된 것은 그것에 該當되는 프로파일의 檢索을 위하여 Main Processor가 必要하다는 것을 意味한다. 이러한 모든 Array가 만들어진 다음에는 Preprocessor Array와 排列된 Substring Array의 첫 8字가 比較된다. Substring Array가 順序대로 排列되어 있기 때문에 매우 빠른 速度로 比較가 可能하며 그 結果 매치가 있는 경우에는 Preprocessor Array의 Bit 0~2에서 探索 種類를 읽고 Substring Array에서 使用된 檢索項目과 比較하여, 같으면 다시 Preprocessor Array의 bit 3~11에서 TRN을 읽어 이에 對應하는 Flag Array의 Flag를 1로 表示해 준다.

하나의 레코드에 속한 모든 Substring Array와 Preprocessor Array와의 比較가 끝나면 Flag Array를 調査한다. 여기에서 0으로 表示된 프로파일은 이 以上의 檢索에서 無視하여도 좋다는 意味이므로 1로 表示된 프로파일에 限하여 Main Process를 한다.

3.4.2 Main Process

여기에서는 우선 데이터 베이스 레코드의 Bit Pattern과 프로파일의 Bit Pattern을 比較하여 (*L카아드를 쓴 경우) 같지 않은 경우에는 다음 프로파일로 넘어가고 같은 경우 즉 두개의 레코오드가 같은 分類에 속하면 純粹論理프로파일인가 加重프로파일인가를 調査하여 各各에 對應하는 프로그램을 行한다. 檢索過程에서는 프로파일에 나타난 대로 檢索用語를 取하여, Substring Array와 Pointer Array를 使用함으로써 用語의 前後關係를 對照하고, 加重프로파일인 경우에는 加重値도 함께 比較하여 現在 레코드가 프로파일의 論理를 滿足시키는 경우에는 그 레코드에, TRN과 그 TRN에 대한 히트의 一連番號로 構成된 Sort-Key, 그리고 檢索되도록 하게 한 檢索用語의 情報를 包含하여 히트파일에 收錄한다. 이와 같이 만들어진 히트파일은 Control Data의 패키지인 SORT/MERGE에 의해 Sort-Key順으로 排列된다.

3.5 SDIACC

GENER 8를 遂行한 다음 檢索統計를 내기 위한 프로그램으로 두가지의 情報를 提供한다. 하나는 檢索이 進行되는 동안의 모든 情報를 包含하고 다른 하나는 各 프로파일當 所要된 檢索時間과 Preprocessor에 의해 除外된 數에 대한 情報를 준다.

3.6 PRNSDI

排列된 히트파일과 SPLIT에서 따로 만들어진 利用者의 姓名, 住所 등이 담긴 파일을 읽고 檢索結果를 定해진 形式대로 프린트해 주는 프로그램이다.

3.7 SELECT

프로파일 作成의 準備段階에서, 必要한 重要語를 찾아내기 위하여 使用되는 프로그램으로 適合文獻을 手作業에 의해 選擇하여 文獻番號로 表示해 주면 테스트 테이프에서 찾아내어 프린트해 준다.

3.8 GNRSTRT

GENER 8가 遂行途中 非正常的으로 끝나버린 경우 끝난 部分부터 다시 시작하도록 하는 프로그램이다.

3.9 READSDI

이 프로그램은 利用者의 姓名, 住所, 서어비스 始作日時 등의 情報를 包含한 파일을 維持하는데 必要한 것으로 利用者 管理에 使用된다.

4. 結 言

컴퓨터情報檢索의 結果는 크게 두가지로 나누어 評價할 수 있다. 첫째로는 檢索文獻에 關係되는 事項이고, 둘째로는 檢索所要時間에 관한 事項이다. 檢索된 文獻에 關係되는 것으로는 흔히 情報科學에서 말하듯이, 再現率을 높이면서 適合率을 높이는데 그 目的이 있고 檢索所要時間은 될 수 있는한 짧게 하는 것이 經濟性을 높인다 하겠다. 그러나 檢索所要時間을 줄이기 위하여 프로파일 作成에 너무 制限을 加하다가는

〈p. 119에 계속〉

必要情報를 如何히 選擇할 것인가가 重要한 問題로 登場된다. 必要情報의 選擇은 情報管理시스팀構成下에서 보다 合理性과 客觀性이 賦與될 것이지만 本稿에서는 주로 情報의 價値問題를 中心으로 考察하였다. 意思決定에 있어 必要한 情報의 價値를 評價함에 있어 情報가 가지는 特性 즉 普通財貨와 달리 어느 누구에게도 同時에 所有 使用될 수 있으며, 使用에 의하여 破壞나 磨耗가 되는 일이 없다는 것이다. 따라서 情報에 대한 價値評價가 곤란하게 된다. 다행히 最近에는 情報도 物的 財貨로 취급, 分析코자 하는 傾向이 濃厚하여 지고 있다. 그러나 지금까지는 主觀的 評價方法에 의하여 情報를 選擇하므로서 信賴性, 正確性 및 適切性이 缺與되 있었다고 볼 수 있다. 經營科學의 發達로 客觀的인 價値評價方法이 繼續 進展되고는 있지만 非可視的 特性을 가지고 있는 不完全情報를 客觀的이고 可視的이며 計量化된 完全情報로 轉換하는 方法은 아직까지 未開拓狀態에 있다고 볼 수 있다.

引 用 文 獻

1. R. Schlaifer, "Decision and Theory", McGraw-Hill Book Co, 1964, p. 68.
2. Herbert A. Simon, "Administrative Behavior", The Free Press, N. Y, 1957, p. 324.
3. C. I. Barnard, The Function of the Executive", Havard University Press, 1958, p. 202
4. H. A. Simon, "The New Science of Management Decision", p. 1.
5. H. A. Simon, op. cit., p. 2.
6. 刀根伸太郎, "經營情報活用に對する一考察" 大阪產業能率研究所刊, p. 47.
7. 平井治, "中小企業における經營情報整理への一考察", 大阪產業能率研究所, p. 90.
8. 金鍾義, "經營情報管理体制에 대한 Approach", 國民大學論文集 1971. pp. 246~248.
9. 小島敏彦, 中小企業におけるビジユアル化をめざす情報管理の進め方, 大阪產業能率研究所, p. 75.
10. 多田和夫編, 企業と情報, 1965, p. 21.

〈p. 114의 계속〉

適合文獻을 漏落시킬 危險이 있다. 그러므로 컴퓨터情報檢索의 效率은 같은 檢索效率을 維持하는 內에서 所要時間을 줄일 때 진정 높아 질 것이다. 그리고 이것은 프로그램의 理解로 이루어 질 수 있다. 쉬운 例로서, 어떤 프로파일이 파라미터가 두개이고 둘다 Preprocessable Tag에 속하는 경우에는 첫번 파라미터에 頻度가 낮은 檢索用語를 使用하고, 만약 파라미터중의 하나만이 Preprocessable Tag인 경우에는 그 파라미터에 頻度가 낮은 檢索用語를 使用함으로써, 될 수 있으면 많은 데이터 베이스 레코드를 無視하도록 하여 所要時間의 比重이 큰 Main Processor의 負擔을 적게 할 수 있다. 그러므로 檢索效率을 높이기 위하여는, 主題에 대한 背景 및 檢索시스팀의 充分한 理解에 根據한 프로파일 作成技法이 考慮되어야 할 것으로 본다.