KORSTIC의 情報處理機械化計劃

崔 成 溶
韓國科學技術情報센터

〈問題點〉

1. 漢字情報處理시스템導入에 있어서의 問題點

漢字情報處理시스템導入에 있어서의 問題點과 그 解決方案은 다음과 같다.

(1) 漢字鍵盤入力裝置

1) 使用可能 字種數의 制限

지금까지 "速報"에 使用하던 文字를 全部 採擇한다면, 使用字種의 過多로 打鍵能率이 低下되고, 誤字가 增加할 可能性이 많으므로, 字種數의 制限이 必要하다.

그리고 實際로 漢字鍵盤에 收容可能한 字種數도 約 3,000字種 뿐이다. 따라서 漢字, 한글, 漢文字, 英文字, 數字, 및 各種 記號를 合하여 3,000字種 以內로 字種數를 制限하지 않으면 안된다.

이와같은 文字種의 制限을 위해서는, "速報"에 있어서의 使用頻度가 높은 順으로 文字를 採擇하되, 非常用文字라 하드라도 使用頻度가 높거나, 人名, 地名 등에 자주 使用되는 것은 採擇하고, 常用文字라 하더라도 使用頻度가 극히 적은 것은 採擇하지 않는 것을 原則으로 하여, 漢字는 常用漢字와 科學技術用語로서 자주쓰이는 漢字를 合하여 約 2,000字 內外로 制限할 必要가 있다.

그리고 한글도 作業能率을 위하여, 使用頻度가 극히 높은 主要文字는 모아쓰기式으로 入力할 必要가 있다. 한글入力은 풀어쓰기도 可能하나 모아쓰기式은 1回의 打鍵으로 한 글자가 入力 可能하므로 使用頻度가 높은 글자는 보다 效率的이기 때문이다.

KORSTIC에서는 이를 위하여 1974年 8月末까지完了豫定으로 文獻速報와 特許速報의 漢字 및 한글 各 文字의 使用頻度를 調査하고 있다. 調査方法은 文獻速報 各編과 特許速報 1個月分으로부터 等間隔으로 標本을 抽出하여, 여기에 包含된 한글과 漢字의 各文字의 使用頻度를 調査하는 方法을 採擇하고 있다.

2) 人間工學的인 鍵盤 文字配列

作業能率을 위하여 漢字鍵盤의 文字配列에는 文字의 使用頻度에 따르는 人間工學的인 配慮가 要求된다. 따라서 KORSTIC에서는 使用頻度가 높은 文字는 앞쪽(下段) 中央과 오른쪽에 配列하고, 使用頻度가 높지 않은것은 上段 中央과 오른쪽에 使用頻度가 극히 적은것은 왼쪽 上下段에 配列할 計劃이다.

3) 한글—漢字變換시스템 開發

漢字鍵盤入力裝置의 使用은 現段階에 있어서는 不可避하나, 장차에는 한글入力用 打字機에 의하여 한글로 入力한後, 電算機處理에 의하여 漢字로 變換하여 出力하는 한글—漢字變換시스템의 開發이 要求된다.

한편 現段階로는 歐文의 入力도 漢字鍵盤入力裝置에 의하는 것이 不可避하나 장차 抄錄誌化할 경우에는 原文標題와 數字로 表現되는 Data의 入力에는 歐文入力用 打字機를 使用할 必要가 있다.

(2) 校正問題

1) 눈의 疲勞

CRT Display 裝置에 의한 校正作業이, 校正用Print를 作成하여 校正하는것 보다, 校正所要時間을 크게 短縮시키고, 校正잘못 發生率을 극도로 減少시키는것은 事實이나, 눈의 疲勞가 甚하여, 實際로 速報의 繼續的인 校正作業이 可能할른지, 그리고 校正作業에 長期從事할때 視力障碍가 없을는지는 아직 알수 없다.

그러나 이 問題는 校正作業의 交代制에 의하여 쉽게 解決될수 있을 것으로 생각된다.

2) 完全校正의 可能性

한편, 校正用 Print 없이 CRT Display 만으로 速報의 完全한 校正이 可能할른지도 아직 알수 없다. 그러나 이 問題도 必要할 경우, 最終校正段階에서 校正用 Print를 作成하여 校正하면, 經濟的인 負擔은 多少 생겨도, 容易하게 解決될 수 있을 것으로 본다.

(3) 自動編輯 處理

1) 外部 電算機 依存

電算機를 갖추지 못하였고, 가까운 時日內에 導入될 可能性도 없으므로, 當分間 外部 電算機에 全的으로 依存하여야 한다. 그러나 장차 速報를 抄錄誌化 할 경우에는, 이것 만으로도 相當한 作業量이 될 것이므로, 單獨的인 電算機의 導入이 不可避하다.

2) 速報分類體系의 完備

電算機에 의한 速報 自動編輯處理를 위해서는 速報分類體系의 完備가 必要하다. 特許速報는 各國別 特許分類 및 國際特許分類가 完備되어 있어, 이것을 利用할 수 있으므로 問題가 없으나, 文獻速報는 分類體系가 細分類까지 完備되지 못하였으므로 早速히 이것을 補完할 必要가 있다.

(4) 自動植字 裝置

1) 植字의 品位

自動植字裝置의 植字의 品位가 現在의 速報印刷原版(紙版) 作成用 즉 輕印刷用으로는 充分하나, 高級 印刷用으로는 약간 不足하다. 32×32個의 點으로 表現되므로 부득이 한글이나, 장차 글자體의 改良등에 의하여 약간의 品位向上은 可能할 것으로 생각된다.

2) 필름印字機能

大量印刷와 印刷物의 質의 向上을 위해서는, 金屬印刷原版에 의한 印刷가 要求되는데, 이를 위한 필름印刷機能이 갖추어지지 않았다. 장차 速報의 發刊部數의 增加에 對備하여 필름 印字機能을 갖출 必要가 있다.

(5) 情報蓄積

情報의 蓄積은 情報檢索(IR)을 前提로 한다. 따라서 檢索이 容易하도록 情報가 蓄積되지 않으면 無意味하다. 漢字情報處理시스템의 導入은 速報, 編輯 組版自動化와 함께, 그 副産物로서 情報檢索이 可能한 情報蓄積File(磁氣테이프)를 얻는데 그 目的이 있다. 따라서 다음과 같은 考慮가 必要하다.

1) 最適의 File 및 Record 形式의 採擇

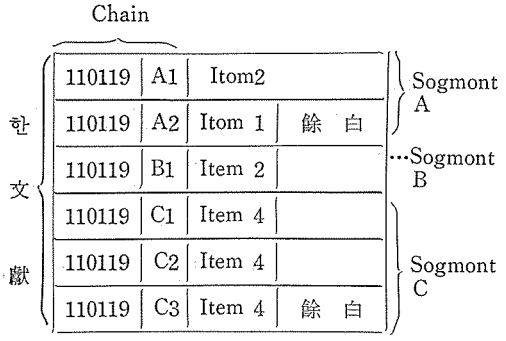
File 및 Record 形式의 採擇에는 Memory의 浪費와 檢索所要時間이 가장 적은 形式이 採擇되도록 考慮하여야 한다.

그런데 速報에 收錄되는 內容은 著者名, 標題 雜誌名, 卷, 號, 페이지 등과같이 Record의 各項目(Item)의 Data의 길이가 一定치 않다. 장차

抄錄誌化할 경우에는 더욱 그러하다. 따라서 固定長 Record 形式은 融通性이 없고 Memory의 浪費가 많아 不適當하다. 또 可變長 Record 形式은 Memory의 浪費는 없으나 校正에 不便하여 適當치 않다.

따라서 兩者를 混合한 半固定長 形式(Segment 方式)이 適合하다고 생각된다. 이 方式은 可變長의 情報를 可變個數의 固定長으로 分割하고, 이것을 便宜上 한 Record로 간주하는 方式으로서 한 文獻 (한 Record)을 Item 別로 몇개의 Segment로 나누고, 이 Segment를 Item의 길이에 따라 다시 몇개의 固定長의 Record로 나누어서 記錄하는 方式이다.

(文獻番號, Sogmont ID)



2) 用語 Keyword의 標準化

情報檢索에 있어서 自然語 그대로의 用語(Keyword)의 使用은 檢索效率을 극도로 低下시키므로, Keyword의 標準化는 그 前提條件이 된다. 따라서 檢索效率을 높이려면, 情報蓄積段階로부터 標準 Keyword로 統一하여 蓄積하지 않으면 안된다. 그러므로 Keyword의 標準化는 情報蓄積을 前提로 하는 漢字情報處理시스템의 導入에 있어 時急한 問題이다. 따라서 다음과 같이 Keyword의 標準化 作業을 推進할 必要가 있다.

① Keyword의 抽出 및 選定

文獻速報, 特許速報의 1年分으로부터, 年間 3回以上 出然한 自然語 그대로의 用語를 抽出하며, 이것을 分類別로 整理한다.

② Keyword間의 關係의 整理

語(USE, UF)를 整理하고, 上位槪念과 下位槪念分野別로 同意念(NT, BT 및 關聯語(BT)를 整理하여 標準用語(Descriptor)와 非標準用語로 區分한다.

③ Thesaurus(檢索用語辭典)의 作成

關係의 矛盾을 체크하고, 上下關係를 擴張한 다음, 主題 Category, 註記, 關係語 등 Entry를 記載하고, Keyword의 ABC順 또는 가나다順으로 配列하여 Thesaurus를 作成한다.

④ Thesaurus의 整備

使用頻度가 적은 用語를 削除하거나 새로운 用語를 追加하고, 이에 따르는 關係의 變化와 主題, Category, 註記 등 補足的인 情報를 整備한다.

이와같은 Thesaurus의 作成은 手作業만으로는 不可能하므로 電算機를 利用하여 Thesaurus를 自動作成하기 위한 "用語管理시스템"의 開發이 時急하다.

3) Keyword의 抽出 및 記載

現在의 情報에 記載되는 것은 記事番號·著者名, 分類, 標題, 雜誌名, 卷, 號, 페이지, 年度 등 書誌的事項 뿐이다.

따라서 漢字情報處理시스템에 의하여 이와같은 情報가 蓄積되었다하여도 情報檢索에 使用하기에는 不足하다. 그것은 分類나 標題가 文獻의 內容을 充分히 代表할수 있다고는 생각할 수 없기 때문이다. 따라서 從來의 速報의 記載事項外에, 그 文獻의 內容을 代表할 수 있는 Keyword를, 原文 또는 抄錄으로 부터 적어도 5個以上 抽出하여, 이것을 速報原稿에 함께 記載하고 漢字鍵盤入力裝置에 의하여 "한글"로 入力시켜 둘 必要가 있다.

2. CA Condensates SDI서어비스 實施에 있어서의 問題點

CA Condensates SDI서어비스 實施에 있어서의 問題點과 그 解決策은 다음과 같다.

(1) 需要者의 確保와 要求의 充足

1) 需要者의 確保

電算機에 의한 이와같은 種類의 서어비스의

最初의 試圖이며, 國內 科學技術의 水準 및 經濟的인 與件 등을 감안할 때, 반드시 需要者의 確保를 樂觀만은 할 수 없다.

따라서 需要者의 確保와 PR을 위하여 最初의 6個月間은 無料로 하여 大量으로 配布하고, 다음 6個月間은 繼續利用希望者에게 有料로 配布하되 最少限의 料金으로 하고, 1年後 부터 適正料金을 算出하여 有料로 提供하는 3段階의 점진적인 實施가 必要하다고 생각된다.

2) 需要者의 要求와 充足

CA Condensates SDI서어비스는 情報性에 있어서는 問題가 없으나 檢索回答書에 抄錄이 包含되어 있지 않다. 따라서 利用者의 要求를 完全히 充足시켜 줄수는 없다.

따라서 必要할 경우에는 抄錄誌를 航空便으로 迅速히 入手하여 "抄錄카아드" 서어비스등 補足的인 서어비스가 必要하다.

(2) 檢索技術의 開發

1) 프로그램의 開發

"CA Condensates"磁氣 Tape에는 다른 旣成情報파일과 같이 完全한 檢索프로그램이 作成되어 提供되지는 않는다.

다만 "Evaluation package"를 提供한다. 따라서 이것을 보고 利用者 自身이 프로그램을 開發하여야 한다.

"Specification Manual"과 "Evaluation Package"를 이미 發注하였으므로 到着되면 KIST와 共同으로 開發할 計劃이다.

CA Condensates 磁氣 Tape의 Record Format는 CAS의 SDF (Stanadrd Distribution Format)로서, Tag와 Value를 그룹化하여 配列하는 Directory 方式(tag value grouping方式)이며, 9-track mode, 800BPI로 記錄되어, "Standard IBM OS/360lables" 또는 "no lables"로서 配布된다.

2) 質問式의 作成

情報를 檢索하려면 利用者의 要求를 分析하여 自然語의 Keyword를 抽出하여 質問을 表現하고 이것을 形式化된 質問式으로 바꾸고, 이것을 入力시켜 檢索處理를 하여야 하는데, 質問式에는 Keyword 間의 論理關係(AND, Or NOT)와 檢索條件(=, ≥, ≥, ≒) Match 方式(完全一致, 前方一致, 中間一致, 後方一致), 무게(Keyword의 重要度를 定量的인 大小로 表現)의 表示등 어려운 問題들이 많다.

3) User profiles의 作成 및 更新

SID서어비스의 檢索效率은 User profiles의 作成과 그 更新의 잘못에 左右된다.

利用者들의 要求를 分析하여 適切한 質問式을 만들어 User profiles을 만들고, 檢索하여 提供한 結果에 대하여 利用者로 부터 Reponse Sheet를 받아서 User profiles를 更新함으로써, 보다 利用者가 願하는 情報에 接近하도록 하여야 하기 때문이다.

(3) 檢索效率

CA Condensater는 用語가 完全히 標準化되어 있지 않다. CA 抄錄作成段階에서 多少 用語를 規制하고 있으나, 다른 시스템들과 같이 用語의 標準化가 되어 있지 않고, 完全히 規制되지 않은 自然語의 用語를 使用한다. 따라서 檢索效率이 그다지 높지 못하다. 따라서 이點을 充分히 考慮할 必要가 있다.