計量群集索引의 模型提示

李　泰　榮
(公州師大 講師)

〈차　례〉

Ⅰ. 序　論

오늘날 現存하는 科學雜誌의 數는 約 6～7 萬種이고 科學者의 數는 約 百萬名 程度로 推算된다. 이들 科學者가 科學雜誌에 發表하는 論文記事 수는 年間 約 3～4 百萬 件이며 單行本은 약 10 萬種에 이르고 있다.[1] 이 이외에 報告書, 書類, 特許資料, 會議錄 등 各種 文獻을 包含하면 그 量은 百萬 科學者의 數 만큼이나 놀랄 만한 것이다. 흔히 말하듯이 情報의 暴發時代라는 말이 그대로 實感이 난다.

드 솔라 프라이스는 科學雜誌의 數量과 科學者의 數는 50 年마다 10 倍씩 增加한다고 하였다.[2] 그의 理論에 依하면 앞으로 세월이 흐를수록 情報量은 想像

1) 崔成眞, 「情報學原論」, 서울 ; 亞細亞文化史, 1976, pp.5-8,98-99.
2) Ibid.

을 초월하여 暴增할 것이며 情報滯症 現象은 점점 심각해 질 것이다. 이러한 情報滯症은 과학자들 自身들에게도 危機가 될 뿐만 아니라 直接 情報流通을 擔當하는 情報實務者와 圖書館 司書에게는 第一의 問題로 등장한다. 情報檢索의 歷史的 變遷을 살펴 보면 檢索方法과 技術 그리고 道具의 變化가 부단하게 이루어져 왔음을 알 수 있다. 各種 分類表, 索引方式, 目錄, 書誌類는 方法과 技術의 結晶이었고, 手動式, 機械式, 컴퓨터檢索시스템은 道具의 발전에 따른 結果이었다. 이 變化의 主原因은 요컨대 學問, 技術의 발전과 情報量의 增加에 직접적인 聯關을 갖는다.

1950년대 後半부터 實用化되기 始作한 컴퓨터 情報檢索시스템도 그런 變革의 결과이었다. 사실 2차대전을 기화로 發達의 速度를 加速化한 科學技術은 尨大한 情報를 生産하였고 이 情報들은 發達速度에 맞추어 迅速, 正確하게 그리고 廣範圍하게 配布, 利用되기를 要求하였다. 컴퓨터는 바로 그러한 要求를 滿足시켰다고 볼 수 있다. 컴퓨터를 媒介로 한 檢索시스템의 長點은 무엇보다 檢索媒體와 축적文獻이 아무리 방대하여도 迅速히 處理하는 것, 그리고 計量化된 索引補助素를 演算式으로 解析할 수 있는 것이다. 그러므로 情報量이 아무리 많아도 檢索을 媒介하는 索引시스템이 精密하면 利用者가 찾고자 하는 情報資料를 容易하게 探索 提供할 수 있다.

索引시스템은 上記한 分類表, 索引法, 目錄, 書誌類들과 모두 關聯된다고 할 수 있는데 좁게는 索引言語를 指稱한다. 本 索引言語는 圖書館이 出現한 古代 메소포타미아와 애급의 時代서부터 硏究의 對象이 되었으며 그때부터 進化 改良되어 왔다. 그 變遷內容을 살펴 보면 分類方式, 主題名方式, 組合(키워드)方式의 順으로 變化의 主流를 이루었다고 할 수 있다. 제일 최근의 方式인 組合索引은 20C에 들어와 本格的인 發展을 하였는데 여러가지 方法이 出現한 結果, 現在는 後組合方式의 키워드索引이 그 主流를 이룬다.

그러면 앞으로 索引은 어떠한 形態를 이룰 것인가?

그 形像은 정확히 알 수 없으나 분명한 點은 드 솔라 프라이스의 公式대로 情報量이 急增한다면 索引은 머지 않은 將來에 變化를 초래할 것이다. 다시 말하면 情報量의 增加와 學術의 發展에 따라 索引의 方式과 技術上의 變革이 期待된다는 것이다.

現在의 檢索節次는 主題索引語와 기타 探索媒體로 探索한 文獻을 최종적으로

抄錄을 보고 選擇하는 檢索사슬을 갖고 있다. 그런데 現 節次대로 探索을 하였을 때에, 例를 들어 한 質問에 對한 回答문헌이 100件 以上 된다면 그 文獻들의 確認은 적지 않은 時間을 要할 것이다. 앞으로 情報量이 끊임없이 增加하면 이와 같은 事態는 언젠가 닥쳐오게 마련이다. 그때에는 現組合索引시스템은 利用者에게 不合理하고 非經濟的인 시스템으로 烙印을 찍히고 만다.

그러므로 現 檢索시스템을 改善하려는 硏究가 부단히 이루어져야 할 것이다. 本稿는 이러한 努力의 일환으로 歷史的인 索引의 變形樣態를 情報量의 增加와 學術의 發展에 結付하여 고찰하고 그 發生要因들을 分析한 후 當面問題를 把握하고자 한다. 그리고 그 結果에 따라 現 키워드索引시스템내에서 方法論的인 變形을 提示하고 그 模型을 運營할 수 있는 過程과 設計를 提示하고자 한다.

Ⅱ. 情報量과 索引方式

역사적으로 考察하여 보면 情報量에 따라 索引方式이 변해 왔음을 알 수 있다. 즉, 學問·技術의 發展에 따라 새로운 主題意味語들의 多數出現, 그리고 主題의 細分化, 特定化로 말미암아 索引方式은 固定的이고 包括的인 것에서 유동적이고 특정적인 시스템으로 改良되어 왔다. 具體的으로 이것들을 分析 비교하면 확실한 그 變遷內容과 當爲性을 발견하게 되며 미래의 推移를 具想해 볼 수 있다.

1. 分類와 主題名索引의 比較

分類索引은 記號를 사용하고 槪念을 體系的으로 벌려놓은 分類表를 필수적으로 동반한다. 분류표의 각 項目은 하나의 주제범주를 대표하며 그것을 表象하기 위하여 記號를 組立한다. 따라서 기호의 論理에 따라 주제개념들 간의 遠近과 上下가 表意된다. 이에 반하여 후에 出現한 主題名索引은 記號를 使用하는 대신에 直接 語彙를 索引語로 사용하였다.

이 두 索引의 長短點을 分析하면 分類索引에서 主題名方式의 出現을 理解하게 된다. 長點으로서의 分類方式은 隣接主題와 關聯主題를 體系的으로 한 系統에 모아주고 言語와 用語問題를 克服하는 點이며 主題名方式은 語彙를 직접 檢索媒體로 사용하기 때문에 利用이 便利하다. 短點으로서 分類方法은 계통별 체계화에

따른 논리적 思考를 거치지 않고서는 特定主題에 直接 接近할 수 없다는 것이며 主題名方法은 上下관계에 있는 主題 및 체계상 서로 近接한 主題가 字順에 따라 分散되며 같은 主題를 가진 槪念이 國語를 달리 할 때와 異音同意語에 따라 分散되는 點이다.[3)]

예를 들면 「아시아 통계」는 분류기호가 " 315 "로 索引이 되므로 그 체계가 일목요연하게 '사회과학', '통계학', '아시아 통계'의 順序에 잡혀 있다. 따라서 관련주제와 인접주제, 상하주제가 한 곳에 모여 그 도서관의 장서 상황을 한눈에 알 수 있게 하며 필요한 자료를 쉽게 찾아 가게 한다. 그런데 「아시아 통계」는 그 분류기호를 생각할 적에 어느 곳에 있는지 한참 걸려야 찾아 갈 수 있다. 그 불편한 점은 「아시아 통계」를 곧바로 「아시아 통계」로 색인하면 없어진다.

애초에 資料가 별로 없고 主題分枝가 적었을 적에는 적절한 代表的 表象槪念을 抽出하여 그것에 맞는 적당량을 配定하여 分類를 하였고 필요할 時에 그 表象槪念으로 접근하여 어렵지 않게 찾고자 하는 資料를 획득하였을 것이다. 이때에 表象形式이 어휘이던 기호이던 간에 그것이 重要하지는 않았을 것이다. 그러던 것이 세월의 흐름에 따라 학문기술이 발전하고 文明이 發達함에 따라 主題개념이 多樣해지고 情報量이 늘어남으로 해서 分類展開가 점점 複雜해지고 主題包括 範圍가 넓어졌고, 따라서 어휘보다는 統一性있고 統制하기 좋은 記號를 使用하는 쪽으로 기울어 졌을 것이다.

분류는 어느 分類索引이든간에 그 분류체계를 形象化한 分類表를 갖는다. 분류체계는 주제개념의 체계화와 기호의 論理的 展開가 合成한 것이며, 分類表는 記號配定과 자리마련이 거의 固定性을 갖는다. 그러므로 통일성이 있고 통제성이 강하나 반면에 즉각적인 체계변환이 어렵다. 시대에 맞추어 보완을 한다고는 하지만 기호의 固定槪念을 바꿀 수 없으며 따라서 새로운 자리 마련이 쉽사리 이루어지지 않는다. 그리고 보완기간의 간격도 문제로 등장한다.

이같은 문제를 해결하기 위해 신종주제가 날로 출현하며 주제세분화가 점증되던 시기에서 어휘가 갖는 장점인 즉응성과 실용성을 살린 主題名方式을 생각하게 되었고 그 단점인 비체계성과 동의어분산을 보완하여 사용하게 되었다.

---

3) 이 재철, "신문기사 색인법의 이론과 실제", 「人文科學」, Vol.22, 1969, pp.83-99.

2. 主題名과 키워드索引의 比較

組合索引은 20세기 前葉부터 각광을 받기 시작한 方式으로서 主題名의 單一標目形式을 타파한 보다 進一步한 시스템이었다.

組合索引은 主題名과는 달리 한 표목에 여러개의 독립된 색인어를 사용한다. 예를 들면 「항공기 엔진의 소음」에 관한 문헌을 主題名索引은 「항공기-엔진」과 같이 單一標目으로 索引하고 組合索引은 「항공기, 엔진, 소음」과 같이 3個의 단어를 갖는 多重標目形式으로 表現한다.[4] 이 形式에서 보이는 바와 같이 「항공기-엔진」보다는 「항공기, 엔진, 소음」이 보다 文獻을 細部的으로 特定性있게 索引할 수 있는 形式이며 自由로운 語彙選擇을 할 수 있다.

우리가 "꿩"이라는 特定主題語를 主題名으로 擇하였다고 하자. 그후 "꿩"에 관한 硏究가 急增하여 "꿩"이란 主題名으로 그 關聯文獻을 全部 索引하였을 時에는 適合文獻의 區別이 잘 안되는 狀態에 온다. 그러면 부득이 「꿩-내장」, 「꿩/제주도」와 같이 聯合標目을 使用해야 한다. 또한 꿩에 관한 情報量이 늘어 날수록 「꿩-내장-기생충」과 같이 三重으로 또는 「꿩-내장-기생충-흡입판」과 같은 四重聯合標目을 使用해야 한다. 이렇게 聯合사슬이 많아지면 主題名索引의 長點을 상실하게 된다. 主題名索引은 語彙를 使用한 簡單性, 即應性이 그 特徵인데 四重, 五重의 사슬을 갖게되면 主題名이라기 보다는 어휘로 쓰여진 분류형태를 나타낸다. 즉, 「꿩 내장 기생충 흡입판」이란 從屬的 複合槪念의 分類의 한 項目과 흡사하다.

한편 主題名 索引은 主題名을 設定함에 있어 異領域간의 主題를 함께 다루어지지 못한다. 즉, 異주제를 다룬 論題는 주제명을 /나 -로 連結하여 索引하기가 곤란해 진다. 예를 들면 곰팡이와 細菌에 關한 記事는 「곰팡이-세균」, 「곰팡이/세균」, 「곰팡이, 세균」으로 처리할 性質의 것이 아니다.

상기의 두 가지 문제를 해결하는 길은 主題名의 사슬을 풀어 놓는 방법뿐이다. 1930年代末에 등장한 Peek-a-boo方式이나 1950年代 初의 Uniterm方式은 이러한 主題名의 사슬을 풀어 놓은 組合索引이었다. 組合索引은 前記하였듯이 문헌을 「꿩, 내장, 기생충, 흡입판」의 形式으로 索引한다. 그래서 各各의 單語는

---

4) 사공 철, 「情報檢索論」, 서울; 亞細亞文化社, 1977, pp.64-65.

獨立된 索引語 子실을 하여 組合能力을 갖는다. 이 形式은 바로 主題名 표목처럼 從屬的 複合概念의 單語形式을 갖지 않고 원래 主題名의 長點인 간단성을 살리며 異種主題도 「곰팡이, 세균」의 組合形式으로 처리되어 方法論上의 問題는 提起되지 않는다.

다시 말해서 組合索引은 主題名이 갖는 上下 從屬關係의 體系性(예, 복-말복)과 複合語를 上位概念으로 類聚하는 一體性(예, 보험/생명)의 굴레를 타파하여 자유스러운 方法으로 索引語를 使用하도록 유도한 시스템이다. 결과적으로 情報量이 급격히 늘어나고 主題의 細分 및 異領域間의 主題結合이 進行되던 時期에 마땅히 出現해야 할 索引方式이었다.

그리고 本 시스템은 主題名索引에서 처럼 異音同意語의 分散을 막고자 同意語를 묶어 주며 아울러 上·下관련어와 인접어도 서로 참조할 수 있게 語彙表(Thesaurus)를 統制手段으로 使用한다.

3. 統制語와 自然語키워드索引의 比較

組合索引이 出現한 후 세월이 흐름에 따라 또 하나의 변화가 抬頭하였다.

1960年代 末부터 本格的인 움직임을 보인 自然語 키워드索引시스템이 바로 그 주인공이다. Klingbiel 같은 이는 DDC[5]에 報告書를 제출하면서 自然語의 時代가 왔음을 强調하였다.[6] 이 自然語시스템은 組合索引에 對하여 索引作成方法과 標目形式을 變化시킨 것이 아니라 단지 索引語彙에 대한 統制를 풀어 自由롭게 索引語를 選擇하도록 유도하는 主張이었다.

이제까지의 語彙索引은 主題名, 組合(키워드)을 막론하고 語彙表 (主題名標目表, Thesaurus)를 마련하여 異音동의어, 관련어, 上下관계어등을 유취시켜 주었다. 그 目的은 어휘가 갖는 단점인 분산성을 최소한으로 막자는 의도에서였다. 그런데 이 統制하는 사슬을 끊어 버리자는 주장이 바로 자연어시스템이었다. 그 주장의 근저에는 自然科學, 技術科學 分野의 급속한 發展이었다. 학술의 발전에 따라 主題概念은 特殊專門化되었고 이영역간의 결합은 新種學問의 誕生으로 이어졌다. 그 결과 많은 主題概念이 출현하였고 점점 主題意味語들은 特定化하였

---

5) Defence Documentation Center.

6) R.A.Wall," Intelligent Indexing And Retrieval : A Man-Machine Partnership", *Information Processing &Management*, Vol.16, pp.73-90.

다. 따라서 索引語들도 相對的으로 特定性있는 主題語로 대체되었다. 그런데 이들 새로 부상한 索引語들이 同意語를 별로 갖고 있지 않은 것이 自然語시스템이 등장하게 된 主要 이유이다.

어느 학문에서 同意색인어들의 분산이 극히 적다면 굳이 컴퓨터 내부에 디소러스를 소장시켜 일일이 確認 統制하는 形式을 거칠 필요가 없다. 다시 말해서 불필요한 프로그램을 두어서 쓸데 없는 過程을 수행케 하는 非經濟性을 제거함은 당연한 理致이다.

오늘날 自然語시스템은 統制語시스템과 어깨를 나란히 하여 多數의 電算化檢索시스템에서 並行 運營되고 있다. 따라서 現在는 변하는 過程에 있다고 할 수 있다.

이상으로 우리는 분류에서 자연어 키워드색인까지 비교 분석을 하였다. 그 결과 색인방식은 고정적이고 통제적인 것에서 유동적이고 自由的인 方法으로 변화하였음을 알수 있었고, 그 변화의 主原因은 학술의 발전과 그에 따른 情報量의 增加임을 고찰하였다. 그러므로 앞으로 情報量이 늘어남에 따라 색인방식에 變化가 오리라고 짐작할 수 있다.

Ⅲ. 計量群集索引의 提起

1. 提起의 當爲性

앞장에서 색인의 변천양식을 情報量의 增加와 聯關지어 분석하였다. 그로써 索引言語는 主題의 特定化와 情報量의 증가로 말미암아 自然스럽고 實際性이 强한 체제로 變하여 가는 事實을 발견하였다. 사실 최근의 方式인 自然語 키워드시스템은 그만큼 편리한 시스템이라고 할 수 있다. 그러나 편리한 만큼 부작용이 따르는데 바로 검색률에 對한 爭點이다. Salton의 實驗으로 보면 통제어시스템에 비해 정도 재현율이 共히 약 20% 정도 낮았다.[7] 이것은 Medlars의 統制語시스템의 檢索效率이 精度率 61%, 再現率 31%인 것과 比較하면 꽤 낮은 것임을 알 수 있다. 그러나 몇몇 硏究의 例를 보면 그렇게 效率이 떨어진다고 말할

---

7) G. Salton, *Dynamic Information And Library Processing*, Englewood Cliffs : Prentice-Hall Inc., 1975, pp.105.

수 없다.[8] 問題는 現 組合索引이 共通的으로 안고 있는 索引作成 狀況이다. 그것은 索引이 歷史的, 傳統的으로 갖고 있는 숙명이기도 하다. 첫째는 學術의 發展에 따라 特定主題語들이 多數 出現을 하여 索引語彙들이 증가되는 點, 둘째, 情報量의 急增으로 인해 各 索引標目에 附與하는 索引語 數가 水平的 또는 水直的으로 많아지는 點[9] 이 있으며, 세째는 索引語彙 選擇이 索引者, 探索者, 著者에 따라 人爲的인 相異性이 深大하여 가는 現象을 피할 수 없는 것이다. 따라서 그 세 가지 要因으로 말미암아 檢索效率이 下落하게 된다. 例를 들면, 첫째번의 索引語彙의 增加는 그 增加比率에 따라 反比例로 再現率이 떨어진다. 100 個의 索引語彙가 運營될 때와 150 個의 索引語彙가 運營될 때는 두 群의 各 索引語의 選擇確率은 Shannon이 말했듯이 前者는 1/100, 後者는 1/150로 선택확률이 돌아온다.[10] 그러므로 語彙가 많을 수록 선택될 機會는 줄어 든다. 물론 各 索引語의 出現頻度에 따라 全體 索引語의 平均 選擇確率은 상승이 되나[11] 어쨌든 索引語彙가 늘어날수록 再現率이 떨어진다. 둘째, 文獻에 附與되는 索引語가 많아질수록 再現率은 減少한다. 실예로써, "벼 흰빛잎마름균을 利用하여 Streptomycin系 藥劑인 Agrepto에 對한 耐性形質의 選拔效果를 調査한 硏究"는 「흰빛잎마름균, Streptomycin, 耐性」으로 索引하면 綱維性이 있으며 索引語도 3個면 足하다. 그런데 이 索引標目으로 出力된 文獻이 100個 以上된다면 여기서 特定性있는 索引語를 몇 個 더 부과해야 할 것이다. 그래서 「흰빛잎마름균,Streptomycin, 耐性」에 「벼, Agrepto, 選拔」을 추가시키면 「벼, 흰빛잎마름균, Streptomycin, Agrepto, 耐性, 選拔」로 索引이 된다. 이 特定性 있는 後者의 索引標目은 前者보다 再現될 確率은 굉장히 떨어진다. 그러므로 索引語彙가 늘

---

8) G. Jahoda, " A Comparison of a Keyword from Title Index with a Single Access Point per Document Alphabetic Subject Index", *American Documentation*, Vol. 20, No. 5, 1969, pp. 377-380.

9) 水平的 증가는 同一水準의 어휘가 늘어나는 것을 말하며 垂直的 증가는 下位 特定語들이 추가됨을 말한다. 즉, 「벼」에 「벼, 보리」와 같이 同位語가 추가됨을 水平的증가. 「농약」에 「농약, 살균제」와 같이 下位語가 투여되는 것이 수직적증가이다.

10) N. J. Belkin, " Information Concept for Information Science," *Journal of Documentation*, Vol. 34, No. 1, (1978), pp. 55-85.

11) $E = -\sum Pi \log pi$ 公式에 의해 선택확률이 상승될 수 있음.

12) 語彙가 組合되어 출현할 확률을 비교하면 다음과 같다. 어휘가 10개 있을 때 그중 3개 어휘가 조합되는 경우는 $10C_3$ 이며, 6개가 조합되는 것은 $10C_6$ 이다. 여기서 보면 전자는 240 가지 후자는 1260 가지이다. 그러므로 전자의 출현확률은 $\frac{1}{240}$ 이고 후자는 $\frac{1}{1260}$ 이 되므로 전자와 후자의 출현비율은 21 : 4 이다. 즉 6개 표목의 색인어보다 재현이 덜 될 확률은 5배나 된다.

어날수록, 文獻에 많이 附與할수록 再現率은 減少한다. 反面에 精度率에 상승하여야 하나 그렇지 않은 것은 바로 세번째 現象에서 起因한다. 索引者와 探索者, 著者와 讀者의 四角 關係에서 語彙使用에 問題가 있는 것이다. 즉, 意味論的인 面에서 고찰이 되어야 한다. 첫째, 水平式 關聯語, 둘째, 垂直式 關聯語를 들 수 있다. 첫째 문제는 앞의 例에서 "벼"와 "흰빛잎마름균"은 水平式으로 관련된 索引語라고 할 수 있으며, 양 單語 中 어느 하나가 省略되면 精度率이 減少한다.[13] 둘째로 垂直式 問題는 "Streptomycin"과 "Agrepto"의 境遇인데, 같은 계열의 上·下語彙이므로 둘 다 또는 둘 중 어느 하나만 索引할 수 있다. 索引者가 판단하여 「Streptomycin」으로 索引을 하였을 때, "Agrepto"를 생각하지 않고 「Streptomycin」으로 접근하는 利用者는 문헌의 特定性 때문에 精度실패를 보게 된다. 한편 「Streptomycin, Agrepto」로 둘 다 索引作成이 되었을 境偶도 「Streptomycin」으로 接近하는 사람은 문헌의 特定性 때문에, 「Agrepto」로 接近하는 사람은 文獻의 網羅性 때문에 精度失敗를 體驗할 수 있으므로 效率이 떨어진다.

이와 같이 색인어휘 수와 문헌에 부여되는 量과 어휘선택에 따라 검색효율-정도·재현율이 下落하는 理由를 살펴보았다. 이것은 情報量이 늘어남에 따라 여과를 最適으로 行하기 위해 索引語를 추가하는데 따른 相馳되는 副作用이다. 그러므로 語彙를 늘리지 않고 特定性을 갖게하는 方式의 考案이 必要하다.

2. 計量群集索引의 提起

特定性과 網羅性을 고루 갖춘 索引이라면 먼저 列擧式 分類시스템을 생각하게 된다. DDC나 KDC는 문헌의 특정성 정도에 따라 체계적 망라성을 해치지 않고 索引을 할 수 있기 때문이다. 그렇다면 分類方式으로 索引體制를 轉換하면 問題는 簡單히 解決되지 않을까? 그러나 앞 章에서 論하였듯이 分類方式은 그 記號의 固定性과 體系展開의 把握의 어려움으로 인해 現代 索引에서는 맞지가 않는다. 그러면 分類와 組合索引을 同時에 並行하는 方法이 떠오른다. 事實 各種

---

13) "벼의 흰빛잎마름균의 상태"를 추적할려고 할시에 문헌의 색인작성이 「벼」나 「흰빛잎마름균」중 한쪽만 되어 있을 때와 질문항에 어느 한 쪽만 기재할 경우는 예를 들어서 「보리, 흰빛잎마름균」, 「벼, 도열병균」같은 문헌이 출력하게 된다.

데이터베이스에서는 分類코드와 키워드를 主題索引言語로 同時 運營한다.[14] 包括的인 質問은 分類記號로, 또 特定的일 때에는 키워드로 探索을 試圖한다. 이 方法은 便利한 方法이나 키워드 自體의 모순을 解決한 것은 아니다. 여기서 筆者는 特殊한 方法으로 특정성과 망라성— 語彙와 記號를 結合시키는 方法論을 생각하였다. 즉, 分類記號가 固定的이고 不便한 것이지만 어휘를 包括하는 長點이 있다. 語彙는 便利하고 即應性이 있으나 反面에 "물"이면 "물"로 接近해야 매칭(matching)이 되는 非融通性이 있다. 그래서 文獻에 配當된 索引語가 많아질 境遇에는 그 非融通性 때문에 逆效果를 發生하는 것은 이미 살펴본 바이다. 그렇다고 特定化되어진 索引표목 중 적당량의 어휘를 삭제하는 方法도 결코 좋은 결과를 초래하지는 못한다. 「벼, 농약」으로 索引된 문헌이 特定化로 인해서 「벼, 피, 살포제, 제초제」로 어휘를 늘렸을 때 再現率을 높이는 名目下에 「피, 살포제」나 「피, 제초제」로 각색이 된다면, 當 文獻의 正確한 內容을 대변한다고 할 수 없으며 따라서 再現·精度率이 떨어지겠다.

그런 관계로 語彙는 줄여서는 아니되며 부득이 줄일 경우는 그 補完을 해주어야 한다.

바로 補完을 하는 方法을 分類記號의 融通性에서 찾는 것이 筆者가 제시하는 方式이다. 상술하면, 특정주제어들은 어휘索引하지 않고 그 特定性 程度를 아라비아 숫자로 나타내어 上位概念인 一般索引語와 結合시켜 使用한다. 이렇게 特定的인 語彙를 數値로 變換시킴으로써, 일단 特定的 어휘의 數量과 이에 따른 非融的 매칭 결과를 無視할 수 있다. 特定的 語彙로써 表意되는 特定度는 數値를 高低로 調整하여 나타낸다.

예를 들면, 病徵은 左側과 같은 系譜를 갖는다. 사실 第二, 第三의 下位語는 特定主題語로써, 색인어로 使用하기에는 너무 망라성이 缺如되어 있다. 그래서 索引語彙는 病徵과 變色만을 選定하고 蒼白 以下는 計量化한다. 즉, 第二下位語는 (5) 第三下位語는 (10)으로 特定度를 매긴다. 그後 索引作成을 할 때 同系列 主題를 包括的으로 다룬 文獻은 「병징」으로 索引하고 次下로 다룬 主題內容은 그 程度에 따라 「變色(0)」,

病徵……項目語
↓
變色……第一下位語
↓
蒼白……第二下位語
↓
淡綠化, 黃變,
銀色化, 白化, 委黃
軟白……第三下位語

14) KORSTIC 발행의 「ORBIT」 manual 참조.

「變色(5)」, 「變色(10)」과 같이 索引한다. 즉, 變色에 대한 一般的인 主題이면 「變色(0)」, 蒼白에 대한 內容이면 「變色(5)」, 그 以下의 特定的인 淡綠化 等을 다룬 文獻이면 「變色(10)」과 같이 索引한다. 이 方法으로 우리는 索引語彙의 垂直的 增加를 막을 수 있으며, 垂平的 인접同位語의 增加도 제지한다.

外型的으로 보아서 本 시스템은 統制語시스템이 運營하는 디소러스內의 同意語 統制形式에 數値를 첨가하여 놓은 것이다. 그러나 內容的으로는 統制語시스템과 다른 點이 두 가지 있다. 첫째, 統制시스템의 어휘통제 목적은 단순히 同意關係에 있는 어휘들이 예를들면 "銀色化"와 "白化"가 自然 그대로 索引이 되어 同一主題文獻이 分散되는 點(自然語시스템의 短點)을 막는데 있다. 그러나 本시스템은 情報의 特定性 있는 여과를 위하여 追加하는 索引語彙의 肥大化를 防止한다. 즉 「창백」에 「은색화」 또는 「백화」가 첨부되어 再現率을 減少시키는 原因을 제거하는데 目的이 있다.

둘째, 統制索引은 水準이 다른 同主題의 文獻이 한 索引語로 集約되는 모순이 있다. 例로서, 「병징, 변색」의 內容이 「변색」으로 統制되고 「창백, 담록화」도 「변색」으로 索引이 되어 精度率이 下落한다. 그런데 本 索引은 그 特定度를 數値로 變換 附與함으로써 精度失敗를 극복하여 주는 點이 統制索引과 다른 長點이다.

本 方法의 特徵은 다음의 몇 가지로 要約된다(앞 章에서 제기된 난점과 비교하여 서술한다).

(1) 索引語彙의 增加를 억제하는 效果가 있다. 새로 나타나는 特定度가 深한 語彙들은 數値로 變換을 시켜 그 上位主題語에 傳加함으로써, 急增하는 索引語彙의 數量을 제어한다. 즉, "銀色化, 白化, 軟白, 蒼白"같은 特定的인 語彙를 "變色"에 蓄約시키므로 그만큼 索引語彙를 절약할 수 있고 相對的으로 選擇確率이 높아져 再現效果가 높아진다.

(2) 文獻에 附與하는 索引語 數를 特定性을 해치지 않고 줄일 수 있다. 情報量이 尨大하여지면 適合文獻을 찾기 위하여 索引語를 追加하는데 그 결과 再現率을 떨어뜨리므로 特定性은 있으되, 어휘量이 늘어나지 않는 本 方法이 必要하다. 즉, 「변색, 창백, 은색화」의 수직적 증가나 「은색화, 백화, 연백」과 같은 수평적 증가를 「변색(5)」, 「변색(10)」과 같이 한 索引語와 特定度 數値로 나타내어 特定性이 있으면서 망라성을 겸유한 색인표목을 형성한다. 그러므로 情

報源을 걸러 주면서도 재현율을 최대한 살린다.

(3) 語彙選擇의 垂直的 混同을 피할 수 있다. 예를 들면 "病徵, 變色, 蒼白 淡綠化"의 順으로 이어지는 系列語들을 그 特定度에 따라 明確히 숫자화하기 때문에 索引者나 探索者의 趣向에 따른 어휘선택에서 오는 精度失敗는 最大한 防止하게 된다.

Ⅳ. 語彙集 構成과 情報價 算定方法

앞서 提起한 시스템을 成功的인 것으로 만들기 위해서는 키워드 索引語 및 特定主題語와 그 情報價를 收錄한 語彙集이 컴퓨터 內에 所藏되어 있어야 한다.

따라서 語彙集의 構成體系와 키워드의 選定, 그리고 特定主題語의 選別 및 情報價 算定方法이 대두된다. 本 章에서는 기존의 디소러스(索引語彙集) 및 自動式 索引語 拔萃方法을 利用하여 語彙集의 편성과 語彙選定을 고찰한다.

情報價의 측정은 自動分類(Automatic Classification)의 클러스터링(Clustering) 技法을 應用한 系譜順 方法과 Yu[15]와 Jones[16]가 제시한 精度加重法(Precision Weighting), 그리고 Salton의 文獻分離價를 利用한 方法을[17] 통해 算出한다.

1. 語彙集 構成

디소러스는 문헌 또는 質問의 內容을 分析하여 索引作成을 할 때에 어휘를 參照하는 도구이다. 그러므로 그 構成體系가 索引者 및 利用者가 쉽게 判讀할 수 있는 形態로 作成되어야 한다.

대체로 데이터베이스(Data base : 컴퓨터 情報檢索 시스팀)에서 使用하는 디소러스의 構成要素를 文字, 數字로 조립되는 키워드와 키워드 간의 上·下 同等

---

15) C.T.Yu, "Precision Weighting-An Effective Automatic Indexing Method", *JACM*, Vol.23, No.1, pp.76-88 (1976).

16) K.S.Jones, "Search Term Relevance Weighting Given Little Relevance Information" *J. of.Doc.*, Vol.35, No.1, pp.30-48 (1979).

17) G.Salton, "A Theory of Term Importance in Automatic Text Analysis", *JASIS*, Vol.26, No.1, pp.34-44 (1975).

관계를 表示하는 特殊記號로 이루어진다. 이 特殊記號와 또 그 記號에 따라 묶어지는 키워드간의 群集形式은 디소러스에 따라 여러 種類가 있다. 本 語彙集에는 넓은 通用性을 갖는 TEST [18] 와 AGRIS [19] 의 特殊記號와 總系譜를 그 群集形式에 포함하고 있는 AGRIS의 方式을 따라 體系化한다. AGRIS의 디소러스는 아래의 예와 흡사한 總系譜를 갖고 있다.

植物……系譜語(Top term)

BT : 生物……上位語

NT_1: 經濟植物……等一下位語

uf : 作物……NT_1의 同意語로 NT_1을 索引語로 使用

NT_2: 콩科作物……等二下位語

uf : 豆類作物

NT_3: 콩……等三下位語

RT : 豆腐……NT_3의 關聯語

<BT, NT, uf, RT가 관계를 표시하는 특수 기호임>

위 총계보에서 출현한 각각의 키워드들은 그 키워드를 項目語로 삼고, 그 外 키워드들의 上·下 關聯관계를 적당히 類聚시켜준

콩科作物

uf : 豆類作物

BT_1 : 經濟植物

NT_1 : 콩과 같은 項目系譜를 編成하고 있다. 이 항목계보와 총계보는 디소러스 내에서 계보어와 항목어의 字母順으로 배열되고 보통 항목계보는 키워드들은 4～5個 類聚하여 준다. 그러므로 본 어휘집도 총계보와 항목계보를 병존시키면서 항목계보의 키워드를 4～5개로 限定한다.

한편, 特定主題語는 그것이 소속하는 총계보의 마지막 키워드에 連結을 시켜 그 下位語로 定하며 표시기호는 알파벳 大文字 UF로 한다. 예를 들면,

病徵

NT_1 : 解部的 變化

NT_2 : 畸形

---

18) J. Aitchison, *Thesaurus Construction* , London ; Aslib, 1972, pp.51—

19) AGRIS Coordinating Center, *AGROVOC* Rome ; the Center, 1981.

uf : 贅生

UF : 貫生(5)、

: 捲葉(5)----- 情報價

: 帶化(5) 와 같다.

특정 주제어들도 항목어로 내주어 키워드와 같이 혼합 배열한다. 그러나 항목계보에서 단지 특정주제어와 계보어(Top term) 그리고 마지막 종속 키워드를 "貫生:USE崎形(5):TT병징"과 같이 병기하는데 그친다.

키워드와 特定主題語의 선정방법은 먼저 키워드는 現在 그 主題分野의 데이터베이스에서 活用되고 있는 디소러스 안의 어휘들로 定한다. 특정주제어는 自然言語시스템에서는 출현하나 디스러스에 실리지 않은 어휘를 우선 발췌한다. 그리고 문헌에서 多頻度 發生하면서, 이 논문은 '결론', '요약', '결과' 등과 같이 나타나는 主題意味語와 표제에 나타난 특정주제어들을 두번째 선택한다. 그 외에 後述할 Salton의 문헌분리가가 높은 단어들, Harter의 two-poisson 確率함수에 따라 선별된 어휘들 중에서 특정주제어를 세번째로 선택한다.

한편 본 어휘집은 색인어, 즉 키워드의 정보가는 기록하지 않고 특정주제어들의 정보가만 앞의 방법으로 병기한다.

2. 情報價 算定法

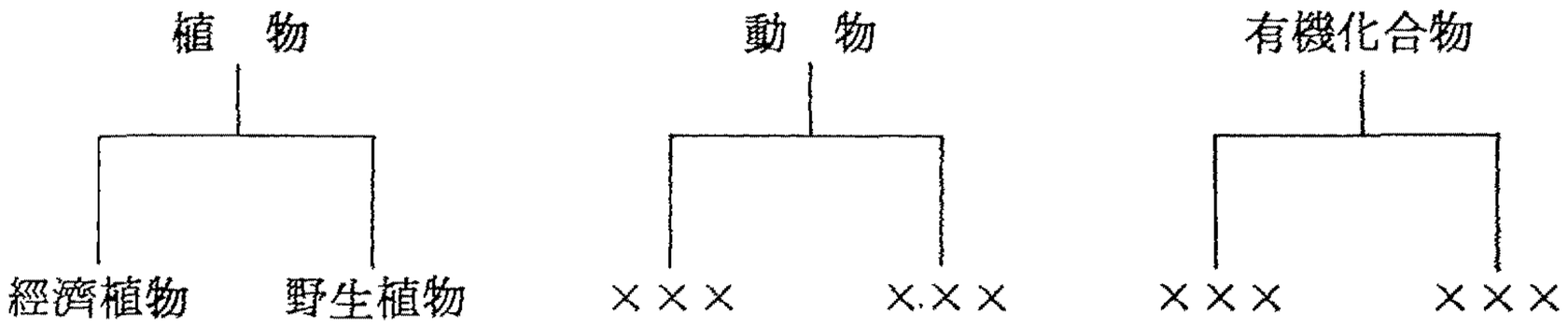
(1) 系譜順 方法

本 系譜順方法에서는 測定에 있어 網羅的 語彙(上位語)에서 特定的 語彙(下位語)로 내려 갈수록 情報價가 많은 것으로 規定한다. 그러나 이 順序가 그대로 通用되는 것은 아니다. 감자, 暮類作物은 감자가 特定的이지만 暮類作物이 더 어려운 語彙로 認識될 수 있으며 또는 同等하다고 볼 수도 있다. 더구나 系譜가 다를 때에는 各系譜들의 키워드 序列은 相互 影響力이 없다. 다시말하면 A系譜는 3階層, B系譜는 6階層일때 A와 B 系譜의 序列은 서로 調整이 필요하여 진다. 이 調整을 위해 自動分類에서 使用하는 클러스터링技法을 導入하여 各 키워드들의 同類를 把握하고 그 同類에 모인 키워드들은 서로가 同等序列이라고 간주하여 같은 情報價를 附與한다. 그 方法은 다음과 같이 進行한다.

1) 語彙集表 完成

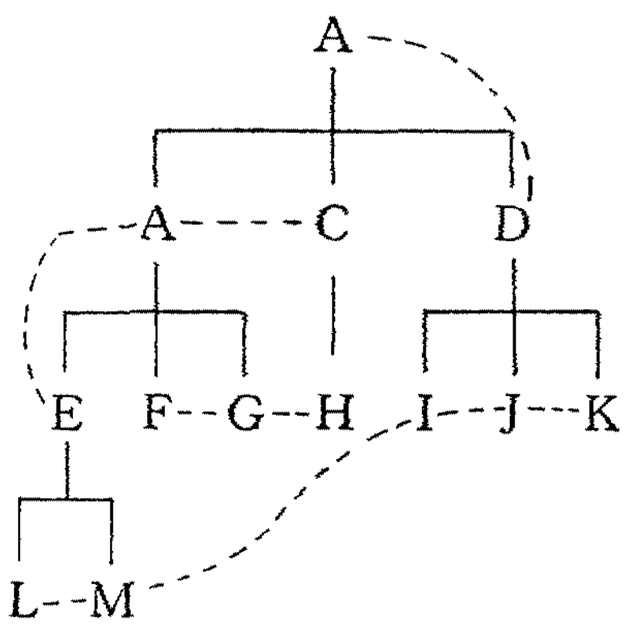
2) 各 系譜別로 組織圖表를 使用하여 整理한다.

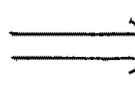
例)



3) 各 系譜內의 同等價를 算出하는데 클러스터에 의하여 그려진 그래프를 使用하여 同類를 묶어 整理한다.

例) "가" 계보





"가" 계보

A · D

↓

E · B · C

↓

F · G · H

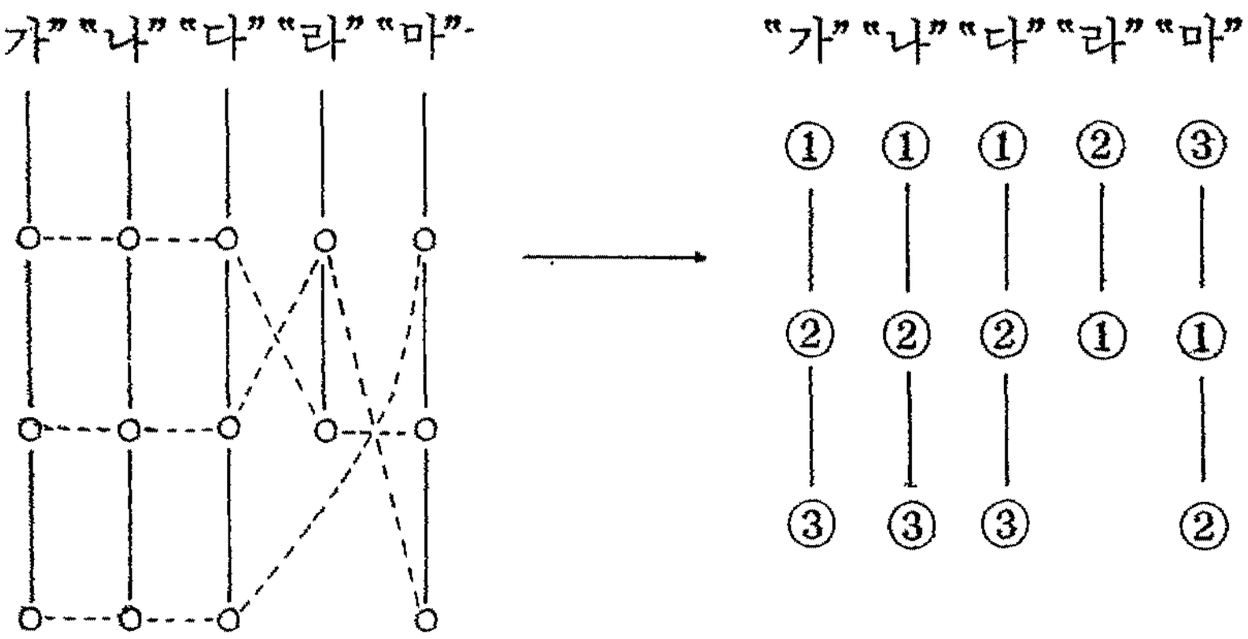
↓

L · M · I · J · K

(點線이 同類로 모인 表示)

4) 異系譜間의 同等價도 클리스터에 의해 그려진 그래프를 利用하여 同類를 묶고 가장 많이 順序的으로 連結된 系譜들의 順序에 따라 情報價를 附與한다.

例)



5) 語彙集表에 情報價를 記錄한다.

(2) 精度加重値利用法

加重値를 주는 法은 대체로 4가지로 分類한다.

① 索引者가 文獻內容을 읽고 拔萃된 索引語의 比重을 測定하는 法

② 拔萃된 索引語가 該當文獻內에서 出現한 頻度에 따라 比重을 計算하는 法

③ 文獻群內에서 그 索引語의 檢索效率에 따라 比重을 주는 法

④ 문헌군내에서 그 索引語가 配當된 文獻數의 比率에 依한 法이 있다. 이중 ③번의 方法을 應用하면 그 單語의 重要性을 測定할 수 있다. 어느 主題分野에서 한 單語가 保有한 檢索效率은 그 索引語의 重要度, 즉 利用者의 만족도이므로 이를 C. T. Yu와 K. S. Jones는 精度加重値라고 하였다. 예를 들면, 곰팡이 眞菌, 子囊菌이라는 索引語가 있다고 하자. 研究者가 이들 索引語로 문헌을 探索하였을 때 各 索引語에 따라 出力한 情報源에 대한 利用者의 滿足度는 곰팡이→진균→자낭균의 順으로 精度率이 높아질 것이다. 왜냐하면 곰팡이 眞菌은 너무 包括的이어서 再現·精度率이 같이 下落할 것이나 子囊菌은 細分化된 特定性이 있으므로 보편적인 만족도를 줄 것이다. 따라서 精度率은 곰팡이(2), 眞菌(4), 子囊菌(6)이라는 임의의 숫자로 추정할 수 있다. 그러므로 索引語가 갖는 검색효율치는 그 단어의 중요도이며 情報價라고 할 수 있다.

精度加重値는 아래의 C·T·Yu와 K.S.Jones의 精度加重値算定公式에 의해 計算할 수 있다.

Yu의 精度加重値 算定公式

$$Pi = \frac{ri}{R - ri} \cdot \frac{I - hi}{hi}$$

Jones의 공식은

$$W = \log\left(\frac{r + 0.5}{R - r + 0.5} \cdot \frac{N - n - R + r + 0.5}{n - r + 0.5}\right)$$이다.

여기서 R은 適定문헌수, I와 (N−n)은 비정적 문헌수, ri와 r은 索引語 i를 포함하고 있는 비적정 문헌수, n은 색인어 i를 포함하고 있는 문헌수, N은 문헌의 총수임.

(3) 文獻分離價에 依한 法

Yu와 그외 동료들이 제시한 단어의 문헌분리가에 의한 索引語選定技法은 또 索引語의 重要度－情報價를 測定할 수 있는 根據를 提供한다. 임의의 單語 K의 分離價 計算은 문헌 클러스터링 중 단어간의 類似性을 測定하는 公式으로 그 절차는 다음과 같다.

Salton의 文獻分離價 算定 公式

$$SIM(Di, Dj) = \frac{\sum_{k=1}^{t} dik \cdot djk}{\left[\sum_{k=1}^{t} (dik)^2 \cdot \sum_{k=1}^{t} (djk)^2\right]^{\frac{1}{2}}} \quad \cdots\cdots ①$$

$$\overline{SIM} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} SIM(Di, Dj) \quad \cdots\cdots ②$$

$$DVk = \overline{SIMk} - \overline{SIM} \quad \cdots\cdots ③$$

여기서, Di →문헌군내의 문헌 i, Dj →문헌내의 문헌 j, dik →문헌 Di 에 출현한 k 단어의 출현 빈도, dik →문헌 Dj 에 출현한 k 단어의 빈도임.

위 공식에서 문헌분리가는 ①식에서 두 문헌 간의 類似係數 $SIM(Di, Dj)$를 구하고 ②식에서 전체 문헌의 平均 밀집도 $\overline{SIM}$을 산정한 후 單語 k를 제거하고저 똑같은 방법으로 문헌의 밀집도 $\overline{SIMk}$를 내어 ③식에서 문헌 분리가 $\overline{DVk}$를 구한다. 이때 값이 양수이면 좋은 색인어, 음수이면 나쁜 색인어로 區別한다. 이 분리가는 그 단어가 제거된 상태에서 얼마나 문헌들의 유사성이 높아지는가를 측정한 것으로 실험결과 문헌을 밀집시키면 自動分類를 沮害하는 단어로 생각하여 색인어로 부적합하게 여긴다. 이것을 응용하면 문헌을 분류할 수 있는 힘에 따라 그 단어의 중요도,즉 정보가를 매길 수 있다.

Ⅴ. 結　　論

이와 같이 계량화된 群集索引語의 필요성을 제기하고 그 특정주제어의 정보가

를 계량화하는 방법론을 서술하였다. 그러나 Ⅲ章의 3가지 方法으로 측정된 정보가의 효율성은 실제로 운영한 후 평가와 인과 분석에 의한 재수정 작업, 즉 피드·백이 중요하다. 다시 말하면 본 방법은 키워드색인에서 문제가 되는 재현·정도율의 보강문제와 초록의 문장을 일일이 확인하여야 하는 이용자의 수고를 덜어 주기 위한 방법의 모색으로 제시가 되어진 색인 시스템이란 점을 강조하고 싶다. 즉 정보량의 급증으로 말미암아 필연적으로 변화를 겪어야 하는 키워드색인 내에서 특정성있는 주제 의미어를 현 키워드에 부가하는 고정적 方法에서 전환하여 그 특정성에 알맞는 정보가를 대치함으로써 가변적 융통성 있는 시스템을 제시하고자 하였다. 그리고 순수한 측면에서 문헌의 정보량을 어휘의 의미적인 면에서 측정하려는 노력의 일환으로도 생각할 수 있다.

〈參 考 文 獻〉

1. 사공 철, 「情報檢索論」, 서울 ; 亞細亞文化社, 1977.
2. 이 재철, "신문기사 색인법의 이론과 실제",「人文科學」, Vol.22 (1969), pp.83 - 99.
3. 최 성진, 「情報學原論」, 서울 ; 亞細亞文化史, 1976.
4. AGRIS Coordinating Center, *AGROVOC*, Rome ; the Center, 1981.
5. Aitchison, J., *Thesaurus Construction*, London ; Aslib, 1972.
6. Belkin, N.J., " Information Concept for Information Science ", *Journal of Documentation*, Vol.34, No.1 (1978), pp.55-85.
7. Jahoda, G., " A Comparison of a keyword from Title Index with a Single Access Point per Document Alphabetic Subject Index ", *American Documentation*, Vol.20, No.5 (1969), pp.377-380.
8. Jones, K. S., " Search Term Relevance Weighting Given Little Relevance Information ", *Journal of Documentation*, Vol.35, No.1 (1979), pp.30-48.
9. Salton, G., *Dynamic Information And Library Processing*, Englewood Clifts ; Prentice-Hall Inc., 1975.
10. Salton, G.," A Theory of Toun Importance in Automatic Text Analysis ", *JASIS*, Vol.26, No.1 (1975), pp.33-44.
11. Wall, R. A., " Intelligent Indexing And Retrieval : A Man-Machine Partnership ", *Information Processing & Management*, Vol.16, pp.73 -90.
12. Yu, C.T.," Precision Weighting-An Effective Automatic Indexing Method", *JACM*, Vol.23, No.1 (1976), pp.76-88.