自動索引方法과 自動索引 시스템 性能

宋 美 蓮
(情報資料室)

〈 차 례 〉

Ⅰ. 序 論

컴퓨터가 우리 生活의 一部分이 되면서 機械가 무엇을 할 수 있을 것이며, 人間과 機械間의 問題를 어떻게 解決할 것인가 하는 것이 오래전부터 問題가 되어 왔다. 특히 人間의 知的 作業을 機械, 즉 컴퓨터가 얼마나 對等하게 遂行해 낼 것인가 하는 것이 중요한 疑問이었다. 과거 20여년동안 情報學分野에서도 自然語 文章의 知的 處理作業에 컴퓨터를 使用하려는 試圖로 많은 硏究가 행해졌고 오늘날에도 많은 關心을 기울이고 있다.

이러한 硏究들은 人間이 索引을 하는 복잡한 知的 作業을 컴퓨터가 대신 遂行함으로써 보다 바람직한 情報檢索을 하려는데 그 目的을 두고 있다. 그래서 본

研究에서는 먼저 지금까지 發表된 自動索引方法을 살펴보고 自動索引시스템과 手作業에 의한 索引시스템간의 檢索效率을 先行研究를 통해 比較해 보았다.

Ⅱ. 自動索引시스템

1. 索引의 意義와 役割

索引의 語源은 라틴어의 'Indicare'로 '가리킨다,' '지시한다'의 意味를 갖는 것으로 現代의 索引의 槪念도 情報의 位置를 指示하고 찾아 주는 道具인 것이다. 맥콜빈(L.R.McColvin)[1]은 索引의 役割을

① 特殊한 項目에 대해 參照를 해 주고

② 하나의 統一된 表에 따라 文獻에 收錄된 內容을 順序대로 排列해 주고

③ 各 項目들간의 關係를 밝혀 줄 수 있어야 하며

④ 省略된 內容도 밝혀 주는 것이라 했다.

보르코(H. Borko)[2]는 이에 덧붙여

⑤ 質問에 解答을 提示하고

⑥ 主題 分野에 대한 包括的인 槪要를 提示하고

⑦ 案內의 役割까지 한다고 했다.

이처럼 索引은 情報源과 情報를 얻고자 하는 情報入手者 사이에 位置하여 同一한 主題의 情報를 選別해 주고,그 選別된 資料의 位置를 指示해 주는 채널로[3] 尨大한 量의 情報로부터 원하는 情報만을 걸러내 주는 필터의 役割을 한다.

文獻을 索引한다는 것은 文獻이 包含하고 있는 중요한 槪念들을 적절한 코우드, 즉 索引語로 變換시켜 주는 일을 말하며 索引 自體는 索引 作業 結果 生産되는 것으로 索引語와 文獻의 識別要素(Descriptor)들의 集合이 된다. 文獻의 識別要素란 特定한 索引語 아래 索引된 文獻의 請求番號를 비롯한 書誌事項이 되며 機械可讀型파일에서는 文獻의 登錄番號가 識別要素가 된다.

理想的인 索引은 雜音이 전혀 없는 커뮤니케이션 채널과 같아서 雜音이 가능

---

1) L.R. McColvin, "The Purpose of Indexing," *The Indexer*, vol.3, no.2, pp.31-35.

2) Harold Borko, *Indexing Concepts and Methods*, New York : Academic Press, Inc.,1978.

3) 정영미, "색인이론과 실제,"「연세논총」, vol.7, 1980. pp.21-35,

한 한 排除되어야 文獻의 主題가 정확히 表現되며 利用者가 情報源에 대해 갖고 있는 不確實性을 最小化시켜 주는 것이어야 한다. 이렇게 적절한 索引을 生産하기 위해서는 索引者의 정확한 主題 分析과 主題概念 抽出, 符號化, 索引 技術의 正確性 등이 要求되며 이와 더불어 索引 語彙의 適合性과 索引政策의 妥當性 與否도 索引에 影響을 미친다.

2. 手作業索引시스템

索引은 圖書館에서 文獻의 取扱過程中 어떤 目錄을 할 것인가를 決定하기 위해 考案한 것으로 傳統的인 圖書館에서의 索引語는 內容指示를 해 주는 主題 分類와 索引上의 位置를 알려주는 적절한 請求番號가 이에 該當될 것이다.

手作業에 의한 索引 節次[4]는

① 情報의 內容을 나타내는 用語 選定

② 重要度에 따라 加重値 附與

③ 各 用語의 文章內에서의 役割에 따라 적절한 機能 表示

④ 用語들간의 關係(類似語, 關聯語 등의 階層的 關係)를 糾明해 주어야 한다.

實際에서 索引作業은 거의가 訓練된 사람이 手作業으로 행하는데 이때 辭典이나 補助 리스트, 또는 索引語와 單語들간의 關係를 정하는 方法과 順序를 정해 놓은 補助 道具를 使用하게 된다. 索引作業에는 상당한 人的 作業이 필요하며 索引者는 索引語彙와 그 實際를 잘 알고 있어야 할 뿐만 아니라 文獻集團의 性格도 잘 알고 있어야 한다. 더 나아가 效果的인 索引은 시스템에서 豫想되는 利用者의 質問 形態도 反映해야 한다. 그런데 手作業索引의 경우 索引者가 各 文獻마다 각각의 判斷을 하게 되므로 索引者가 經驗이 不足하면 不適當한 用語나 잘못된 加重値의 附與, 중요하지 않은 用語들간의 關係 規定같은 問題가 發生하여 일관된 索引 作業이 不可能하다. 原則的으로 經驗과 熟練된 索引者만 있으면 매우 成功的인 索引을 만들어 낼 수가 있어야 하는데 實際 索引作業은 그렇지가 못하다. 人間에 의한 傳統的인 索引方法의 問題點은 같은 文獻을 서로 다른 索引者가 索引하게 되면 다른 索引語가 選定되고 또 索引作業이 정확하게 遂行되

---

4) G. Salton, *Dynamic Information and Library Processing*, Englewood Cliffs : Prentice-Hall, Inc., 1975.

었다 하더라도 한사람의 索引者가 索引하기에는 資料의 量이 너무 많다. 이러한 索引作業의 不一致는 檢索性能에도 害를 미치게 된다. 그래서 생각하게 된 것이 컴퓨터를 이용한 자동 색인 시스템이다.

3. 自動索引시스템

自動索引은 컴퓨터가 符號나 順序를 認知할 수 있다는 데서 出發한다.[5] 컴퓨터는 自然語로 된 原文을 단순한 符號의 連續으로 보아 알고리즘(Algorithm)을 써서 각 文獻에 똑같은 索引作業을 해 줌으로써 그 알고리즘이 變更되지 않는 한 同一 文獻에 同一한 索引語를 符與하여 手作業에서 發生하는 不一致의 問題를 解決해 준다. 自動索引은 原文의 本文이나 書名抄錄 등을 그 分析對象으로 하여 單語의 發生頻度, 單語나 文章의 要素, 文章內에서의 同時 發生程度, 單語를 構成하는 文字 形態 등을 利用한다. 自動索引은 科學·技術關係 文獻에 더 빨리 使用되었고 情報爆發을 統制管理할 目的으로 시작되었으며 그 바탕은 컴퓨터의 利用이 손쉬워졌고 컴퓨터가 符號 操作으로 數字나 單語를 處理할 수 있다는 認識과 함께 " Computational Linguistic "이라 불리우는 새로운 學問의 出現으로 言語의 構造와 意味를 컴퓨터가 分析 可能하게 되었다는데 있다.

이처럼 컴퓨터가 文獻을 自動으로 分析하고 索引語를 附與하며 質問을 分析하여 適合文獻을 檢索하는 自動索引시스템의 利點이라면

① 더욱 일관된 索引作業

② 저렴한 作動費用

③ 신속한 文獻의 利用

④ 自體開發시스템의 可能性

등을 들 수 있겠다.

Ⅲ. 自動索引方法

自動索引方法은 1950년대 중반 룬(H.P.Luhn)[6]이 文獻과 質問에 포함된 語彙

---

5) Susan Artandi, "Machine Indexing : Linguistic and Semantic Implications," *JASIS*, vol.27, no.6, July/Aug. 1976, pp.235-239.

6) H.P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM J. of Research and Development*, vol.1, no.14, Oct.1956, pp.309-317.

를 內容分析의 目的으로 利用한 것이 그 효시로 單語의 數를 세어서 索引語를 選定하는 가장 基本的인 것이었다.

살튼(G.Salton)[7]은 우선 索引方法을 " Word Indexing "과 " Subject Indexing "으로 나누어서 說明하고 있는데 " Word Indexing "은 " Derived Indexing "이라고 하며 著者가 本文에서 사용한 單語中에서 索引語를 選擇해서 그것을 單語와 構文으로 連結해서 사용하는 方法이며, " Assigned Indexing "라고도 하는 " Subject Indexing "은 著者가 사용한 단어를 特定한 主題名으로 表現 修正하여 文獻의 主題로 이끌어 주는 것으로 이 또한 單語로 表現되는 것이나 Word Indexing "과는 엄연히 區別된다. 그러나 " Subject Indexing "도 사람의 知的인 勞力을 加하고 發展된 言語學 技法, 즉 自動的인 語義分析, 構文分析을 利用해서 할 수 있다. 여기서는 보르코[8]가 區分한대로 統計的 方法(Statistical or Frequency Analysis)과 構文·語義的 方法(Syntatic and Semantic Analysis)으로 크게 나누어서 說明하고자 한다.

1. 統計的 方法

(1) 頻度(Frequency)測定方法

이 方法은 한 文獻에서 여러번 나오는 單語가 그 文獻의 主題를 가장 잘 나타내리라는 假定下에 컴퓨터 프로그램에 의해 한 文獻에 收錄된 모든 單語를 뽑아내어 이들을 頻度數에 따라 그룹을 짓고 그 안에서 字母順으로 排列하며 이때 前置詞나, 代名詞, 冠詞, 接續詞 등의 機能語는 不用語로 定義하여 세지 않도록 한다. 이처럼 單語의 數를 세는 方法은 1957년에 룬[9]이 가장 먼저 提案한 것으로 自動索引의 基本 技法이다. 그런데 이때 어느 程度의 頻度를 갖는 單語가 중요한 單語인가를 定하는 問題가 發生한다. 著者는 같은 單語를 계속 重複해서 쓰려 하지 않으며 索引者가 索引語로 選定할 單語의 頻度數를 任意대로 設定하기 때문이다. 어떤 경우에는 빈도리스트의 가장 중간에 오는 範圍의 單語를 選擇하기도 하나 이는 시스템마다 다르다. 룬은 文獻 i에 대한 用語 k의 出

7) G. Salton, *op.cit.*
8) Harold, Borko, *op.cit.*
9) G. Salton, *op.cit.*

現 頻度 F^k를

$$F^k = \sum_{i=1}^{n} f_i^k$$

(n = 문헌집단내의 총 문헌수)

라고 規定하였으며 가장 頻度數가 많은 單語와 가장 頻度數가 적은 單語는 除外시켰다. 높은 빈도의 用語는 보편적인 것으로 檢索時 質問과 文獻의 索引語間에 一致되는 것이 많아져서 適合한 文獻이 많이 檢索되어 再現率은 增加하나 正確率은 떨어진다. 相對的으로 낮은 頻度의 用語는 덜 매치(match)되어 再現率은 떨어지나 正確率은 높아진다. 이처럼 索引語로 選擇할 單語의 頻度數를 정할 基準이 없어 單語의 數만을 세는 方法으로는 未洽하다. 이러한 問題를 解決하기 위한 方法이 여러가지가 있는데 우선 相對頻度(Relative Frequency)를 測定하는 方法을 들 수 있다. 이는 1965년 다메로우(F.J.Damerou)[10]가 提案한 것으로 文獻의 標本 集團을 정해 用語의 頻度數를 정해 놓고 索引하고자 하는 文獻에서의 發生 頻度數가 많으면 그 單語를 索引語로 選定하는 方法으로 이 또한 實際로 標本의 單語와 比較할 文獻 集團의 處理가 어렵기 때문에 사용이 쉽지 않다.

(2) Signal-Noise Calculation

샤논(Shannon)의 情報理論에서 類推한 것으로[11]쿠퍼(W.S. Cooper)[12]가 提案한 方法이다. 이는 n개의 文獻集合에 대한 k 用語의 雜音 N^k는

$$N^k = \sum_{i=1}^{n} \frac{f_i^k}{F^k} \cdot \log \frac{F^k}{f_i^k}$$

로 나타내며 Signal S는

$$S^k = \log F^k - N^k$$

가 된다. 그래서 용어 k가 각 문헌에 한번씩만 나오면 모든 $f_i^k = 1$이 되어

---

10) F.J. Damerou "An Experiment in Automatic Indexing," *American Doc*, vol.16, no. 4, Oct.1965, pp.283-289.

11) 평균정보량 $= -\sum_{k=1}^{t} Pk \log Pk$

12) W.S. Cooper, "Is Interindexing Consistency a Hobgollin ?" *American Doc.*, vol.20, no.3, July 1969, pp.268-278.

$$N^k = \sum_{i=1}^{n} \frac{1}{n} \cdot \log \frac{n}{1} = \log n$$

으로 $F^k = n$ 인 경우 Signal S^k 는 " 0 "가 된다.

반대로 集中 分布를 갖는 用語는 頻度 F^k 를 갖고 한 文獻에만 나타난다면 雜音이 " 0 "가 되고 Signal 은 最大가 된다. 이처럼 Signal 과 雜音의 關係를 測定하여 索引語를 選定하기도 하는데 이의 基準은 시스템의 性能, 즉 檢索效率 정도에 따라 달라지게 된다.

(3) Discrimination Value

文獻集團내에서의 用語間의 거리를 測定하는 것으로 文獻 集團內의 文獻들이 얼마나 떨어져 있는가에 따라 용어의 값을 정해 높은 값, 즉 分離度가 높은用語를 索引語로 選定하는 方法이다.[13),14)] 우선 文獻 集團內의 모든 文獻雙들간의 類似値(Similarity)를 測定하여 그 文獻集團의 平均 類似値 $\bar{S}$ 를 구한다.

$$\bar{S} = C \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} S(D_i, D_j)$$

〈$S(D_i, D_j)$는 문헌 Di 와 Dj 간의 유사치, C는 상수〉

n 개의 문헌 모두가 獨創的인 것일 경우의 각 文獻雙들간의 類似値는 "1"이 되어 이때의 平均 類似値가 最大가 될 것이다.

용어 k 가 평균 頻度數를 갖는 문헌 $\bar{D}$를 Centroid로 정하고 각 문헌을 문헌 $\bar{D}$와 비교하여 平均 類似値를 구하면 더욱 效果的이다.

$$\bar{S} = C \sum_{i=1}^{n} S(\bar{D}, D_i)$$

용어 k 의 Discrimination 값 V_k 는

$$V_k = S_k - \bar{S}$$

로 V_k의 값이 높은 것이 좋은Discriminator로 좋은 索引語가 될 수 있다.

---

13) G. Salton, and C.S. Yang, " On the Specification of Term Values in Automatic Indexing, " *J. of Doc.*, vol.29, no.4, Dec.1973, pp.351-372.

14) Jones K. Spärck, " A Statistical Interpretation of Term Specificity and Its Application in Retrieval ," *J. of Doc.*, vol.28, no.1, 1972, pp.11-21.

(4) 加重値(Weighting)

단어의 중요도에 따라 加重値를 附與하여 索引語를 選定하는 方法으로 加重値 附與 方法에는 여러가지가 있다. 첫째, 文獻에 나타나는 單語의 位置, 즉 書名, 각 文章의 처음, 끝 등의 位置에 따라 加重値를 附與하는 方法으로 本文에 나오는 單語보다는 書名에 나오는 單語에 더 높은 加重値를 준다든가 하는 方法이다. 또는 特定 單語의 頻度數에 따라 加重値를 附與하기도 하는데 이 또한 發生 頻度가 높은 單語가 그 文獻의 主題를 잘 나타내리라는 假定下에서 出發한다. 또 한가지 方法은 文獻頻度, 즉 各 文獻에서 빈번히 發生하는 單語보다 頻度가 낮은 單語에 높은 加重値를 附與하는 方式[15]으로 이는 索引語의 特定性은 文獻의 數와 關係가 있으라는 槪念에서 出發한다.

(5) 相關係數(Association Value)

單語間의 關聯性 程度를 計算하는 方法으로[16],[17] 같은 文獻內에 두개의 單語가 밀접하게 나타나는 程度에 따라 索引語로 選定해 준다. 알고리즘에 의해 한 文獻에서 두 用語가 同時發生하는 頻度를 計算하여 相關係數를 구하는데 用語 A와 B가 文獻에서 여러번 함께 나오면 두 用語는 높은 값을 갖게 된다.

用語 j와 k간의 상관계수 $f(j, k)$는

$$f(j, k) = \frac{\sum_{i=1}^{n} f_i^j \cdot f_i^k}{\sum_{i=1}^{n} (f_i^k)^2 + \sum_{i=1}^{n} (f_i^k)^2 - \sum_{i=1}^{n} f_i^j \cdot f_i^k}$$

(f_i^j : j 용어의 발생빈도
f_i^k : k 용어의 발생빈도)

의 公式에 의해 구하고 그 결과 나온 "0"에서부터 "1"까지의 相關係數로

---

15) *Ibid.*

16) L.B. Doyle," Indexing and Abstracting by Association," *American Doc.*, vol.13, no. 4, Oct. 1962, pp.378-390.

17) H.E. Stiles, " The Association Factor in Information Retrieval", *J. of ACM*, vol. 8, no.2, Apr. 1961, pp.271-279.

두 用語雙間의 用語相關行列을 만들어 索引語를 選定하는 方法이다.

2. 構文·語義的 方法

앞에서 언급한 統計的 方法은 1960년대 중반에 發展하여 自動索引法의 基本이 되고 있다. 물론 構文分析과 語義分析의 利點은 認識하고 있었으나, 言語學者와 情報學者間에 긴밀한 접촉이 없었던 관계로 덜 발전했다. 統計的 方法은 많이 사용된, 즉 頻度數가 높은 用語가 文獻의 內容을 가장 잘 나타내리라는 假定下에서 出發한데 반해 構文 分析은 文章內에서 單語의 役割, 즉 그것이 名詞로 쓰였는가 혹은 動詞로 쓰였는가하는 文法的인 區分, 役割을 糾明하는 것이며 文章들간의 關係를 糾明하는 것이다. 촘스키(N. Chomsky)[18]가 提示한 言語學的 모델은 言語의 表面的 構造와 內部的 構造를 區分하고 있는데, 예를 들어 "Mary went home with John"과 "Mary and John went home together"라는 文章을 볼 때 表面的 構造는 다르지만 그것이 포함하는 意味(deep structure)는 같다는 것이다. 自動索引의 경우 이 問題는 더욱 복잡해진다. 語義分析의 目的은 文獻이나 本文의 位置, 즉 抄錄같은 것의 內容을 含蓄하는 單語나 主題를 分析하고자 하는 것으로 몇가지 方法이 있다. 우선 檢索時 再現率을 높혀 주는 傳統的인 方法으로 用語를 切斷하는 方法이 있다.[19] 이것은 같은 語根을 갖는 用語들을 모아 줌으로써 매치되는 用語의 數를 늘려 주는 方法인데 用語의 切斷을 잘못할 경우 檢索失敗의 原因이 되고 있다. 英語의 경우에는 같은 意味의 用語에서 語根의 變化가 있기는 하나(예, mouse-mice, absorb-absorption, hop-hopping, ease-easier 등등) 自動 接頭(尾)辭 切斷 프로그램에서는 별다른 問題가 없다. 이 方法에는 右側 切斷(Right-truncation), 左側切斷(Left-truncation),兩側切斷(Both-truncation)그리고 Infix-truncation[20]이 있다.

그 밖에 本文의 單語나 句를 文獻의 索引語로 정하기 위해 이미 만들어 놓은 디소러스(Thesaurus)를 사용해서 그 안의 單語나 句와 比較하여 選擇하는 方法

---

18) N.Chomsky, "Three Models for the Description of Language," *IEEE Transaction on Information Theory*, vol.2, no.1, 1956, pp.113-124.

19) Jones, Sparck, "Automatic Indexing," *J.of Doc.*, vol.30, no.4, Dec. 1974, pp.393-432, Word Truncation, Keyword Nomalization이라고 한다.

20) 단어의 가운데를 절단하는 방법으로 Wom*n이면 woman, women이 모두 처리 된다.

으로 이때에는 本文의 單語에서 索引語로 連結해 주는 폭넓은 類似語 統制가 필요하다. 그밖에 辭典을 參照하거나 關聯語들을 묶어 주는 다양한 分類方法을 사용할 수 있다. 自動索引에서의 語義 分析 技法은 앞으로도 많은 硏究가 필요하고 이러한 分析方法을 위해서 言語의 組織과 作用을 더 잘 理解하기 위한 言語學 硏究가 더욱 필요하다.

Ⅳ. 自動索引시스템의 評價

自動索引시스템은 컴퓨터를 이용한 온라인 檢索시스템이 계속 증가함에 따라 그 이용이 늘고 있고 手作業索引에서의 索引作業의 不一致의 問題가 解決될 뿐만 아니라 사람이 索引作業하는 것보다 抄錄이나 本文의 一部를 檢索 目的으로 機械可讀型으로 變換시켜 索引作業에 사용함으로써 더욱 經濟的이다. 그러나 自動索引의 경우 表題나 抄錄 또는 文獻의 本文에 나오는 單語만을 索引語로 採擇하게 되므로 만일 사람이 한번 本文을 읽기만 해도 알 수 있는 명백한 主題라 할지라도 本文에 그 用語가 나타나지 않거나, 題目에서만 索引語를 골라낼 때 그 題目이 文獻의 內容을 나타내지 못할 경우 索引語 採擇에 問題가 發生한다.

스파크(J.Sparck)[21]는 自動索引시스템의 評價를 두 가지 方法으로 나누고 있는데 手作業索引시스템과 比較해 보는 것과 또 다른 自動索引시스템과 比較해 보는 것이 그것이다. 또 다른 評價方法은 巨視的 評價와 微視的 評價로 나누는 것이다. 巨視的 評價(Macro Evaluation)란 自動索引시스템의 全 性能을 評價하는 것으로 手作業索引시스템과의 比較時에는 手作業索引시스템의 디소러스와 自動索引시스템에서의 자동키워드 抽出方法과 自動分類를 比較하고 微視的 評價(Micro Evaluation)는 特殊한 變數만 比較하는 것으로 手作業索引시스템과의 比較時에는 手作業의 키워드 抽出方法과 自動索引시스템의 技法을 比較할 수 있겠다. 自動索引시스템의 評價는 再現率과 正確率이라는 檢索 性能面에서 언급이 되어야 한다. 그러면 實際 比較實驗에서 나타난 結果를 살펴 보고자 한다. 自動索引生産物로 順列表題索引(Permuted Title Index)인 KWIC, KWOC, 그리고 表題語處理시스템, 自然語處理시스템(Full Text Processing System) 등이

21) *Ibid.*

있다. 自然語處理시스템도 실제 많이 利用되고는 있으나 많은 사람들이 費用도 덜 들고 더 빠르며 폭 넓은 範圍를 다루고 있다는 점에서 自動索引을 더 選好한다.

評價硏究를 크게 세 그룹으로 나누어 볼 수가 있는데, 첫째, 表題語 硏究로 主題書名에서 自動으로 抽出된 索引項目을 比較하는 것으로 醫學[22], 化學[23], 法學[24]같은 여러 주제 분야에서 행해졌다. 둘째, 自動索引語와 手作業에 의한 用語를 比較하는 것으로 自動索引과 手作業索引의 相關係數를 測定하는데[25] 상관계수 Q는

$$Q = \frac{C}{A+M-C}$$

A : 자동추출한 용어의 수
M : 수작업으로 추출한 용어의 수
C : 공통적으로 추출된 용어의 수

이다. 이 相關係數를 비교한 結果 手作業시스템과 自動索引시스템에 의해 抽出된 用語의 60 %가 一致된다는 結論을 얻었다.[26] 그리고 세번째 硏究 形態로는手作業·自動索引法을 사용한 檢索시스템을 비교한 實驗을 들 수 있는데, 이 경우 여러 側面에서 많은 硏究가 행해지고 있다. 作動費用이나 速度같은 經營的인 基準, 質問에서부터 出力때까지의 時間, 出力物의 形態, 藏書의 範圍, 利用者의 勞力 등의 基準에서 비교할 수도 있다.

1960 년대 스완슨(R. Swanson)[27]이 처음으로 傳統的인 索引시스템과 自動文獻處理시스템을 비교했는데, 이는 傳統的인 主題名索引과 文獻의 本文에서 自動的으로 單語나 句를 抽出한 시스템을 비교한 것이다. 그 結果 自動文獻分析에 根據한 시스템의 平均檢索性能이 더 낫다고 나타났다. 후에 나온 몇개의 硏究에

---

22) C. Montgomery and D.R. Swanson, "Machine-like Indexing by People," *American Doc.*, vol.13, no.4, Oct. 1962, pp.359-366.

23) M.J. Ruhl, "Chemical Documents and Their Titles : Human Concept Indexing vs. KWIC Machine Indexing," *American Doc.*, vol.15, no.2, Apr. 1964, pp.136-141.

24) D.H. Kraft, "A Comparison of Keyword in Context (KWIC) Indexing of Titles with a Subject Heading Classification System," *American Doc.*, vol.15, no.1, 1964.1, pp.48-52.

25) G. Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*, New York : McGrow-Hill, 1983.

26) T.N. Shaw and H. Rothman, "A Experiment in Indexing by Word Choosing", *J. of Doc.*, vol.24, no.3, 1968.9, pp.159-172.

27) D.R. Swanson, "Searching Natural Language Text by Computer," *Science*, vol.132, no. 3434, 1960.10.21, pp.1099-1104.

서도 스완슨의 結果를 立證하고 있다.

살튼은 Cranfield시스템과 SMART시스템을 比較 分析했는데[28], 우선 Cleverdon의 Cranfield Ⅱ에 관한 硏究[29]를 보면 Cranfield Ⅱ 實驗은 많은 言語學的 "道具(Device)"를 측정했다. 이것은 모든 索引作業을 訓練된 索引者가 手作業으로 遂行했으나 索引規則이 매우 細分되어 있고, 몇 가지는컴퓨터의 役割을 振作시키는 方法을 사용해서 空氣力學 分野의 1,400개 文獻을 279건의 探索質問으로 다음 세 가지 側面에서 調査하였다.

① 文獻의 本文에서 選擇한 單一語(Single Term)

② 디소러스에 의해 修正된 統制語(Controlled Term)

③ 單一 槪念(Single Concepts)

이 그것이다.

調査結果 單一語를 포함한 統制하지 않은 索引語의 檢索性能이 더 좋았으며 統制語彙를 사용한 경우 잘못된 探索結果가 나왔다. 다시 말해 간단한 索引 節次에 의한 探索 結果가 더 效果的이라는 것이다. 이러한 結果가 나온 것은 單一語가 복잡한 過程을 거쳐서 統制한 語彙보다 덜 效果的일 것 같으나 自動으로 索引하기는 더 쉽고 單一語가 보통의 利用者 水準에서 特定性 水準이 정확하기만 하다면 手作業索引法보다 自動索引法이 더 效果的이고 費用도 低廉하기 때문이다.

Cranfield에 대한 實驗結果는 SMART시스템의 硏究結果로 더욱 立證되었는데[30] SMART 시스템은 IBM 7094, 370을 사용하는 實驗的인 自動文獻檢索시스템으로 同意語 辭典, 階層構造, 링크 등의 여러가지 다양한 自動分類方法과 索引方法을 이용한다. 그리고 質問式과 文獻間의 類似係數(Similarity Coefficient)를 계산해서 적은 순서대로 놓아 이용자가 원하는 수만큼 검색해 준다. 이 연구에서도 컴퓨터工學, 空氣力學, 情報學 등의 分野에서 이용되는 文獻 集團을 調査한 結果, 單一語가 가장 낮고 어떤 복잡한 分析道具도 期待한 것보다 效果가 적었다고 나타났다. 그리고 SMART와 手作業에 의한 Cranfield의 性能

---

28) G. Salton and M.E.Lesk, "Computer Evaluation of Indexing and Text Processing," *J. of ACM*, vol.25, no.1, Jan. 1968, pp.8-36.

29) C.W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings*, vol.19, no.6, June 1967, pp.174-194.

30) G.Salton and M.E. Lesk, *op. cit.*

評價結果 完全自動索引시스템이 傳統的인 索引方法보다 그다지 뒤떨어지지 않는다는 結果가 나왔다. Medlars와 SMART를 비교한 Salton의 硏究結果는 다음과 같다.[31]

분 석 방 법	재 현 율	정 확 률
Medlars(통제색인)	0.3117	0.6110
SMART(빈도가중치 사용)	0.2622	0.4901
SMART(Discrimination 사전 사용)	0.3223	0.6106

여기서 自動索引시스템의 結果가 좋지 않은 理由는 內容分析節次가 너무 간략했기 때문이고 手作業過程에서의 失敗는 잘못된 索引語 選定과 組合의 잘못, 또 探索質疑가 너무 網羅的이였거나, 너무 特定的이었기 때문이라고 볼 수 있다. 그러나 이 調査結果에서도 傳統的인 統制語彙를 사용한 手作業索引보다 간단한 自動分析法이 더 나은 結果를 갖는다는 것을 알 수 있다.

Van der Meulen[32]은 手作業主題索引인 INSPEC과 自動索引方法을 사용하는 DIRECT를 比較 評價했다.

INSPEC은 英國電子工業會의 한 분과로 科學·技術關係文獻을 分類·抄錄·索引하고 매월 索引 抄錄誌를 發刊하고 있으며 SDI업무까지도 하고 있다. 이용자는 키워드와 카테고리번호를 불리언(Boolean)論理를 이용하여 檢索할 수있다. DIRECT는 SMART시스템과 유사한 自動檢索시스템으로 자동으로 文獻本文에서 索引語를 選定하며 類似語 등을 시스템 辭典에 포함하고, 各 槪念에는 單語의 關聯語에 따라, 發生頻度에 따라, 位置에 따라 加重値를 附與하는 시스템이다. 불리언 논리도 사용하나 이 實驗에서는 사용하지 않았다. 이 實驗에서는 동일한 데이터 베이스를 사용해서 探索質問 두가지를 주어본 結果 手動시스템인 INSPEC이 DIRECT보다 正確率과 再現率이 각각 20%씩 높았다. 이러한 探索性能에 影響을 미친 것을 分析해 본 結果 文獻의 索引語作成에 있어서는

31) G. Salton, "A New Comparision between Conventional Indexing (MEDRARS) and Automatic Text Processing (SMART), *JASIS*, vol.23, no.2, Mar/Apr. 1972, pp.75-84.

32) Van der Meulen and P.J.F.C.Janssen, "Automatic versus Manual Indexing," *Inf. Proc. and Mgt.*, vol.13, 1977, pp.13-21.

自動이나 手動시스템이 비슷한 結果를 보였다. 따라서 이러한 性能 結果가 나타난 理由는 探索質問 形成에서 緣由한다고 볼 수 있다. 즉 불리언 操作子를 使用했는가 안했는가에 따라 探索性能이 달라졌으며 自動索引시스템도 探索質問作成時에 불리언 조작자를 사용할 경우 手作業시스템만큼의 性能을 期待할 수 있다.

Ⅴ. 結　論

이상에서 살펴본 것처럼 自動索引시스템의 檢索性能이 手作業에 의한 索引시스템의 性能보다 더 낫다고 斷定하기는 어렵다. 그러나 앞으로 爆發的인 情報量과 온라인 情報檢索시스템에 따른 데이터 베이스가 開發되고 發展해 감에 따라 컴퓨터를 이용한 自動索引法이 費用의 減少와 手作業의 短點인 일관된 索引作業 등의 利點으로 더 發展하리라는 期待는 할 수 있다. 그러므로 앞으로도 自動索引시스템의 檢索效率을 높일 수 있는 自動索引方法이 더욱 硏究開發되어야 할 것이다.

〈參 考 文 獻〉

1. 정 영미, "색인이론과 실제," 「연세논총」, vol.17, 1980.11, pp.21-35.
2. Artandi, Susan. "Machine Indexing : Linguistic Implications, " *JASIS*, vol.27, no.6, July/Aug. 1976, pp.235-239.
3. Borko, Harold. *Indexing Concepts and Methods*, New York : Academic Press, Inc., 1978.
4. Chomsky, N. " Three Models for the Description of Language ", *IEEE Transaction on Information Theory*, vol.2, no.1, 1956.
5. Cleverdon, C. W. " The Cranfield Tests on Index Language Devices", *Aslib Proceedings*, vol.19, no.6, June 1967, pp.174-194.
6. Cooper, W.S. " Is Interindexing Consistency a Hobglollin ? ", *American Doc.*, vol.20, no.3, July 1969, pp.268-278.
7. Damerou, F.J. " An Experiment in Automatic Indexing, " *American Doc.*, vol.16, no.4, Oct.1965, pp.283-289.
8. Doyle, L. B. " Indexing and Abstracting by Association", *American Doc.*, vol.13, no.4, Oct. 1962, pp.378-390.

9. Kraft, D. H. "A Comparison of Keyword in Cotext (KWIC) Indexing of Titles with a Subject Heading Classification System", *American Doc.*, vol.15, no.1, Jan. 1964, pp.48-52.
10. Luhn, H.P. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", *IBM J. of Research and Development*, vol.1, no.4, Oct. 1957, pp.309-317.
11. McColvin, L. R. "The Purpose of Indexing," *The Indexer*, vol.1, no. 2, pp.31-35.
12. Montgomery, C. and D. R. Swanson. "Machine-like Indexing by People," *American Doc.*, vol.13, no.4, Oct. 1962, pp.359-366.
13. Ruhl, M.J. "Chemical Documents and Their Titles : Human Concept Indexing vs. KWIC Machine Indexing", *American Doc.*, vol.15, no.2, Apr. 1964, pp.136-141.
14. Salton, G., *Dynamic Information and Library Processing*, Englewood Cliffs : Prentice-Hall, Inc., 1975
15. _____, "A New Comparison between Conventional Indexing (MEDLARS) and Automatic Text Processing (SMART)", *JASIS*, vol.23, no.2, Mar/Apr. 1972, pp.75-84.
16. _____ and C.S. Yang., "On the Specification of Term Values in Automatic Indexing," *J. of Doc.*, vol.29, no.4, Dec. 1973, pp.351-372.
17. _____ and M. E. Lesk, "Computer Evaluation of Indexing and Text Processing", *J. of ACM*, vol.25, no.1, Jan. 1968, pp.8-36.
18. _____ and Michael J. McGill, *Introduction to Mordern Information Retrieval*, New York : McGrow-Hill, 1983.
19. Shaw, T. N. and H. Rothman, "An Experiment in Indexing by Word Choosing", *J. of Doc*, vol.24, no.3, Sep. 1968, pp.159-172.
20. Sparck, Jones, "Automatic Indexing", *J. of Doc.*, vol.30, no.4, Dec. 1974, pp.393-432.
21. _____, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *J. of Doc.*, vol.28, no.1, 1972, pp.11-21.
22. Stile, H. E., "The Association Fator in Information Retrieval", *J. of ACM*, vol.8, no.2, Apr. 1961, pp.271-279.
23. Swanson, D. R. "Searching Natural Language Text by Computer", *Science*, vol.132, no.3434, 1960.10.21, pp.1099-1104.
24. Van der Meulen and P. J. F. C. Janssen, Automatic versus Manual Indexig", *Inf. Pro. and Mgt.*, vol.13, 1977, pp.13-21.