統計的 表現法의 適用實驗

金 益 喆·盧 準 植
(雙龍洋灰中央研究所 技術情報室)

〈차 례〉

Ⅰ. 序 論

産業, 技術의 급속한 發展으로 말미암아 情報의 發生量과 對象主題技術分野가 增加되고 있다.[1] 이로 인하여 필요로 하는 文獻을 어떻게 찾는가, 다시 말해서, 適合率과 再現率을 어떻게 높이는가 하는 것이 큰 問題로 대두되고 있다.

情報檢索의 自動化와 더불어 情報의 適合率과 再現率의 향상을 위한 여러 가지 試圖가 있었다. 그중 주목할만한 것은 일본과학기술정보센터(JICST) 에서 試圖中인 ASSIST(Automatic System for Selection and Instruction of Search Terms)와 같은 單語間의 相關性에 의해 旣作成된 검색텀(term)의 組合에 의한 Ready-Made-Profile 形態의 試圖,[2),3)] MRI(Mitsubishi Research Institute)의 USTAR 같은 지식(Knowle dge) 베이스의 시스템,[4)] 自動序列法(Ranking)과 같은 加重値에 의한것,[5)] 統計的 表現法 등이 있다.[6)]

이 報告書는 上記한 方法中 統計的 表現法에 관하여 쌍용양회중앙연구소의

STIS (Ssangyong Technical Information System) 데이터베이스를 對象으로 하여 그 適用可能性을 檢討하여 본 것이다.

Ⅱ. SITS 데이터베이스

STIS [7] 데이터베이스는 쌍용양회 중앙연구소에서 1976년부터 自體製作해온 서지데이터베이스로서 그 내역은 〈表 1〉과 같다.

또한 이 System의 特徵은 한글 全文檢查(Full-Text Searching)가 가능한 COHIRES (Conversational Hangul Information Retrieval System) 패키지의 開發로 社內 情報 및 한글情報의 電算化가 가능케 된 것이다(83년 11월에 COHIRES가 開發 完了되어 현재 약 4000건의 데이터를 蓄積하고 運用 및 試驗中임).

〈表 1〉 **STIS Data Base 내역**

총 Data 수	56,000건 (2만건/년 갱신)
주제범위	시멘트, 콘크리트 및 신요업(Ceramics)
검색어	통제어 및 자연어 (Thesaurus는 TEST 사용)
형 태	서지적 (Bibliographic)
수록년도	1950년-현재
검색 program	STAIRS(영문 검색용) COHIRES(한글 검색용)
COMPUTER	IBM 3083 (Remote Terminal 서울-대전간 연결)

Ⅲ. 統計的 表現法

1. 概 要

이 方法은 網羅的인 檢索을 必要로 하지않고, 적당한 수의 必要 文獻을 얻기하는 경우에 適用한다. 다시말해 再現率보다 適合率에 力點을 둔것이다.

기존의 불리안 논리(Boolean logic)에 의한 檢索 프로파일의 作成이 아니라,

檢索語에 주어진 特性値를 基準値와 對比하여, 選定된 特性値까지의 論理合을 檢索하는 것으로 다음의 事項을 前題로 한다.

① 檢索對象인 데이터베이스는 檢索語로서 自然語를 포함하고 있다. 다시 말해 統制語와 自然語로 構成되어 있는 경우이다.

② 말의 有用度는 使用頻度와 反比例 한다.[8)]

③ 이러한 여러 用語의 論理合에 의한 檢索에 比較的 높은 適合率이 얻어지는 것이 가능하다면, 檢索用語의 選擇과 프로파일의 作成도 컴퓨터에 의해 自動化가 가능해 질 것이다.

2. 對象 데이터베이스

① 檢索對象은 當社의 데이터베이스인 STIS를 STAIRS로서 檢索 했음.

② 對象 데이터는 1950～1983년분의 시멘트, 콘크리트 및 신요업(New Ceramics)관계의 데이터 56,000건임.

③ 단 5번째 테마(〈表 2〉 參照)는 CA(Chemical Abstracts) 데이터에 국한시켜 검색 했음.

3. 檢索遂行

다음과 같은 過程으로 進行하였음.

제 1 단계 : STIS 데이터베이스의 特性에 맞는 5개의 테마를 選定했음(〈表 2〉 參照).

제 2 단계 : 각 테마를 具體的으로 記述하고, 問題의 側面을 分析하고, 對象方法, 現象 등에 관해 具體的인 單語를 列擧 羅列하여 記錄한다. 텀의 갯수는 몇 개라도 상관 없으나 본 檢索遂行에서는 6～10개 정도이었다.

제 3 단계 : 前段階에서 抽出된 單語의 出現頻度를 檢索하여 確認한다.

제 4 단계 : 出現頻度에 의해 劃數가 적은 單語를 K_1, 그 다음을 K_2 식으로 K_n까지의 特性値를 附加한다. 다시 말하면 가장 頻度가 높은 單語가 K_n이 된다.

제 5 단계 : 이렇게 주어진 K_n을 〈表 3〉과 같이 整理한다. 다음 〈表 4〉, 〈表 5〉와 같이 出力累計特性値인 f_i 및 q_i를 計算한다. 여기서 f_i는 $K_1 + \cdots + K_i$까지의 累計値, q_i는 $f_n / f_n \times 100$으로 계산되는 寄與値이다.

〈表 2〉 수행 Theme 및 검색 Term

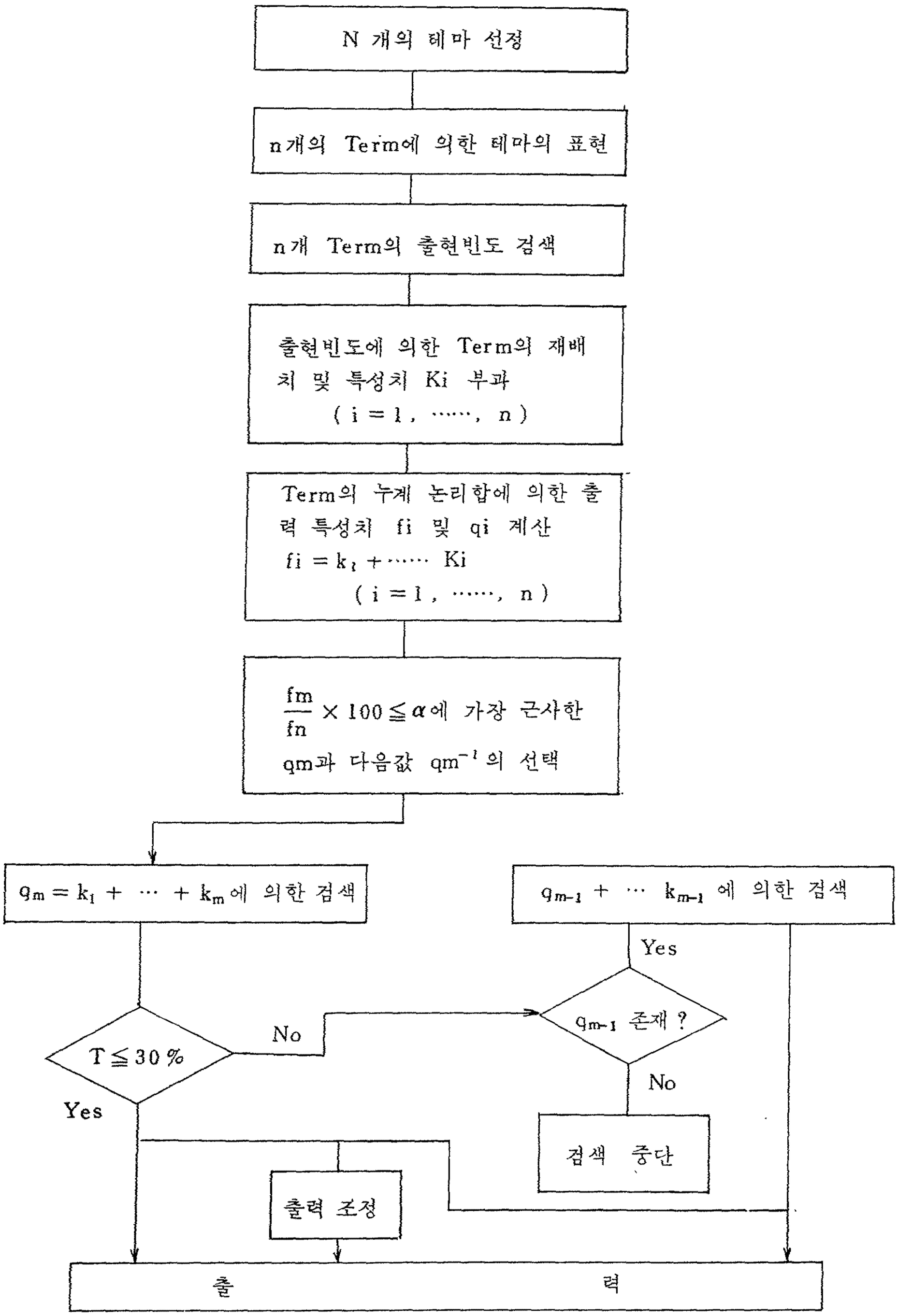
테마내용	검색 Term Ki	테마내용	검색 Term
1. 알칼리 골재 반응에 대한 문헌	K_1 : Pop out K_2 : Alkali silica K_3 : Alkali silica K_4 : Alkali aggregate : reaction K_5 : Aggregate reaction K_6 : Crack		K_5 : Heat exchange K_6 : Energy daving K_7 : Energy conservation K_8 : Cooler K_9 : Utilization K_{10} : Gases
2. 인공사 제조를 위한 원석 종류 및 설비	K_1 : Stone type K_2 : Manufactured aggregate K_3 : Pit sand K_4 : Sand stone K_5 : Manufactured sand K_6 : Artificial aggregate K_7 : Artificial sand K_8 : Crushed stone K_9 : Manufacturing	4. TiO_2가 시멘트 물성에 미치는 영향	K_1 : Minor component K_2 : Burnability K_3 : TiO_2 K_4 : Grindability K_5 : Physical property K_6 : Cracking K_7 : Titanium oxide K_8 : Effect K_9 : Strength
3. Cooler 배기의 회수 이용방법	K_1 : Waste air K_2 : Dust collection K_3 : Cooling rate K_4 : Waste gas	5. MHD 발전기용 전극재료	K_1 : Magnetohydrodynamics K_2 : Electroconductive K_3 : Generator K_4 : Mechanical strength K_5 : MHD K_6 : Electrode K_7 : Material

제 6 단계 : 미리 選擇한 基準値 α에 가장 近似한 값을 지닌 q_m과 그 아래의 값 q_{m-1}을 선택한다. 여기서 주의할 점은 근사값 q_m이 α보다 커서는

〈表 3〉 4단계 수행 결과 (I)

예제	K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8	K_9	K_{10}
1	0	15	153	176	221	2024				
2	0	0	3	5	8	13	20	223	833	
3	4	55	65	208	276	276	328	587	764	1287
4	95	196	198	354	556	2135	2224	13503	22129	
5	4	25	101	171	176	563	9304			

〈圖 1〉 통계적 표현법 수행 flow chart



〈表 4〉 4단계 수행 결과 (Ⅱ)

예제	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
1	0	15	168	344	565	589				
2	0	0	3	8	16	29	49	272	1105	
3	4	59	124	332	608	884	1212	1799	2563	3850
4	95	291	189	843	1399	3534	5758	19261	41930	
5	4	29	130	301	477	1040	10344			

〈表 5〉 q_m의 계산 및 선정

예제	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}
1	0	0.6	6.5	13.3	21.8	100				
2	0	0	0.3	0.7	1.4	2.6	4.4	24.6	100	
3	0.1	1.5	3.22	8.6	15.8	23	31.5	47	67	100
4	0.2	0.7	1.2	2	3.4	8.5	4	46	100	
5	0.04	0.3	1.3	2.9	4.6	10.5	100			

안된다는 것이다. 여기서 α는 3%로 選定 使用했는데 이는 우리가 目的으로 하는 適合率 30% 정도를 滿足시킬 수 있으리라고 假定한 값이다.[6] α를 增加시키면 再現率은 增加하고 適合率이 減少한다.

제 7 단계 : 〈表 6〉과 같이 각 테마별로 q_m에 따라 檢索을 하고, 檢索된 文獻을 20건이 초과할 때는 임의로 20건만을 브라우징(Browsing)하여 適合率을 判斷하고 미만일 때는 全文獻을 出力하여 判讀하고 適合率을 把握했다. 한편, 20건만 出力할 경우 데이터의 分散으로 인한 誤差發生의 경우도 고려하여 全體檢索文獻을 出力하여 全體의 適合率도 調査해 보았다(〈表 7〉 參照). 금번 수행에서 4번과 5번 테마가 1차 수행에서 適合率이 目標보다 낮아 재차 q_{m-1}로 再檢索을 했다. 그 結果는 〈表 8〉과 같다.

〈表 6〉 선택된 K_m과 q_m

예 제	q_m	q_{m-1}	K_n (검 색 Term)
1	q_2	없 음	K_2
2	q_6	q_5	$K_3 + K_4 + K_5 + K_6$
3	q_2	q_1	$K_1 + K_2$
4	q_4	q_3	$K_1 + K_2 + K_3 + K_4$
5	q_4	q_3	$K_1 + K_2 + K_3 + K_4$

〈表 7〉 1차검색에 의한 적합률

예 제	총 출 력 수	출력조정적합률	총 적 합 률	비 고
1	13	11(85 %)		
2	25	8(40 %)	9(36 %)	
3	43	7(35 %)	15(34 %)	
4	394	3(15 %)	7(18 %)	q_{m-1} 검증
5	78	1(5 %)	15(19 %)	

〈表 8〉 2차검 에 의한 적합률

예 제	총 출 력 수	출력조정적합률	총 적 합 률	비 고
4	211	3(15 %)	5(2.2 %)	기준미달
5	25	6(30 %)	9(36 %)	

Ⅳ. 結 論

1) 本 方法의 遂行結果 4개의 테마가 目標値인 30 % 보다 큰 값으로서 統計的 表現法의 客觀性을 認定할만 했다.

2) 이 方法을 DIALOG 등의 海外 데이터 뱅크 利用時와 같이 主題分析인 容易치 않고, 데이터베이스의 構造를 파악키 힘든 경우에도 適用可能할 것 같다.

3) 이 방법의 가장 큰 弱點은 檢索用語의 選定이 主題에 대한 高度의 理解가 없이는 不可能하다는 것이다.

結論的으로 USTAR 種類의 지식 베이스 시스템만이 情報檢索의 難題를 解決하리라 생각된다.

〈參 考 文 獻〉

1. 笹本光雄, 「Chemical Abstracts の使い方」, 1979.
2. 佐藤 雅之외 4人, "キーワード自動選擇 システムの 開發," 1981, 「18回情報科學技術研究集會發表論文集」, pp. 51-60.
3. 佐藤 雅之외 2인, "キーワード 自動選擇 システム(ASSIST)を 用いた檢索 評價實驗", 「19回情報科學技術研究集會發表論文集」, 1982. pp. 111-120.

4. MRI 計劃 システム部, "知的な 情報檢索システムを 目指して— USTAR システムの 研究開發," *MRI New Letter*, no.24, 1983, p.7.

5. 海老沼幸夫, "自動 ランキング法の 信賴性の 基礎",「ドワメンテーション 研究」, vol. 32, no. 4, 1982, pp. 173-181.

6. 海老沼幸夫, "オンライン 簡易自動文獻 探索法",「情報管理」, vol. 26, no. 9, 1983, pp. 726-735.

7. 雙龍 研究所,「STIS 정보 서비스 이용 안내」, 1983.

8. Sparck Jones, K., "Statistical Interpretation of Term Specificity and Its application in Retrieval", J. *Doc*, vol. 28, no. 1, 1972, pp. 11-21.

9. 相川進, "オンライン 情報檢索入門"「ドクメンテーション 研究」, vol. 31, no.3, 1981, pp. 105-112.

절약하는 아빠마음 밝아지는 엄마얼굴