自然語 데이터 베이스에 대한 인덱스 키의 分析

石 川 徹 也
金 廣 明 譯
(KIET 電算室)

〈차 례〉

Ⅰ. 目 的

情報檢索 시스템을 運用함에 있어서 데이터베이스 管理者는 限定된 電算機 資源을 效果的으로 活用한다는 意味에서 對象이 되는 데이터의 特性에 대하여 充分히 認識한 후에 運營管理(Maintenance)를 할 必要가 있게 된다.[1]

특히 데이터의 增加分 및 데이터 길이가 可變的인 書誌데이터를 對象으로 하는 경우에는 데이터 파일(Data File) 및 인덱스 파일(Index File)에 대한 파일 容量의 最適値를 事前에 推算하는 것이 하나의 課題가 된다.

따라서 여기서는 圖書館情報大學에서 利用되고 있는 情報檢索 시스템(TOOL-IR on ORION)의 對象 데이터 베이스인 LCMARC 및 JAPAN-MARC중

1) 穗鷹良介, 「データベースの論理設計」(情報處理雙書 6), 情報處理學會, 1981, p.108.
酒井博敬, "データベース技術應用の過去・現在・未來," 「情報處理」, vol.23, no.10, 1982, pp.893～897.

에서 LC-MARC를 對象으로 인덱스 키의 增加特性을 分析하고 데이터의 增加分에 대한 인덱스 파일容量의 增加係數를 算出하고자 시도하여 보았다.

이 實驗的 分析은 自然語處理 시스템에 있어서 必須的인 機械可讀辭典의 엔트리語數 및 辭典파일容量의 推算을 위한 基礎的인 데이터를 얻을 수 있으리라는 意圖에서 이루어진 것이다.

Ⅱ. 데이터 베이스化를 위한 考察事項

1. 書誌 데이터의 特性

書誌 데이터의 特性을 整理해보면 아래와 같다.

① 書誌 데이터는 時系列的으로 發生한다. 아울러 그 데이터 內容은 항상 固有하다. 따라서 過去의 데이터도 항상 固有한 存在가 된다.

② 書誌 데이터의 發生量은 推定이 곤란하다. 또 各 主題分野에 따라 增加傾向이 서로 다르다.

③ 書誌 데이터는 상당히 많은 데이터 項目으로 構成되어 있다.

④ 데이터 項目의 記述形式을 크게 나누면 다음과 같다. 즉, 하나의 레코드 內에 數値 데이터, 自然語 데이터, 統制語 데이터, 識別 데이터와 같이 記述形式이 다른 데이터가 混在한다.

⑤ 위에서 말한 항목들에서 엔트리 데이터로부터 차례대로 인덱스 키를 抽出한 경우, 특히 自然語 데이터가 對象으로 되기 때문에 인덱스 키의 길이가 一定하지 않게 된다.

⑥ 마찬가지로 엔트리 데이터에서 차례대로 인덱스 키를 抽出한 경우 識別 데이터를 除外한 數値 데이터, 自然語 데이터, 統制語 데이터에서 抽出된 인덱스 키는 모든 엔트리 데이터 內에서 重複하여 發生하게 된다. 이 結果 固有한 인덱스 키 數의 增加量이 一定하지 않게 된다.

2. 인덱스 파일 內의 인덱스 키의 更新

書誌 데이터의 데이터 베이스化에 있어서 運營管理 데이터가 아닌 한 항상

소스 데이터 파일(source data file)에 대하여 更新을 要하게 된다. 마찬가지로 書誌 데이터 內에서 抽出한 인덱스 키도 인덱스 파일에 대하여 更新을 要하게 된다. 단 인덱스 키가 같을 때에는 인덱스 키 項을 更新하는 것이 아니라 포인터(pointer)項 만을 更新하게 된다. 즉, 서로 다른 인덱스 키만을 更新하게 된다. 여기서 인덱스 키項의 數를 N, 모든 포인트數를 P*라 하고 하나의 데이터 內에서 抽出한 다른 인덱스 키 數 n을 更新하는 경우 n을 更新할 때마다 인덱스 키項에 新規로 登錄된 다른 인덱스 키의 數를 x라고 하면, x가 更新될 確率係數는 다음과 같이 나타낼 수가 있다.

$$P(x) = \frac{\left(\frac{P^*}{x}\right) \left(\frac{N(1-p)}{n-x}\right)}{\left(\frac{N}{n}\right)} \quad \cdots\cdots ①$$

(-(1 − p) : (x − n)의 確率)

여기서 n에 비해 N이 크게 되면(實驗値는 6,016,801키), 하나의 데이터의 更新에 의해 P는 거의 變化하지 않는다(實驗値는 1～2포인터). 따라서 X의 更新分布는 아래 ②式의 二項分布가 된다고 할 수 있을 것이다.

$$\lim_{n\to\infty} P(x) = \left(\frac{n}{x}\right) p^x (1-p)^{n-x} \quad \cdots\cdots ②$$

이 점 때문에 書誌 데이터의 데이터 베이스化에 있어서 인덱스 키의 出現現象이 問題가 된다.

自然語의 出現現象에 대해서는 Zipf의 텍스트 데이터(text data)에 대한 分布相關法則이 알려져 있다.

3. 데이터 베이스化 要件

書誌 데이터의 데이터 베이스化를 實現하기 위한 實務的인 要件을 整理하면 아래와 같다.

① 1의 ① 特性에 대하여, 時系列的인 것 이외에도 定期的으로 데이터 파일(data file)에 대하여 運營管理를 하여야 할 必要가 있다.

② 1의 ② 特性에 대하여, 有限한 蓄積資源으로 保存·維持한다는 점에서 데이터 베이스化의 保有期限을 限定하고, 同時에 데이터 파일에 대하여 相當量의 豫備領域을 事前에 確保하여 놓아야 할 必要가 있게 된다.

③ 1의 ③ 特性에 대하여, 各各의 데이터 項目의 意味內容을 데이터 베이스 內에 적용할 때 事前에 評價하여 보아야 할 必要가 있게 된다.

④ 1의 ④ 特性에 대하여, 인덱스 키의 記述形式이 다르기 때문에 各 記述形式마다 檢索 커맨드 신텍스를 設計할 準備가 必要하게 된다.

⑤ 1의 ⑤ 特性에 대하여, 인덱스 파일의 인덱스 키項을 可變長化 할 必要가 있다.

⑥ 1의 ⑥ 特性에 대하여 인덱스 파일의 인덱스 키의 엔트리項 및 액세스 포인터의 엔트리項을 事前에 推定하여 確保해 놓을 必要가 있다.

위에서 말한 데이터 베이스化 要件中에 시스템의 設計次元에서 對處하여야 할 問題와 對象데이터에 대하여 事前에 分析하고, 그 特性을 利用하여야 하는 問題로 구분할 수가 있다.

前者의 問題에 대해서는 圖書館情報大學의 情報檢索 시스템 TOOL－IR on ORION[2),3)]에서 實現하고 있으므로 여기서는 省略하기로 한다. 後者의 問題에 대한 對處案에 대하여 간단히 整理하면 아래와 같다.

① 위에서 말한 3의 ① 및 ②의 問題에 대하여, 對象 데이터의 增加特性을 時系列的으로 分析하고, 적어도 항상 데이터 保有期限을 기초로 데이터 數를 豫測分析하고 增加豫測値를 利用하여 데이터 파일 容量의 確保에 利用하고 있다.[4)]

② 上記한 3의 ③ 및 ④의 問題에 대하여, 對象 데이터 베이스의 特徵을 기초로 데이터 項目別로 인덱스 키의 抽出方法을 檢討하여 그 方法을 근거로 檢索 커맨드 신텍스를 設計하였다.[5)]

---

2) HITACマニュアル, "プログラム・プロダクト ── 情報檢索システム ORION 利用の手引 ──".

3) 山本毅雄, " TOOL－IR/ORIONの使い方 ── OS7 TOOL ──IRとの差を中心に ──" 「東大大型計算機センターニュース」, vol.12, no.10, 1980, pp. 58～68.

4) 石川徹也, " LC-MARCⅡデータ・ベースの言語別書誌データ件數と年間收錄レコード件數の豫測 "「圖書館短期大學紀要」, no.13, 1977, pp.25～33.

5) TOOL－IR/ORION LC ──MARC, JAPAN ──MARC檢索マニュアル／檢索實例集, 1983年 (圖書館情報大學內部資料).

③ 上記한 3의 ⑤의 問題에 대하여, 인덱스 키 길이項의 最大 길이에 대한 決定을 위해서는 단어 길이에 대한 分析을 할 必要가 있다. 識別 데이터, 統制語 데이터 길이에 대해서는 對象 데이터 베이스 안내책자를 參考로 하여 判定할 수 있으나, 數値 데이터 및 특히 自然語 데이터에 대해서는 단어 길이와 對應하는 發生頻度分布와 함께 分析할 必要가 있다.[6]

④ 上記한 3의 ⑥의 問題에 대하여, 인덱스 키 內의, 識別 데이터에서의 인덱스 키를 除外한 다른 모든 인덱스 키를 對象으로 인덱스 키의 增加特性을 데이터의 增加値와 對比하여 分析하고, ①과 같이 데이터의 保有期限을 기초로 인덱스 키 數의 增加豫測分析을 하여 增加豫測値를 利用하여 인덱스 파일 容量의 確保에 적용할 必要가 있다.

따라서 情報檢索 시스템 TOOL-IR on ORION을 근거로 利用되고 있는 LC-MARC데이터 베이스를 對象으로 上記한 ④의 問題에 대한 分析結果에 대하여 報告하고 자 한다. 또한, 同機能 시스템에 있어서 우시마루(牛丸) 등은 *Chemical Abstracts Condensates*를 對象으로 分析하고 있다.[7]

Ⅲ. 인덱스 키의 增加特性分析

1. 데이터 베이스 파일容量의 實驗値

LC-MARC 데이터 베이스를 ORION 시스템으로 데이터 베이스化 한 結果, 파일 容量의 實驗値는 〈表 1〉과 같다.[8]

實驗値에 나타난 것처럼 인덱스 파일의 容量은 모든 데이터 베이스 파일 容量의 約 50%를 차지하며 소스데이터 파일에 대해서 대략 1 : 1로 對應하고

---

6) 이 問題에 대해서, 에로서는 아래의 조사·분석 보고가 있다. 더우기, 당 시스템 인덱스 키의 길이에 관한 조사결과는 아래의 표에 나타나 있다.
·田中康仁, "專門用語の解析と應用"「計量國語學」, vol.12, no.8, 1981, pp. 367～376.

7) 牛丸守, 山崎昶, 山田博, 藤原鎭男, 山本毅雄, "TOOL——IRシステムⅡ——データベース內容の統計的解析——",「昭和 49 年度情報處理學會 第 15 回大會豫稿集」.

8) HITACマニュアル, "プログラム·プロダクト——VOS情報檢 インデシクス·キー長の實測値索システム ORION建設と運用マニュアル——".

〈表 1〉 DB파일 容量의 實驗値

DB名	데이터 件數	파일 名稱	파일 容量 (KB)	比 率(%)
LC-MARC (Books - All)	678,382 '78/4～'82/1 6卷1號～9卷 44號	Table File ①	152	0.01
		Head File ②	554,168	51.90
		Index File ③	511,057	47.87
		Range File ④	2,279	0.22
		合 計	1,067,656	100.00

① 데이터 베이스에 관한 情報를 저장하고 있는 파일.
② Source Data를 저장하고 있는 파일.
③ 인덱스 키와 Source Data파일 內의 Source Data에의 포인터를 저장하고 있는 파일.
④ Range 데이터 키를 저장하고 있는 파일.

있음을 알 수 있다.

이는 書誌 데이터를 對象으로 하는 情報檢索 시스템에 있어서 인덱스 파일의 重要性 및 파일 運營管理의 重要性을 말해줌을 알 수 있다.

2. 인덱스 키 數의 實驗値

소스데이터에서 데이터 베이스化 한 인덱스의 出力인 인덱스 파일 트랜잭션 (Index file transaction 以下 ITR이라 稱함)을 時系列的으로 16分한 實驗値를 〈表 2〉에, 또 인덱스 키의 抽出對象 데이터項目의 一覽을 〈表 3〉에 나타내었다.

3. 自然語表記 인덱스 키의 實驗値

LC-MARC로부터 인덱스 키를 抽出하기 위한 데이터 項目을 〈表 3〉의 키 種別로 나타낸 것과 같이 原則的으로 소스 데이터에 대해서 個別的으로 固有하게 부여된 데이터 項目(U記號로 表示)과 書誌 데이터로서의 意味內容을 가진 데이터 項目(N記號로 表示)으로 나눌 수가 있다. 後者의 데이터 項目은 著者名, 團體名, 會議名 등을 포함하여 原則的으로 自然語表記 데이터 項目이라고 간주할 수가 있다. 그리고, 또 前者에서 固有하게 부여된 데이터 項目中에 定型表記 데이터 項目(C記號로 表示)이 存在한다.

그리하여 〈表 2〉의 인덱스 키의 實驗値를 위에서 말한 세 가지 系列로 나눈 實驗値는 〈表 4〉와 같다.

〈表 2〉 인덱스 키 數의 實驗値

ITR 分割 No.	데이터 數		Total 인덱스 키 數		異種 인덱스 키 數	
	ITR單位의 데이터 數	累積 데이터 件數	ITR 單位의 인덱스 키數	累積 인덱스 키 數	ITR單位의 인덱스 키數	累積 인덱스 키 數
1	45,060	45,060	3,278,935	3,278,935	727,114	727,114
2	46,105	91,165	3,399,163	6,678,098	742,257	1,252,503
3	44,580	135,745	3,281,084	9,959,182	729,040	1,704,918
4	45,486	181,231	3,363,330	13,322,512	748,824	2,131,296
5	42,858	224,089	3,155,821	16,478,333	704,365	2,501,586
6	37,671	261,760	2,841,327	19,319,660	665,601	2,839,689
7	42,633	304,393	3,315,794	22,635,454	720,453	3,193,793
8	40,624	345,017	3,049,855	25,685,309	709,483	3,534,408
9	43,729	388,746	3,358,241	29,043,550	770,144	3,897,354
10	42,131	430,877	3,233,804	32,277,354	725,375	4,219,218
11	40,683	471,560	3,167,650	35,445,004	720,100	4,528,220
12	43,070	514,630	3,421,776	38,866,780	770,534	4,858,562
13	40,996	555,626	3,339,433	42,206,213	747,708	5,170,099
14	43,166	598,792	3,433,437	45,639,650	731,366	5,476,906
15	44,427	643,219	3,548,879	49,188,529	750,715	5,780,258
16 (合計)	35,163	678,382	2,815,901	52,004,430	620,920	6,016,801

〈表 3〉　　LC-MARC의 인덱스 키 抽出對象 데이터 項目一覽

Tag	對象 데이터	키
No.	項　　目	種別
001	LC Control 番號	U
008 /07	刊年 1	U,C
/11	刊年 2	U,C
/15	出版國 코드	U,C
/35	言語 코드	U,C
020	ISBN	U
022	ISSN	U
050	LC 請求記號	U
060	NLM Call 番號	U
070	NAL Call 番號	U
082	DC 分類番號	U
100	基本標目・個人名	N
110	基本標目・團體名	N
111	基本標目·團體名·會議名	N
130	基本標目・統一書名	N
210	略書名	N
212	異書名	N

Tag	對象 데이터	키
No.	項　　目	種別
222	要約書名	N
240	統一書名	N
241	로마字化書名	N
242	翻譯書名	N
243	統一書名	N
245	書名表示	N
246	別書名	N
247	前書名 or 異書名	N
400	叢書表示：個人名書名	N
410	叢書表示：團體名書名	N
411	叢書表示：會議名 叢書名	N
440	叢書表示：叢書名	N
490	叢書表示：Tracing 無	N
600	件名副出：個人名	N
610	件名副出：團體名	N
611	件名副出：會議名	N
630	件名副出：統一書名	N

Tag	對象 데이터	키
No.	項　　目	種別
650	件名副出：一般書名	N
651	件名副出：地名	N
700	副出記入：個人名	N
710	副出記入：團體名	N
711	副出記入：會議名	N
730	副出記入：統一書名	N
740	別形式으로 副出한 叢書名	N
800	叢書副出：個人名	N
810	叢書副出：團體名 叢書名	N
811	叢書副出：會議名 叢書名	N
830	叢書副出：統一書名	N
840	叢書副出：叢書名	N
870	別形：個人名	N
871	別形：團體名	N
872	別形：會議名	N
873	別形：統一書名 標目	N

U…固有 데이터 項目.

C…定型表記 데이터 項目.

N…自然語表記 데이터 項目.

〈表 4〉 인덱스 키 數의 內譯

ITR 分割No.	累積데이터 件數	異種인덱스키數 固有키數	定型表記 키數	自然語表記 키數	合計
1	45,060	112,418	75	614,621	727,114
2	91,165	217,282	90	1,035,131	1,252,503
3	135,745	311,302	99	1,393,517	1,704,918
4	181,231	404,638	107	1,726,551	2,131,296
5	224,089	489,400	118	2,012,068	2,501,586
6	261,760	565,682	142	2,273,865	2,839,689
7	304,393	649,336	147	2,544,310	3,193,793
8	345,017	727,630	156	2,806,622	3,534,408
9	388,746	811,325	169	3,085,860	3,897,354
10	430,877	892,741	178	3,326,299	4,219,218
11	471,560	968,449	182	3,559,589	4,528,220
12	514,630	1,048,929	188	3,809,445	4,858,562
13	555,626	1,128,425	192	4,041,482	5,170,099
14	598,792	1,218,582	197	4,258,127	5,476,906
15	643,219	1,307,637	201	4,472,420	5,780,258
16 (合計)	678,382	1,378,292	201	4,638,308	6,016,801

① 固有 키…固有 데이터 項目에서 抽出한 인덱스 키.
② 定型表記 키…定型表記 데이터 項目에서 抽出한 인덱스 키.
③ 自然語表記 키…自然語表記 데이터 項目에서 抽出한 인덱스 키.
①과 ②는 一部가 重複된다.

4. 인덱스 키의 特性

① 圖書館情報大學 시스템의 하나의 書誌 데이터當 平均 인덱스 키의 抽出件數는 〈表 5〉에서 볼 수 있는 것처럼 約 76.66件으로 상당히 많다. 이는 〈表 3〉의 LC-MARC의 抽出對象 데이터 項目一覽에서 볼수 있는 것처럼 對象 데이터 베이스의 데이터 엔트리 項目의 거의 모두에 대해서 인덱스 키를 抽出하고 있는 것에 起因한다고 생각된다. 이는 本 시스템의 設計趣旨에 의한 바가 크다. 그것은 데이터 베이스를 단순히 一般의 檢索要求를 위하여 데이터 베이스化하는 것이 아니라 本 大學의 研究·教育現場을 支援하기 위하여 利用하는 것도 目的으로 하고 있기 때

〈表 5〉 인덱스 키의 特性

데이터件數 ①	Total 抽出 ② 인덱스 키 數	異種抽出 ③ 인덱스 키 數	重複率 ④	1 데이터 당 ⑤ 平均抽出 인덱스 키 數
678,382	52,004,430	6,016,801	8.64	76.55 + 0.5

④ = ③ / ②
⑤ = ② / ①

문이다.[9] 具體的으로는 圖書館業務에서 主로 圖書의 選別 · 目錄業務에서의 데이터 提供 및 敎育現場에서의, 主로 分類 · 目錄의 敎授學習에의 데이터 利用에도 充分히 使用可能할 수 있도록 하는 것을 目的으로 하고 있다.

② 抽出된 인덱스 키에 대해서 現在로는 〈表 5〉에서 처럼 單純히 平均해서 9 포인터를 維持하고 있는 것을 알 수 있다.

5. 自然語表記 인덱스 키의 增加現象

앞에서 말한 〈表 4〉의 인덱스 키 數의 實驗値의 內譯을 그래프化 하면 〈圖 1〉과 같이 된다(단 定型表記 키는 誤差의 範圍에서 省略). 또 增加分에 대하여 그래프化 하면 〈圖 2〉와 같이 된다.

① 自然語表記 인덱스 키의 增加現象은 〈圖 2〉에서 처럼 指數函數特性을 나타냄을 알 수 있다. 따라서 自然語 表記 인덱스 키의 增加特性을 最小自乘法에 의하여 實驗式을 구하면 다음과 같다.[10]

$$y_1 = 197.6\ x^{0.75} \qquad \cdots\cdots ①$$

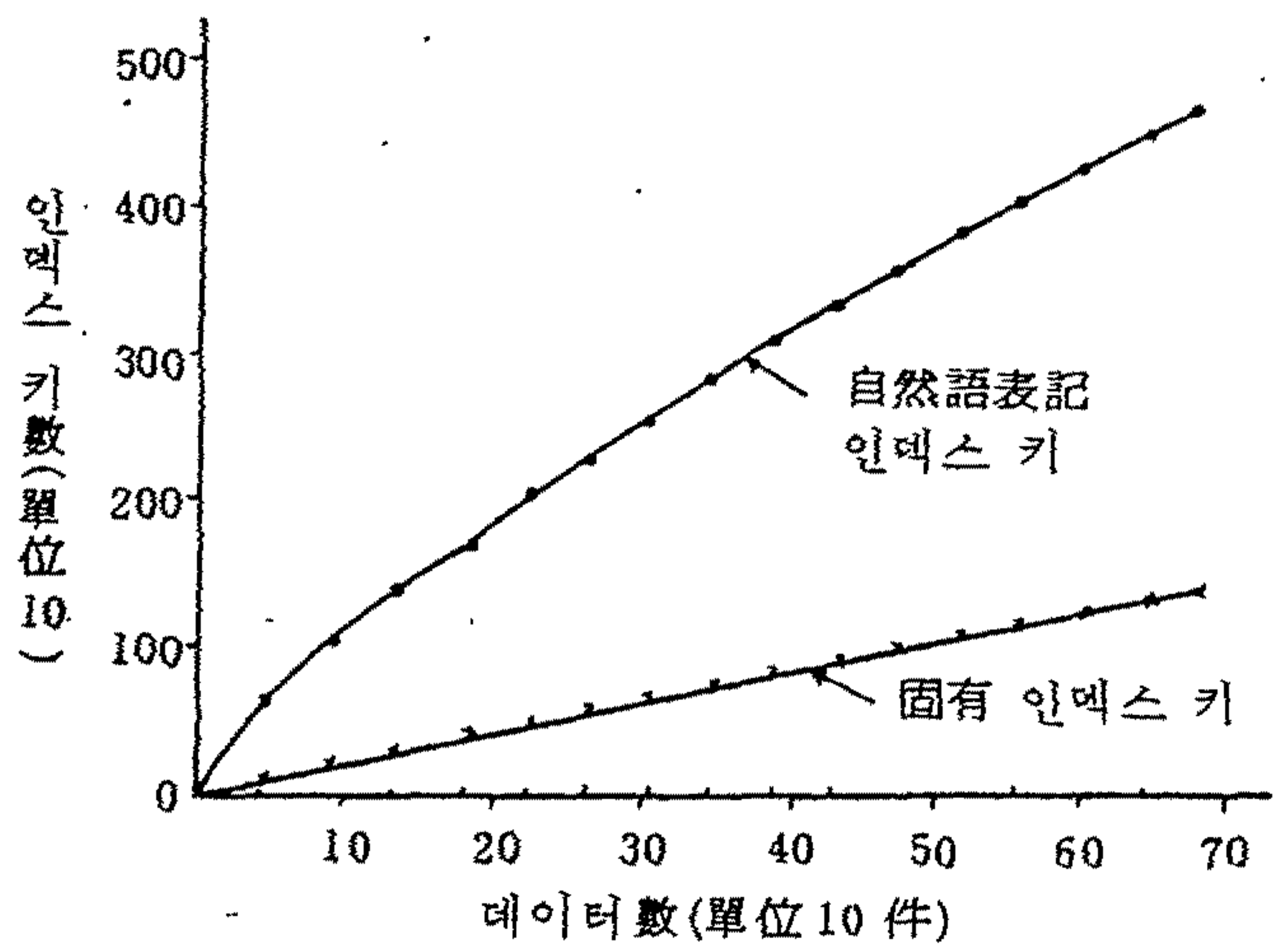
y_1 : 自然語表記 인덱스 키 數

x : 데이터 件數

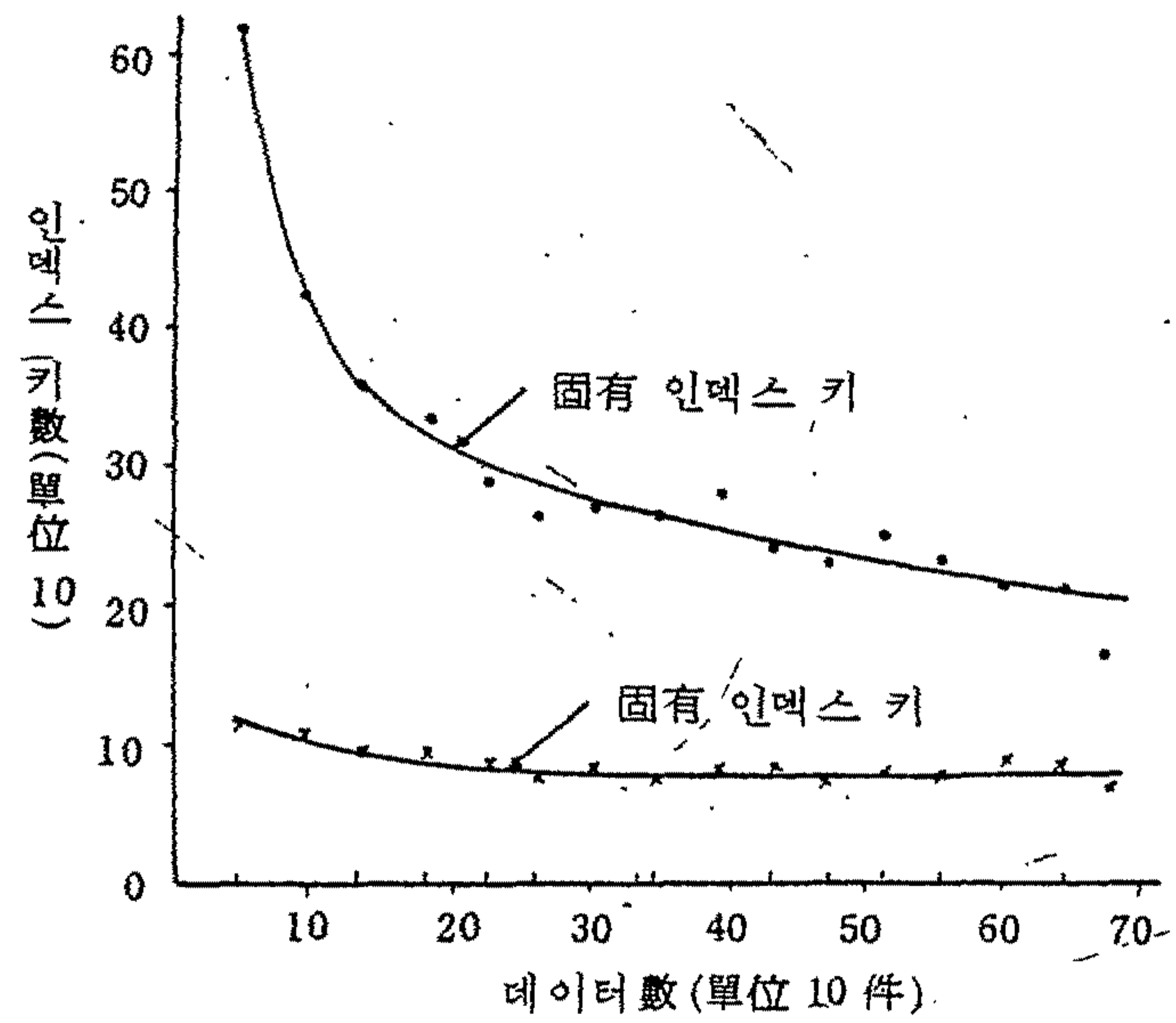
9) 山本毅雄, "圖書館情報大學の情報システム", 「大學圖書館硏究」, vol.19, 1981, pp.11 ~ 17.
——, "圖書館情報大學の情報システムについて", 「情報管理」, vol.25, no.11, 1983, pp.985 ~ 991.

10) 灘保夫 編, 「確率統計現象 Ⅰ, Ⅱ(岩波講座 基礎工學 3)」, 岩波書店, 1973, Ⅰ : p.114, p.231, Ⅱ : p.251.
三重野博司, 「情報處理のための數學」, 電氣學會, 1971, p.245.

〈圖 1〉 自然語表記 인덱스 키의 增加傾向



〈圖 2〉 인덱스 키의 增加特性



①式을 5 %의 危險率로서 分散分析에 의해 檢定하면 意味가 있다고 判定이 된다.

② 固有 키의 增加現象은 〈圖 2〉에서 처럼 一次函數로서 나타나진다. 이것은 固有 키의 內容, 즉 하나의 데이터에 대해 原則的으로 1件 出現한다는 점에서도 推測될 수 있다. 따라서 앞에서 말한 ①과 같이 固有 키의 增加特性을 實驗式으로 나타내면 아래와 같이 된다.

$$y_2 = 2.06\ x \quad \cdots\cdots ②$$

y_2 : 유니크 키 數

x : 데이터 件數

②式을 5%의 危險率로서 分散分析에 의해 檢定하면 意味가 있는 것으로 判定된다.

③ 以上과 같은 점에서 異種 인덱스 키의 增加特性의 實驗式은 ①式과 ②式의 合으로써 表現될 수 있다.

즉, $y = 197.6\ x^{0.75} + 2.06\ x$

y : 異種 인덱스 키 數

x : 데이터 件數

Ⅳ. 結論 및 考察

本 報告의 目的의 하나는 大容量으로서 可變長 데이터項目이 많은 書誌 데이터 베이스를 데이터 베이스化하여 감에 있어서 特히 인덱스 파일의 容量確保를 將來에 대비하여 어느 程度로 고려하는 것이 좋을 것인가를 LC-MARC를 對象으로 하여 實驗한 것과, 또 하나는 이 實驗結果를 토대로 自然語處理 시스템에서 要求되는 機械可讀辭典製作을 위한 參考 데이터를 얻을 수 있었다. 結論的으로 自然語의 出現指數係數 0.75를 얻었다.

實驗式의 指數가 異種 自然語表記 인덱스 키數 4,638,308에 대하여 (〈表 4〉 참조) 0.75로서 큰 것임을 감안할 때 역시 異種 自然語表記 인덱스 키는 增加하리라는 것을 豫測할 수 있다.

이에 대해서 히프스(H.S.Heaps)는 LC-MARC의 主題標目에 대하여 增加指數 0.66을 提示하고 있다.[11] 히프스의 값과 비교하여 筆者가 提示한 實驗値쪽이 큰 것은 本 시스템에 있어서는 著者名, 團體名 등의 데이터項目과 이 項目으로부터 人爲的으로 作成한 短縮 키 및 프리픽스 키(Prefix Key)를 포함하여 實驗한 것에 起因한다고 생각된다.

---

11) Heaps, H.S., "Information Retrieval — Computational and Theoretical Aspects"—, Academic Press, 1978, p.344.

〈表 6〉 인덱스 키 길이의 實驗値

키길이 (B)	異種抽出 키 數	比 率 (%)	合計 키 길이 (B)
1	37	0.00	37
2	872	0.05	1,744
3	7,143	0.45	21,429
4	19,935	1.26	79,740
5	34,736	2.19	173,680
6	60,419	3.80	362,514
7	123,766	7.80	866,362
8	285,292	17.97	2,282,336
9	301,682	19.00	2,715,138
10	182,902	11.52	1,829,020
11	445,853	28.08	4,904,383
12	19,862	1.25	238,344
13	101,379	6.39	1,317,927
14	3,771	0.24	52,794
15	0	0	0
合 計	1,587,649 ①	100.00	14,845,448 ②

· 對象 데이터… '78 年 1 年分.
· 本 시스템에서는 最大 인덱스 키 길이를 15 B로 하고 있다.
· 인덱스 키 길이의 平均…①/②≒9.35 B.
· A,AN, THE를 Stop Word로 脫落시켰다. 또 人爲的으로 生成시킨 短縮 키 및 Prefix키 등을 포함한다. 따라서 모든 自然語의 단어길이라고는 限定할 수 없다.