韓國經營科學會誌
第10券 第2號
1985年 10月

—研究報告—

2차 探索費用函数를 갖는 데이터베이스의 再構成 時期 결정에 관한 研究

A Study on Deciding Reorganization Points for Data Bases with Quadratic Search Cost Function

姜 錫 昊*
金 永 杰*

Abstract

Reorganization is an essential part of data base maintenance work and the reasonable reorganization points can be determined from the trade-off between reorganization cost and performance degradation. There has been many reorganization models so far, but none of these models have assumed nonlinear search cost function. This paper presents the extensions of two existing linear reorganization models for the case where the search cost function is quadratic. The higher performance of these extended models was shown in quadratic search cost function case.

Ⅰ. 序 論

대규모 데이터 베이스 시스팀(Date Base System)의 경우, 挿入(insertion), 削除(deletion)등의 修正(update) 作業들이 일정기간 계속됨에 따라 別途의 再構成作業(Reorganization)을 수행해 주지 않으면 데이터 베이스의 探索費用이 급격히 증가하게 된다.

데이터 베이스를 再構成함으로써 探索費用의 급격한 증가를 막을 수 있으나 재구성 작업에는 적지 않은 비용이 隨伴되므로 그 時期는 任意로 결정될 수 없다.

確定的(Deterministic) 探索費用函數를 假定한 지금까지의 研究는 주로 探索費用의 增加가 시간變數의 一次式으로 표시되는 線形再構成模型에 대해 이루어져 왔다.

즉, 시간이 經過함에 따라 화일(File)의 크기가 線形的으로 증가하고 따라서 원하는 레코드(Record)에 대한 探索時期도 선형적으로 증가한다는 가정이 前提되었었다.

그러나, 實際의 데이터 베이스 시스팀에서는 非線形 探索費用函數도 존재하고 있음이 발표된 바 있으며 [4], [7], 이러한 境遇에 從來의 선형 재구성 모형들을 적용한다면 실제 탐색비용함수와의 適合缺如에 따른 費用損失이 豫想된다.

따라서, 本 論文은 두가지 대표적인 線形 再構成模型들을 擴張하여 探索費用 함수가 二次인 경우에도 적용가능한 두 가지 再構成模型들을 提示하고자 한다.

* 서울大學校 産業工學科

Ⅱ. 研究背景

데이터 베이스의 再構成 時期 결정에 관한 研究는 재구성을 필요케하는 여러 要因 중에서도, 원하는 레코드에 대한 探索費用(또는 使用者 應答時間)이 증가하여 시스템 成果가 低下되는 경우에 대하여 주로 이루어져 왔다.

그 이유는 내부 데이터의 構造變化라든가 物理的 接近樣式(access method)의 변경으로 인한 재구성 요인 등은 일정한 변화 양상을 갖지 않으므로 豫測하기 어려운 반면, 탐색비용의 증가는 어느 데이터 베이스에서나 共通的으로 發生하고 있으며 또 그 變化樣相이 시간에 따라 일정한 형태를 가지는 것이 대부분이어서 計量化 및 分析하기가 비교적 容易하기 때문이다.

시간의 경과에 따른 탐색비용의 증가는 그림-1에서 보는바와 같이 自體 화일 크기의 增加에 의한 부분과 데이터 베이스內 貯藏 構造의 腐敗에 따른 부분으로 區分되며, 재구성 작업에 의해 抑制할 수 있는 費用增加는 後者의 경우라고 하겠다. [8]

탐색 비용의 증가가 任意的(ramdom)인 경우에 대해선 Mendelson[10], Yechiali [10], Heyman[12] 및 Winslow[11] 등이 연구결과를 발표한 바 있으며, 탐색비용의 증가가 確定的(deterministic)인 경우에 대해서는 1973년 Shneiderman의 "Optimun Data Base Reorganization Points" [1], 를 嚆矢로 Yao & Theory [5], 및 Tuel [6] 등이 연구한 바 있다.

이들의 모형이나 알고리듬은 모두 探索費用이 시간에 따라 線形的 增加를 한다고 假定하고 있다.

本 論文은 데이터 베이스 壽命週期, T가 주어진 경우와 그렇지 않은 경우의 대표적인 두 확정적 모형들인 Shneiderman의 「最適再構成 週期 模型(Optimun Fixed-Interval Reorganization Model)」과 Yao의 「動的再構成 模型(Dynamic Reorganization Model)」을 확장하여 탐색비용이 二次的(Quadratically) 으로 증가할 경우의 재구성 시기를 결정하고자 한다.

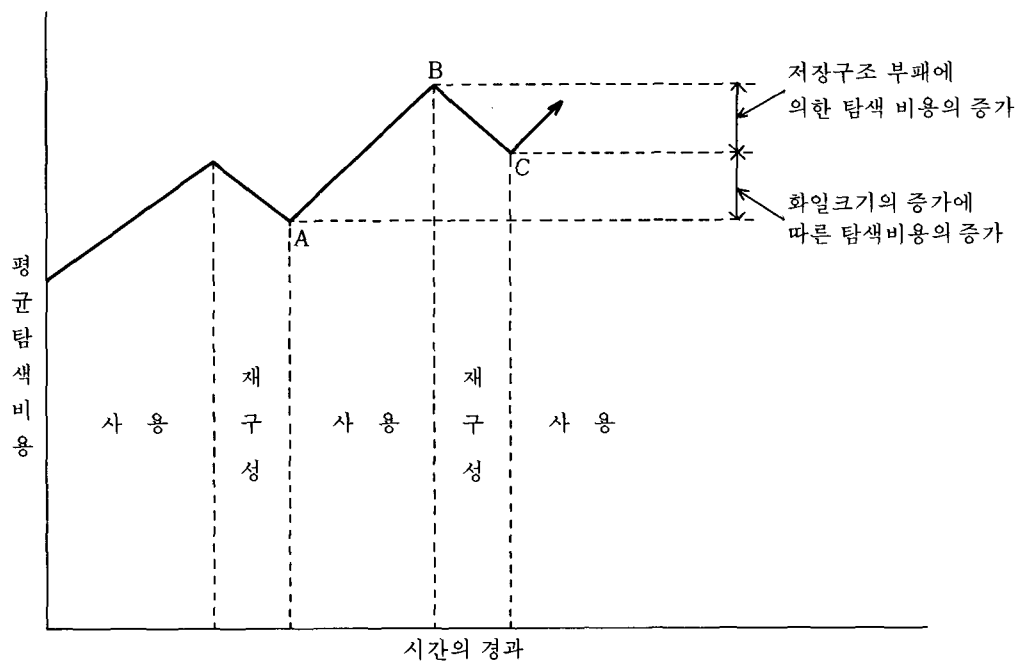


그림-1 탐색비용 및 재구성

Ⅲ. 研究方法

여기서는 探索費用이 二次的으로 정의되는 데이터 베이스의 再構成 時期 決定模型으로써 壽命週期 T 가 주어질 경우에는 FIXED_QUAD 模型을, 수명주기가 주어지지 않을 경우에는 DYNAMIC_QUAD 모형을 제시한다.

Ⅲ-1 FIXED_QUAD 模型

本 模型은 Ⅱ장에서 소개된 바 있는 '最適再構成週期 모형' [1] 의 諸般 假定들을 기초로 하며, 이중 탐색비용, S_t에 관한 가정만이 $S_t = S_0 + \theta_t$에서 $S = S_0 + \theta' t^2 + \theta'' t$로 確張되었다.

따라서, 다음 식들이 가정된다.

i) $\gamma_{t+1}(t) = \gamma_t(t) + \theta_3 t^2 + \theta_4 t$, $\gamma_1(t) = S_0 + \theta_1 t^2 + \theta_2 t$

ii) $\delta_{t+1}(t) = \delta_t(t) + \theta_3 t^2 + \theta_4 t$, $\delta_1(t) = S_0 + \theta_3 t^2 + \theta_4 t$

iii) $R_{t+1}(t) = R_t(t) + \mu t$, $R_1(t) = R_0 + \mu t$

iv) $\theta_1 \geqq \theta_3 \geqq 0$, $\theta_2 \geqq \theta_4 \geqq 0$

(단 $\theta_1 = \theta_3$면 $\theta_2 > \theta_4$)

여기서, iv) 식은 다음 條件에서 派生된다.

$$d(t) = \gamma(t) - \delta(t) = (\theta_1 - \theta_3) t^2 + (\theta_2 - \theta_4) t > 0 \quad \forall t > 0$$

이러한 확장된 가정들 하에서 화일에 대한 探索費用은 그림-2의 二次 曲線下의 면적이다. 總費用은 탐색비용과 재구성 비용과의 合이고 탐색비용 함수는 N개의 區間에서 각기 일정하다.

最少化시키고자 하는 비용, C(T)은 剩餘 探索費用(그림-2의 빗금친 부분)과 每 週期($t, 2t, 3t, \cdots$)마다의 再構成 費用을 합한 값이다.

$$C(T) = \sum_{i=1}^{N} \left[\int_0^t (\gamma_i(t') - \delta_i(t'))\, dt' + R_i(t) \right]$$

$$= N\left[\frac{(\theta_1 - \theta_3)}{3} t^3 + \frac{(\theta_2 - \theta_4)}{2} t^2 \right] + NR_0 + \frac{N(N+1)}{2} \mu t$$

$\left(N = \frac{T}{t} \text{ 대입} \right)$

$$= T\left[\frac{(\theta_1 - \theta_3)}{3} t^2 + \frac{(\theta_2 - \theta_4)}{2} t \right] + \frac{1}{t} (TR_0 + \frac{\mu}{2} T^2) + \frac{1}{2} \mu T$$

한편,

$$\frac{dC(T)}{dt} = T\left[\frac{2(\theta_1 - \theta_3)}{3} t + \frac{(\theta_2 - \theta_4)}{2} \right] - \frac{1}{t^2} (TR_0 + \frac{\mu}{2} T^2)$$

$$\frac{d^2 C(T)}{dt^2} = \frac{2T(\theta_1 - \theta_3)}{3} + \frac{2}{t^3} (TR_0 + \frac{\mu}{2} T^2) > 0 \quad \forall t > 0$$

따라서, C(T)는 t가 양수일 때 t에 대한 凹面(c

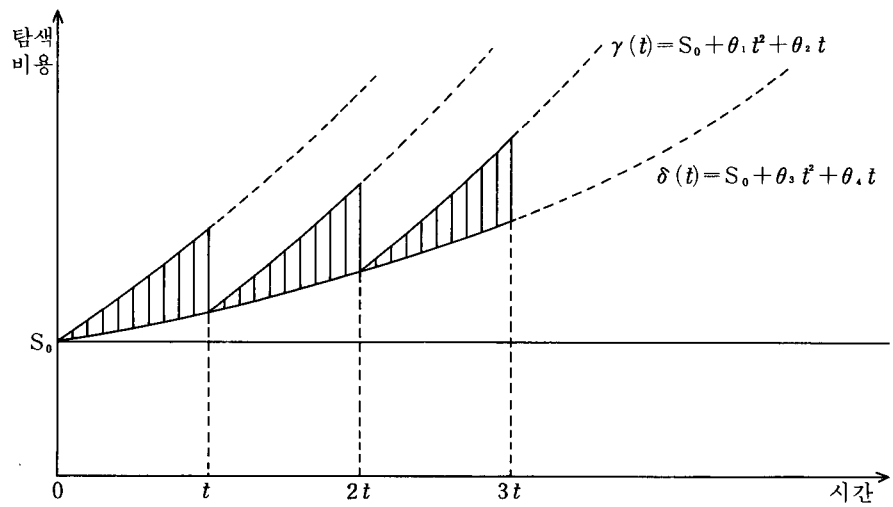


그림-2 비선형 탐색비용 함수

onvex) 函數이다. 그러므로, 구하고자 하는 最適再構成 週期, t_{FQ} 는 다음과 같다.

i) $\theta_1=\theta_3$, $\theta_2>\theta_4$ 일 때

$$t_{FQ}=\left[\frac{2R_0+\mu T}{\theta_2-\theta_4}\right]^{\frac{1}{2}}$$

ii) $\theta_1>\theta_3$, $\theta_2\geqq\theta_4$ 일 때

t_{FQ}는 0과 壽命週期, T 사이의 값으로 다음식을 만족시키는 t가 된다.

$$F(t)=\frac{2(\theta_1-\theta_2)}{3}t^3+\frac{(\theta_2-\theta_4)}{2}t^2-(R_0+\frac{\mu T}{2})$$
$$=0, \quad 0<t<T$$

여기서, 三次式 $F(t)$는

$$t=0,\ \frac{-(\theta_2-\theta_4)}{2(\theta_1-\theta_3)}\ (<0)$$

의 두 점에서 極값을 갖게 되며, 그림－3과 같은 형태를 갖는다.

따라서, 방정식 $F(t)=0$은 $t>0$인 범위에서 하나의 근, t^* 을 가짐을 알 수 있다.

이 根은 1차원 탐색 알고리듬을 통하여 쉽게 찾아낼 수 있으며, 그 값이 수명주기, T, 보다 클 境遇에는 棄却되고(재구성 非수행), 작을 경우에는 구하고자 하는 再構成週期, t_{FQ}로 採擇된다.

Ⅲ－2 DYNAMIC_QUAD 模型

本 模型은 Ⅱ장에서 소개된 바 있는 '動的再構成模型' [5] 및 동적 재구성 基準(Dynamic Reorganzation Criterion)을 토대로 하며, 이 중 탐색 비용 S_n에 대한 假定만이 $S_n=S_0+\theta n$에서 $S_n=\theta' nS_0+\theta' n^2+\theta'' n$으로 確張되었다.

따라서, 다음 式들이 새로 가정된다.

i) $S_n=S_0+\theta_1 n^2+\theta_2 n$

ii) $S_n'=S_0+\theta_3 n^2+\theta_4 n$

iii) $R_n=R_0+\mu n$

iv) $\theta_1\geqq\theta_3$, $\theta_2\geqq\theta_4$ (단 $\theta_1=\theta_2$ 면 $\theta_2>\theta_4$)

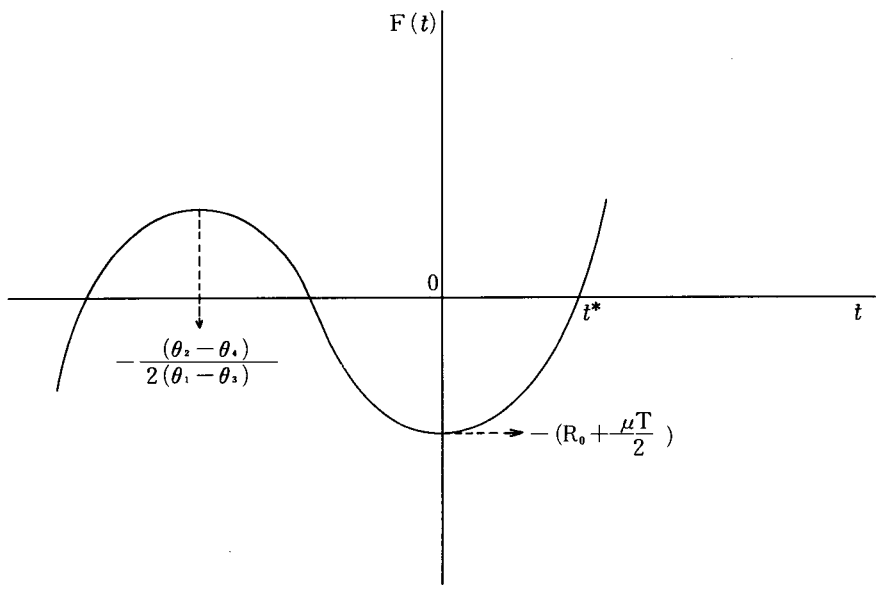
動的 再構成 基準을 따를 경우, 구하고자 하는 재구성시기, t_{DQ} 는 다음 不等式을 만족시키는 最少의 整數 n이다.

$$S_n\geqq S_n'+\frac{R_n}{n} \quad \cdots\cdots(1)$$

(1)式에 새로 가정된 S_n, S_n', R_n 등을 대입하면, (n은 t_n으로 置換)

$$S_0+\theta_1 t_n^2+\theta_2 t_n\geqq S_0+\theta_3 t_n^2+\theta_4 t_n+\frac{R_0+\mu t_n}{t_n}$$

$$D(t_n)=(\theta_1-\theta_3)t_n^3+(\theta_2-\theta_4)t_n^2-\mu t_n-R_0 \quad \cdots\cdots(2)$$



그림－3 $F(t)$ 그래프

이 구해진다.

그러므로, 구하고자 하는 動的再構成 時期, t_{DQ},는 다음과 같다.

i) $\theta_1=\theta_3$, $\theta_2>\theta_4$ 일 때

$$t_{DQ}=\{\mu+[\mu^2+4R_0(\theta_2-\theta_4)]^{\frac{1}{2}}\}/(2(\theta_2-\theta_4)) \quad \cdots\cdots\cdots\cdots(3)$$

ii) $\theta_1>\theta_3$, $\theta_2\geqq\theta_4$ 일 때

t_{DQ}는 다음 式을 만족시키는 t_n이다.

$$D(t_n)=(\theta_1-\theta_3)\,t_n^3+(\theta_2-\theta_4)\,t_n^2-\mu t_n-R_0=0 \quad \cdots\cdots\cdots\cdots(4)$$

$$\forall\ t_n>0$$

여기서, 다음 재구성 시기를 결정하기 위해서는 (2)式의 R_0에 당시의(Current) 재구성 비용, Rn을 代入하여 (3)式이나 (4)式을 새로 適用해야 한다.

$D(t_n)$은

$$t_n=\begin{cases}\dfrac{-(\theta_2-\theta_4)+\sqrt{(\theta_2-\theta_4)^2+3\mu(\theta_1-\theta_3)}}{3(\theta_1-\theta_3)}\ (=t_n'>0)\\ \dfrac{-(\theta_2-\theta_4)-\sqrt{(\theta_2-\theta_4)^2+3\mu(\theta_1-\theta_3)}}{3(\theta_1-\theta_3)}\ (=t_n''>0)\end{cases}$$

의 두점에서 極값을 갖게 되어 그림－4의 곡선 A의 같은 형태를 갖는다. 또한, 다음번 재구성 시기, t_n^{**} 도 R_0만 R_n으로 바꿔준 뒤 같은 방식으로 찾아낼 수 있다.

이때, $D(t_n)$은 그림－4의 곡선 B의 형태를 갖는다.

Ⅳ. 適用事例

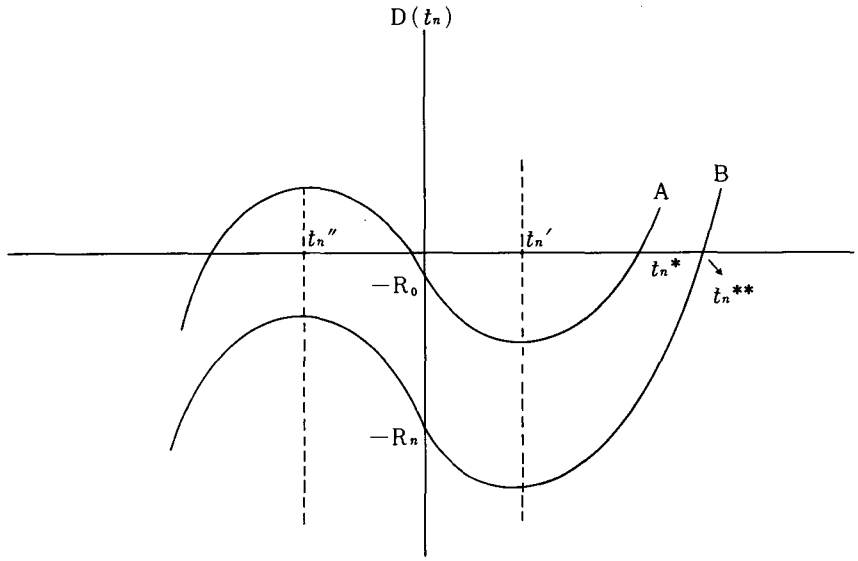
線型 探索費用 함수를 가정한 Yao등의 재구성 모형에서는 다음과 같은 費用파라미터들을 사용해 왔다. [5, 6]

$\theta_1=$ 5,460 (access seconds/day
$\theta_2=$ 3,000 (access seconds/day)/day
$R_0=$75,100 (access seconds)
$S_0=$681,000 (access seonds/day)
$\mu=$ 75,100 (access seconds/day)

여기서 비용은 每 接近(access)시의 待機時間(단위 : 초)으로 가정된다.

본 사례는 二次 탐색비용 함수를 가정하고 있으므로 두 가지 탐색비용 곡선,

$S_t=S_0+\theta_1 t^2+\theta_2 t$와 $S_t'=S_0+\theta_3 t^2+\theta_4 t$가 회귀곡선 $S_t=S_0+5460t$와 $S_t'=S_0+3300t$로 適合되는 θ_i $(i=1,2,3,4)$ 값들을 취했다.



그림－4 $D(t_n)$ 그래프

표－1은 본 사례의 탐색비용계수들로서 기간 T 별로 回歸分析을 거쳐 선택된 파라미터들이고, 괄호 안의 값들은 적합된 회귀직선($y=S_0+\beta t$)의 기울기, β이다.

표－2는 본 사례에 대한 6가지 再構成 正策들의 成果를 수명주기別로 비교해 본 結果이다.

표－2 재구성 정책들의 비교 : 사례A

T	C_{int}	정책	재 구 성 시 기	$C_{dis+reorg}$	% Save
200	199.0	NO		37.3	
		LF	83. 80, 167.61	30.1	19.3
		LD	35.74, 92.87, 169.97	30.7	17.7
		QF	91.54, 155.62	29.2	21.7
		QD	52.36, 131.29	23.8	36.2
		MO	30, 60, 90, 120, 150, 180	51.5	
400	520.7	NO		144.0	
		LF	118.22, 236.45, 354.67	80.5	44.1
		LD	35.74, 92.87, 169, 97.266.39, 381.75	91.3	36.6
		QF	139, 25, 278.49	65.5	54.5
		QD	63.99, 161.08, 286.32	64.3	55.3
		MO	30, 60, 90,……, 390	212.3	
800	1452.1	NO		608.0	
		LF	166.99, 333.97, 500.96, 667.94	166.0	72.7
		LD	35.74, 92.87, 169.97, 266.39, 381.75, 515.79, 668.35	185.7	69.5
		QF	224.30, 449.60, 628.03	155.7	74.4
		QD	116.58, 279.99, 480.29, 712.22	168.1	72.4
		MO	30, 60, 90,……, 780	797.0	
1,600	4751.7	NO		2304.0	
		LF	236.01, 472.01, 708.02, 944.03, 1180.03, 1416.04	452.0	80.4
		LD	35.74～668.35, 839.27, 1028.46, 1235.83, 1461.31	776.3	66.3
		QF	361.17, 722.33, 1083.50, 1372.43	407.4	82.3
		QD	167.45, 401.65, 688.30, 1019.88, 1391.46	390.5	83.1
		MO	30, 60, 90,……, 1590	3233.5	

표－1 사례의 θ_i 추정치

T \ θ	θ_1	θ_2	θ_3	θ_4
200	15	3,000	10	1,800
	(5, 357)		(3, 371)	
400	9	2,500	6	1,500
	(5, 328)		(3, 385)	
800	7	1,000	4	700
	(5, 399)		(3, 214)	
1,600	3.9	500	2.4	300
	(5, 402)		(3, 317)	

여기에 使用된 用語들은 다음과 같다.

T : 데이터 베이스의 壽命週期(단위 : 日)

C_{int} : 화일크기 증가에 따른 探索費用의 증가치(단위 : 10^6 access seconds)

$C_{dis+reorg}$: 재구성 비용＋저장구조 腐敗로 인한 탐색비용의 증가치(단위 : 10^6 access seconds)

% Save : 재구성에 의한 비용 感小 비율(%)

NO : 재구성 非 수행

LF : 最適 再構成 週期 모형 (線型 탐색비용 가정)

LD : 動的 재구성 모형 (선형 탐색비용 가정)

QF : FIXED_QUAD 모형

QD : DYNAMIC_QUAD 모형

MO : 每月 재구성 수행

총 비용, C(T)은 C_{int}와 $C_{dis+reorg}$의 合이며, 이들은 다음과 같이 구해진다.

$$C_{int} = S_0 T + \frac{1}{2}\theta_4 T^2 + \frac{1}{3}\theta_3 T^3$$

$$C_{dis+reorg} = \frac{1}{3}(\theta_1 - \theta_3)\sum_{i=1}^{K+1}(t_i - t_{i-1})^3 + \frac{1}{2}(\theta_2 - \theta_4)\sum_{i=1}^{K+1}(t_i - t_{i-1})^2 + \sum_{i=1}^{K}(R_0 + \mu t_i)$$

여기서, $t_0 = 0$, $t_{K+1} = T$이고 K는 재구성 횟수이다. 또한, 成果比較 基準인 % Save는 다음과 같다.

$$\%\ \text{Save} = 100 \times \frac{C_{dis+reorg}\ (\text{NO}) - C_{dis+reorg}\ (\text{정책})}{C_{dis+reorg}\ (\text{NO})}$$

V. 結　論

本 硏究에서는 從來의 데이터 베이스 再構成 模型들이 線形 探索費用函數를 가정해 온 점을 確張하여, 탐색비용함수가 二次인 경우에도 適用 가능한 2가지 재구성 모형(FIXED_QUAD, DYNAMIC_QUAD)들을 提示하였다.

또한, 적용사례를 통해, 탐색비용함수가二次인 데이터 베이스 시스팀의 境遇, FIXED_QUAD 모형은 (선형) 最適 再構成 週期 모형보다 또 DYNAMIC_QUAD 모형은 (선형) 動的 再構成 모형보다 좋은 시스팀 成果를 가짐을 보였다. 본 연구의 追後 연구 사항으로는 첫째, 탐색비용함수가 일반적인 비선형 함수인 경우의 再構成 모형 定立과 둘째, 二次인 경우의 可變週期 재구성 모형에 대한 硏究 등을 들 수 있겠다.

참　고　문　헌

1. Shneiderman, B. :"Optimum data base reorganization points." Comm. ACM 16, 6 (June 1973), 362－365.
2. Shneiderman, B.:*Data base: Improving Usability and Responsiveness.* Academic press, 1978, 151－189.
3. Das, K. S., Theory, T. J., & Yao, S. B.: "Reorganization points for file designs with nonlinear processing costs (abstract)." *Proc. Inter. Conf. Very Large Data Bases,* Sept. 1975, 516－517.
4. Yao, S. B. & Merten, A. G.: "Selection of file organization using an analytic model." ibid, 255－267.
5. Yao, S. B., Das, K. S., and Theorv, T. J.: " A dynamic database reorganization algorithm," *ACM Trans. Database Syst.* 1, 2 (June 1976), 159－174.
6. Tuel, w. E.: "Optimum Reorganization Points for Linearly Growing Files." *ACM Trans. Database Syst.* 3, 1 (March 1978), 32－40.
7. Söderlund, L.: "Concurrent Database Reorganization- Assessment of a Powerful Technique

through Modelling." *Proc. Inter. Conf. VLDB.* 1981. 499-509.

8. Sockut, G. H., Golaberg, R. P.: "Database Reorganization- Principles and Practices." *Computing Surveys, ACM,* 11, 4 (Dec. 1979), 371-395.
9. Sockut, G. H.: "A performance model for computer detabase reorganization Performed concurrently with usage." *O. R.*, 26, Sept.-Oct. 1978, 789-804.
10. Mendelson, H., Yechiali, U.: "Optimal policies for Data Base Reroganization." *O. R.*, 29 Jan.-Feb. 1981, 23-36.
11. Winslow, L. E., Lee, J. c.: "Optimal Choice of Data Restructuring Points." *Porc. VLDB,* 1980. 353-363.
12. Heyman, D. P.: "Mathematical Models of Database Degradation," *ACM Trans. Database Syst* 7. 4 (Dec. 1982), 615-631.
13. wiederhold, G.: *Database Design,* McGraw-Hill, 1981.
14. Atre, S.: *Data Base: structured Technigues for Design, Performance, and Management,* John-willey, 1980.
15. Date. C. J.: *An Introduction to Database Systems,* Vol. Ⅱ. Addison-wesley, 1983.
16. 강 석호 : 경영정보론 : 개념과 설계. 박영사, 1983