The Journal of Natural Science,
Pai Chai University, Korea
Volume 5(2):25-32, 1992

통계 데이타베이스의 효율적 관리를 위한 관계형데이타베이스 관리 시스템에의 전위시스템 설계

안 성 옥*· 김 용 호**
* 배재대학교 국제산업대학 전자계산학과
** 배재대학교 국제산업대학 전자계산학과

The Design of Front-end System to RDBMS for Effective Management of Statistical Database

Sung Ohk An*and Yong-Hoh Kim**
* *Dept. of Computer Science, Pai Chai University*
** *Dept. of Computer Science, Pai Chai University*

통계 데이타 베이스는 데이타가 단순한 통계치일 뿐만 아니라, 일반적인 통계처리에서 필요한 통계 분석을 위해 주로 사용되는 대량의 데이타 베이스를 말한다.

통계 데이타 베이스를 관리하기 위해 기존의 범용 데이타 베이스 관리 시스템을 그대로 이용하기에는 데이타 저장과 액세스의 비효율성, 사용의 편이성의 부족과 질의어 등의 부족으로 인해 사용자의 요구를 충족시키지 못해, 새로운 관리 방법의 필요성이 요구되어 왔다.

독자적 개발에 의한 새로운 소프트웨어로써 통계 데이타 베이스를 관리할 때의 실제 이용하기 어려운 현실적 제고를 고려하여, 이 논문에서는 관계형 데이타 베이스 시스템에의 전위 시스템인 SM-F 시스템을 설계하여, 이를 이용하여 통계 데이타 베이스를 관리하는 방법을 제시하였다.

이 시스템은 통계 데이타 베이스의 효율을 고려한 시멘틱 모델인 GROS 모델을 사용하며 통계분석을 지원하고 통계 요약 정보를 제공하기 위해, 메타 데이타 베이스와 요약 데이타 베이스를 저장하고 운영한다.

Statistical database(SDB) are large database primarily collected for the purpose of statistical analysis. Commerical database management systems have not been widely used for SDB because of the efficiency problem of storage and access of those systems for SDB.

In this paper, we propose SDB management method to use a front-end system to a Relatianal Database Management System (RDBMS).

We do the design of SM-F system (Stasticical database Management as Front-end system) as a front-end system to a RDBMS.

In this system, we use GROS model specially proposed for SDB, and store and manage summary database and meta database to support statistical analysis and to provide users with statistical summary infomation.

Keywords : 통계 데이타 베이스, SM-F 시스템, GROS 모델, 메타 데이타 베이스, 요약 데이타 베이스

Ⅰ. 서 론

통계 데이타 베이스는 지금까지의 범용 데이타 베이스 시스템이 지원하는 데이타 베이스와는 다른 특징을[5,6] 가지고 있어 이러한 특징을 고려하여 에그리게이션 연산을 명확히 표현하고, 추상 오브젝트 (abstract object) 들을 지원하여 저장과 액세스의 효율성과 이용의 편이성을 가져오기 위하여 독자적 시스템에 의한 통계 데이타 베이스 관리의 필요성이 대두되어 실제 통계 데이타 베이스 관리 시스템을 구현하여 통계 데이타 베이스를 관리 하고자하는 연구가 많이 있어 왔다.[1,5]

그러나, 현실적으로 만약 어떤 기관이나 연구소에서 이렇게 독자적 개발에 의한 새로운 소프트웨어로써 통계 데이타 베이스를 관리할 때, 기억장소나 액세스의 효율성을 높이고 사용자에게 사용의 편이성을 제공할 수 있으나, 한계점은 상업용 데이타 베이스 관리 시스템이 아니므로 인사관리, 급여계산 등을 포함한 그 기관의 전체업무를 총괄하기에는 제한점이 많아 또 다른 상업용 데이타 베이스 관리 시스템의 도입을 초래하므로 생기는 낭비적인 요소와 두개의 데이타 베이스 관리 시스템과 관련된, 데이타 공유 및 디자인의 균형에 어려움을 배제할 수 없다. 그리하여 보다 효율적으로 통계 데이타 베이스를 관리한다 하더라도, 독자적 데이타 베이스 관리 시스템을 이용하기 어려운 현실적 제고를 고려하여, 현재 상업용 데이타 베이스 관리 시스템 중 가장 바람직한 관계형 데이타 베이스 관리 시스템을 목적 시스템으로 한 전위 시스템인 SM-F 시스템 (Statistical database management as front-end system) 을 설계하여, 이를 이용하여 통계 데이타 베이스를 관리하는 방법을 논해 보겠다.

Ⅱ. SM-F 시스템의 구조

통계 데이타 베이스 관리를 위해 관계형 데이타 베이스 관리 시스템을 목적 시스템으로 하여 설계한 전위 시스템을 SM-F 시스템(Statistical database management as front-end system) 이라 명명 하겠다.

SM-F 시스템을 전위 시스템으로 하고 목적 데이타 베이스 관리 시스템을 관계형 데이타 베이스 관리 시스템으로 한 통계 데이타 베이스 관리를 위한 구조는 〈그림 1〉 과 같다.

통계 데이타 관리를 위하여 사용되는 메타 데이타 베이스, 원시 데이타 베이스와 요약 데이타 베이스는, 목적 시스템인 관계형 데이타 베이스 관리 시스템의 저장구조를 이용하여 저장한다. 즉, 전위 시스템인 SM-F 시스템에 의하여 생성된, 요약값들의 개념적 모델을 해석하여 내부적 모델인 릴레이션으로 변환 처리하여, 요약 데이타 베이스를 저장한다. 한편, 요약 데이타 베이스로부터의 질의 요구는 반대로, 릴레이션을 GROS 모델[20]에 의한 테이블 형태로 변형하여 출력한다. 현재 사용되는 관계형 데이타 베이스 관리 시스템에, 몇 개의 모듈로 된 소프트웨어인 SM-F 시스템을 첨가하여, 사용자에게 통계 분석 기능을 지원하고, 요약 데이타의 통계량에 관한 통계적 뷰를 제공할 수 있어, 현실적으로 쉽게 사용될 수 있다.

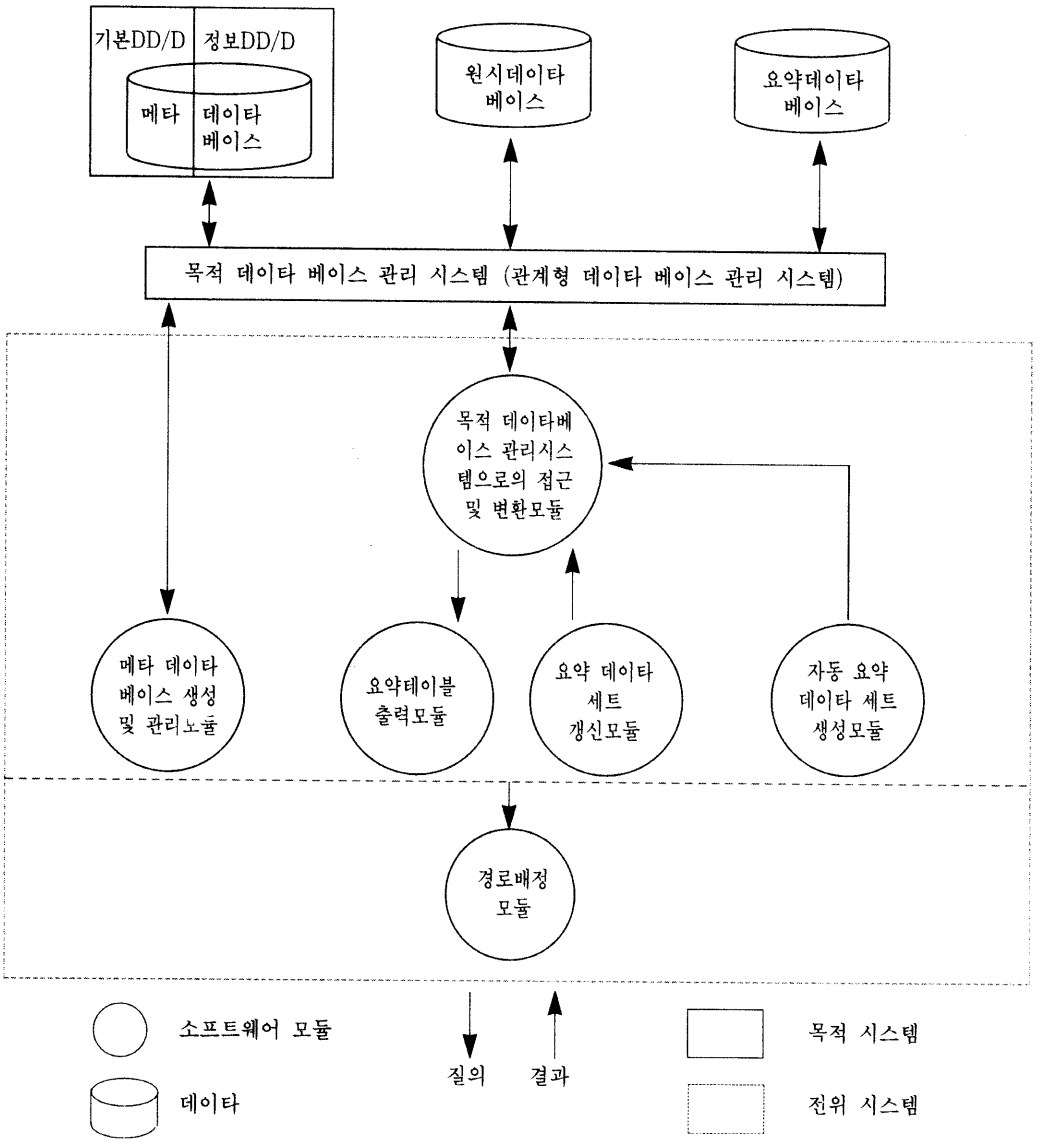
Ⅲ. 관계형 데이타 베이스로의 접근 및 변환

GROS 모델의[20] 계층구조에 의한 요약 데이타가 저장되는 개념적 스키마의 기본 형태인 AST(Automatic Summary Table)[20]는 목적 데이타 베이스 관리 시스템으로의 접근 및 변환 모듈에 의해 AST의 기본구조를 입력으로 받아 릴레이션으로 변환시켜 릴레이션의 집합을 출력으로 내보내어 관계형 데이타 베이스 시스템의 저장구조를 이용하여 저장된다. 이와같은 변환 알고리즘[12,22]은 행과 열에 대한 속성 계층구조와 중간적인 릴레이션을 저장하여야 하며 이러한 단계적 변화를 거쳐야 하므로 수행시간이 길어지며, 중간적인 데이타를 저장하여야 하는 부담을 갖는다. 또한 이렇게 릴레이션으로 저장된 데이타를 반대로 AST로 출력하기 위한 반대 과정의 알고리즘도 필요하다.

통계 데이타인 농업 조수입 데이타를 예로하여 이 과정을 도식으로 표현하면 〈그림2〉와 같다.

Ⅳ. 요약정보 처리를 위한 질의어의 확장

일반적으로 관계형 데이타 베이스 관리 시스템에 사용되는, 대표적인 질의어로는 튜플관계해석을 기초로 한 데이타 언어로서 INGRES에서 사용하는 QUEL 과 관계사상을 기초로 한

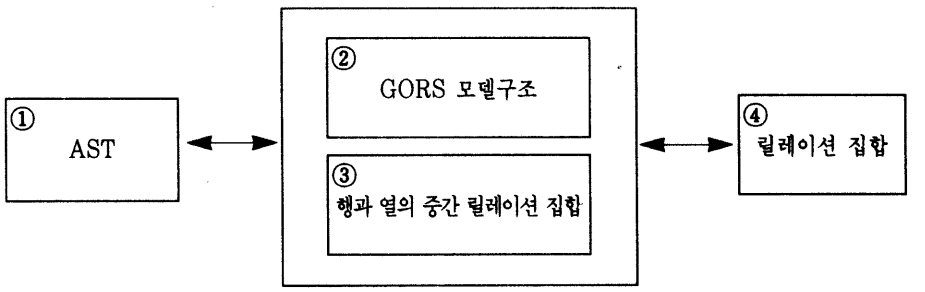


〈그림 1〉 SM-F 시스템을 전위 시스템으로 한 구조

언어이며 시스템 R에 대한 질의어로 소개되어 널리 쓰이고 있는 SQL 등이 있다. 이러한 상업용 관계형 질의어들은 요약 데이타를 다루기 위한 에그리게이션 표현등의 어려움으로 통계적 응용에 적합하지 못하다. 그리하여 참고문헌(13)에서 제안한 질의어와 참고문헌(12)에서 제안한 QBSRT 를 바탕으로, 요약 테이블들인 BAST[20]와 MAST[20]등의 정의 및 접근, 조작에 필요한 질의 기능을 SQL언어의 문법형식을 기준으로 제시함으로써 통계정보를 효과적으로 제공하기 위하여 상업용 관계형 질의언어들이 어떻게 확장되어야 하는지를 제안하고자 한다.

MAST 와 BAST 를 구별하지 않고 어떠한 요약 테이블도 가능하게 두가지 경우를 전부 포함한 경우인 AST(Automatic Summary Table)를 기준으로 정의 및 접근, 조작에 필요한 기능을 살펴 보겠다.

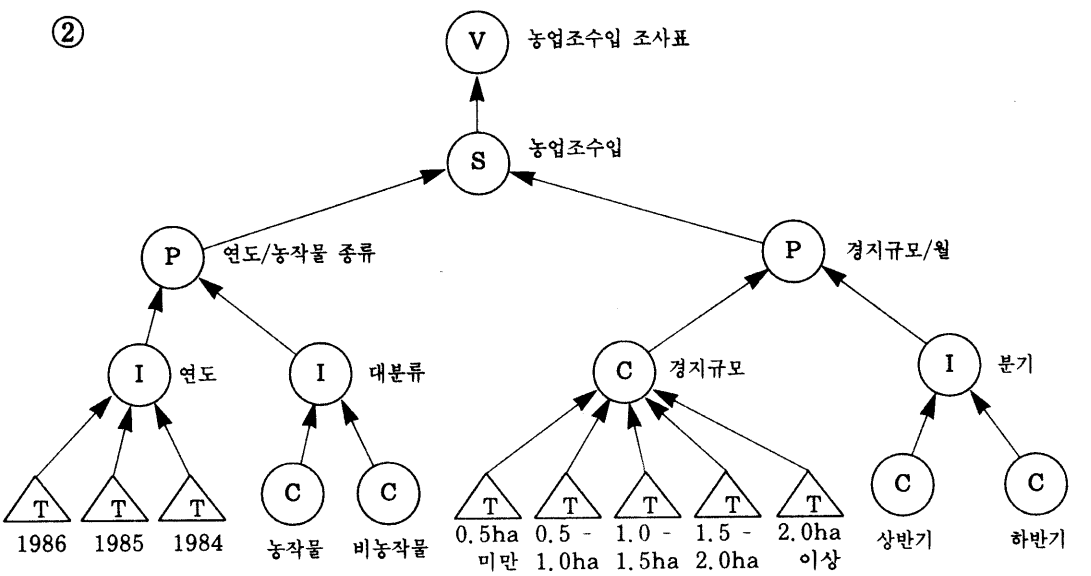
AST 를 생성하기 위한 정의어의 구문은 다음과 같다.



①

조수입		0.5ha 이하		0.5 - 1.0ha		1.0 - 0.5ha		1.5 - 2.0ha		2.0ha 이상	
		상반기	하반기	상반기	하반기	상반기	하반기	상반기	하반기	상반기	하반기
1984	농작물										
	비농작물										
1985	농작물										
	비농작물										
1986	농작물										
	비농작물										

②



③

연 도	대 분 류
1984	농작물
1984	비농작물
1985	농작물
1985	비농작물
1986	농작물
1986	비농작물

경 지 규 모	분 기
0.5 ha 이하	상반기
0.5 ha 이하	하반기
0.5 - 1.0 ha	상반기
0.5 - 1.0 ha	하반기
1.0 - 1.5 ha	상반기
1.0 - 1.5 ha	하반기
1.5 - 2.0 ha	상반기
1.5 - 2.0 ha	하반기
2.0 ha 이상	상반기
2.0 ha 이상	하반기

④

연 도	대 분 류	경 지 규 모	분 기	조 수 입
1984	농 작 물	0.5ha 이하	상반기	
1984	농 작 물	0.5ha 이하	하반기	
1984	농 작 물	0.5 - 1.0ha	상반기	
1984	농 작 물	0.5 - 1.0ha	하반기	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	
1985	비 농작물	1.0ha 이상	상반기	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	
.	.	.	.	

〈그림·2〉 릴레이션과 AST와의 상호 변환 과정

```
CREATE_AST 〈 테이블 이름 〉 FROM〈 기저 릴레이션 이름 〉
        ROM 〈 행 속성 계층구조 순서대로 속성명과 종류명세 〉
        COL〈 열 속성 계층구조 순서대로 속성명과 종류명세 〉
```

이에 대한 예를 들면 다음과 같다.

〈 예 제 1 〉
조수입 조사에서 행은 연도, 열은 경지규모와 상하분기로 구분되는 AST 'TAB #1' 을 정의하라.

```
CREATE_AST  TAB#1    FROM    조수입조사
        ROW  연도 = ( 1984, 1985, 1986 )
        COL  경지규모 = ( 0.5 이하, 0.5-1.0, 1.0-1.5,
                          1.5-2.0, 2.0 이상 )/
             상하분기 = ( 상반기, 하반기 )
```

이와같이 정의한 AST의 측정 데이타의 통계량을 얻어내기 위한 AST의 접근 질의어의 구문은 다음과 같다.

```
SELECT
STATISTICS_NAME    〈 ( 요약 속성들…) 〉
   INTO 〈 디스플레이될 AST 이름 〉
   FROM 〈 AST 이름 셀 명세 〉
```

여기서 셀은 CELL(m,n) 으로, AST의 각 격자를 말하며 행위주(row-wise)의 순서를 갖는다. 만약 셀 명세가 생략되면 default로 전체 셀을 접근 한다.

이에 대한 예를 들면 다음과 같다.

〈 예 제 2 〉
TAB#1 의 조수입의 평균을 구하라.

```
SELECT
   MEAN(조수입)
   INTO  TAB#1
   FROM  TAB#1
```

〈 예 제 3 〉
TAB#1 의 조수입의 셀(1,3)의 빈도수를 구하라.

```
SELECT
   MEAN(조수입)
   INTO  TAB#1
   FROM  TAB#1.cell(1,3)
```

이때, INTO 부분에는 출력시킬 새로운 형태의 AST를 정의해서 그 이름을 써 주어도 좋다.

지금까지의 AST의 정의어 및 접근 질의어 외에 필요한 AST의 조작 질의어 중 통계분석을 위해 필요로 하는 기능인 프로젝션과 에그리게이션 연산에 대한 구문을 살펴 보겠다.

프로젝션(PROJECTION)은 정의된 AST로부터 사용자가 원하는 범주 속성들을 추출하기 위하여 행이나 열의 수를 줄임으로써 부분 정보를 구해주는 연산으로 구문은 다음과 같다.

```
PROJECT 〈 행 : 열  범주속성 계층구조
          순서대로 속성명과 종류명세 〉
          ON COL : ROW
   INTO 〈 디스플레이 될 AST 이름 〉
   FROM 〈 AST 이름 〉
```

이에 대한 예를 들면 다음과 같다.

〈 예 제 4 〉
TAB#1 의 행 범주속성 연도가 1984인 난을 구하라.

```
PROJECT 연도 = ( 1984 ) ON ROW
   INTO   TAB#2
   FROM   TAB#1
```

에그리게이션 연산은 AST의 각 셀에 대하여 원하는 행 또는 열을 에그리게이트 하는 연산으로 구문은 다음과 같다.

```
AGGREGATE 〈 행 : 열  범주속성 계층구조 순서
             대로 속성명과 종류명세 〉
   ON COL : ROW
   INTO 〈 디스플레이될 AST 이름 〉
   FROM 〈 AST 이름 〉
```

이에 대한 예를 들면 다음과 같다.

〈 예 제 5 〉
TAB#1 의 열속성 상하분기에 해당하는 난들을 에그리게이션 하라.

```
AGGREGATE  상하분기  ON  COL
```

```
INTO    TAB#3
FROM    TAB#1
```

〈예 제 5〉처럼 한 속성 전체를 에그리게이션 할 때는 속성명만 명세하고 종류는 필요가 없으나 다음의 예제처럼 한 속성의 일부를 에그리게이션 하는 경우 반드시 속성명과 함께 종류도 명세되어야 한다.

〈 예 제 6 〉
TAB#1 의 행속성 연도의 1984년과 1985년을 에그리게이션 하라.

```
AGGREGATE 연도 = ( 1984, 1985 )
                        ON ROW
    INTO    TAB#4
    FROM    TAB#1
```

V. 결 론

통계 데이타 베이스를 관리하기 위해, 기존의 범용 데이타 베이스 관리 시스템을, 그대로 이용하기에는 통계 데이타 베이스의 특성상 데이타 저장과 액세스의 비효율성, 함수성과 사용의 편이성의 부족 등으로 사용자의 요구를 만족시키지 못해 통계 분석을 적절히 수행할 통계 데이타 베이스 관리 시스템이 필요하게 되었다. 그러나 독자적 데이타 관리 시스템을 이용하기 어려운 현실적 제고를 고려하여, 이 논문에서는 관계형 데이타 베이스 관리 시스템을 목적 시스템으로 한 전위 시스템인 SM-F 시스템을 설계하여, 이를 이용하여 통계 데이타 베이스를 관리하는 방법을 제시하였다.

이 시스템은 통계 데이타 베이스의 효율을 고려한 시멘틱 모델인 GROS 모델을 사용하며 통계분석을 지원하고 통계 요약 정보를 제공하기 위해, 메타 데이타 베이스와 요약 데이타 베이스를 저장하고 운영한다.

또한 관계형 데이타 베이스로의 접근 및 변환 과정을 AST와 릴레이션 집합의 중간 매체로 GROS 모델 구조와 행과 열의 중간 릴레이션 집합을 둠으로써 해결하였으며, 요약 정보 처리를 위해 관계형 질의어들이 어떻게 확장되어야 하는지를 제안하였다.

계속하여 설계된 SM-F 시스템을 구현하여 실제 릴레이션 데이타 베이스 관리 시스템에의 전위 시스템으로 사용하는 후속 연구가 이루어져야 하며 또한 이 시스템에서 제공하는 각종 통계 정보가 의사 결정 지원의 수단으로 적절히 제공되기 위하여 통합적 통계 정보 시스템으로의 실현을 추진하기 위한 연구가 진행 되어야 할 것이다.

참 고 문 헌

1. Becker, R.A.,〈 Data Manipulation in the S System for Interactive Data Analysis 〉, First Workshop on SDB, 1981, pp.155 - 156
2. Bishop, Y.M. and Freeman, S.R., 〈 Classification of Matadata 〉, Second International Workshop on SDB, 1983, pp.230 - 234.
3. Burnett, R.A. and Cowly, B.P. and Thomas, J.J.,〈 Mamagement and Display of Data Analysis Environment for Lage Data Sets 〉,Second International Workshop on SDB,1983,pp.22 - 31.
4. Burnett, B.A. and Thomas, J.J.,〈 Data Management Support for Statistical Data Editing and Subset Selection 〉,First LBL Workshop on SDB, 1983, pp.88 - 102.
5. Chan, P. and Shoshani, A.,〈 SUBJECT : A Directory Driven System for Organizing andAccessing Large Statistical Database 〉, LBL, Pers. on SDB, 1982, pp.29 - 68.
6. Chin, F.Y. and Ozsoypglu, G.,〈 Statistical Database Design 〉,ACM Trans. on Software Engineering, Vol.6, No 1, Mar., 1981.
7. Chin, F.Y. Ozsoypglu, G.,〈 Auditing and Inference Control in Statistical Database 〉, IEEE Tran. on Software Engineering, Vol. 8, No 6, Nov. 1982.
8. Cubitt, Roger E.,〈 Meta Data : An Experience of its Uses and Management 〉, Second International Workshop on SDB, 1983, pp.167 - 169.
9. Curtice, R.M.,〈 Data Dictionaries : An Assessment of Current Practice and Promblem 〉,Proc. of the 7th Int''l Conference on VLDB, 1981, pp.564 - 570.

10. Eggers, S. J. and Shioshani,〈 Efficient Access of Compressed Data 〉, A LBL Perspective on SDB, 1982, pp. 179 - 189.
11. Gey, F.,〈 Data Definition for Statistical Summary Data or Appearances Can Be Deceiving 〉, First LBL workshop on SDB, 1981, pp. 3 - 18.
12. Ghosh, S. P.,〈 Statistical Relational tables for Statistical Database Management 〉, IEEE Trans. on Software Engineering, Vol. SE-12, No 12, DEC 1986, pp. 1106 - 1116.
13. Ozsoyoglu, G. and Ozsoyoglu, Z. M., Mata, F., 〈 A Language and Physical Organization Technique for Summary Tables 〉, Proc. ACM, SIGMOD Conf., Austin, Texas, May 1985, pp. 3 - 16.
14. Ozsoyoglu, G. and Ozsoyoglu, Z. M.,〈 An extention of Relational Algebra for Summary Tables 〉, Second International Workshop on SDB, 1983, pp. 202 - 211.
15. Pistor, P. M.,〈 Designing a Generalized NF Model with an SQR_Type Language Interface 〉, Proc. of 12th Int'l Conference on VLDB, 1986, pp. 278 - 285. (16) Sato, Hideto, Hodaka Rypsuke, 〈 For large Meta Information of National Integrated Statistics 〉, First LBL Workshop on SDB, 1981, pp. 206 - 223.
17. Shamkant, B. N. and Tames, B. F., 〈 Restructuring for Large Database : Three Level of Abstraction 〉, ACM Trans. on DBSystem. Vol 1. No2, June 1976, pp. 138 - 158. (18) 안성옥, 서보환, 박세권,〈 농업 데이타 베이스 구축의 효율적 방향-통계 데이타 베이스적 관점에서 〉, 한국농촌경제 연구소, 연구 보고, 148 - 6, 1988. 12
19. 안성옥, 황종선,〈 통계 데이타 베이스의 효율적 처리를 위한 DSM 시스템의 설계 및 구현에 관한 연구 〉, 한국 정보과학회, 89년 봄 학술발표 논문집, 1989.
20. 안성옥,〈 통계 데이타 베이스의 효율적 관리에 관한 연구 〉, 고려대학교, 박사논문, 1989.
21. 안성옥, 황종선,〈 통계 데이타 베이스의 효율적 관리를 위한, 요약 테이블의 설계 및 구현 〉, 고려대학교 이학론집, 제 30 집, 1990.
22. 최원석,〈 統計分析 지원을 爲 한 統合要約 데이타 테이블의 設計 및 具現 〉, KAIST, 電算學科 석사 논문, 1986.
23. 한상만, 이충일, 정찬진, 송영기,〈 메타 시스템에서의 데이타 베이스 自動生成硏究 〉, 1988 통계 데이타 베이스 학술 세미나 논문집, 제 2 권 제 1 호, 1988.