# 統計/科學 데이타베이스를 위한 個體-側面 模型

○유 철 중
기전여자전문대학 전자계산과

# An Entity-Aspect Model for Statistical and Scientific Databases

Cheol Jung Yoo
Kijeon Women's Junior College

ABSTRACT

This paper analyzes the statistical and scientific entity-aspect model for statistical and scientific databases(SSDB's). The statistical and scientific entity-aspect model(SEAM) is defined an example of the application of the statistical and scientific entity-aspect model is represented.

Finally, the statistical and scientific entity-aspect model as a design tool for SSDB is evaluated and the further research areas are suggested.

## 1. 序 論

統計/科學 데이타베이스(statistical and scientific database,SSDB)는 새로운 연구분야로서 統計 데이타베이스는 統計的 分析을 주로 取級하는 데이타베이스를 뜻하며 計量經濟 자료, 社會環境 자료, 人口 자료, 品質管理 자료등을 포함한다. 科學 데이타베이스는 과학적 實驗이나 시뮤레이션(simulations) 에 의해 收集된 자료를 포함하며 分子 物理學, 流體 力學 시뮤레이션등이 여기에 해당된다. 일반적으로 科學 데이타베이스는 測定된 자료는 물론 實驗機具에 관한 자료, 實驗環境, 實驗 構成등에 관한 자료를 包含하기 때문에 全數(標本) 調査 자료와 같은 엄밀한 통계 데이타보다 複雜하다[2].

대부분의 科學 데이타베이스가 결국 統計的으로 분석되기 때문에 統計/科學 데이타베이스 응용은 類似한 性質을 많이 갖고, 따라서 統計/科學 데이타베이스 應用에 대한 硏究가 함께 이루어지고 있다[2].

本 論文은 統計的 性格을 갖는 데이타와 演算의 適切한 取級을 위한 模型의 考案에 關한 事項을 硏究한다. 統計/科學 데이타베이스의 模型化는 統計的 分析의 描寫가 主를 이루는데 常用 데이타베이스 模型은 複雜하고 多樣하면서도 尨大한 統計 데이타베이스의 特徵을 表現하는데 不適當하다[7,9]. 따라서 別途의 統計 데이타베이스 模型이 硏究되어 왔는데 여기에는 SUBJECT[5], GRASS[4], LOGS[12]와 常用 데이타베이스 模型을 擴張한 SAM*[8], 統計的 個體-關聯性 模型[11]등이 있다. SUBJECT, GRASS, SAM* 등은 統計的 分析 表現에는 强點을 갖지만 設計 道具로서 使用하기에는 問題點이 따른다. 그러한 問題點을 解決하는 側面에서 統計的 個體-關聯性 模型등이 硏究되고 있다.

本 論文에서는 個體 노드(entity node) 와 側面 노드(aspect node) 를 主로 하고, 여기에 汎疇屬性(category attribute)과 要約屬性(summary attribute) 등을 表示하여 階層的 트리구조로 表現함으로써 하나의 個體를 여러 側面에서 觀察할 수 있고 해당 個體의 세부적 階層構組를 나타낼 수 있는 統計/科學 데이타베이스 모형을 제안한다. 이는 SSDB's 의 多次元性(multidimensionality) 및 多重 데이타베이스 (multiple databases)의 개념에 起因한다[2].

이 후 2章에서는 통계/과학 데이타베이스의 특징을 分析하고 기존 통계 데이타베이스 模型의 문제점을 알아보며 3章에서는 통계/과학 개체-측면 모형을 定義한다. 4 章에서는 이 모형을 이용한 통계/과학 자료의 模型化(modeling)의

단계를 설명하고 例를 보이며, 5 章의 結論에서는 통계/ 과학 개체-측면 모형을 개괄적으로 평가하고 앞으로의 研究方向을 提示한다.

## 2. 統計/ 科學 데이타베이스의 特徵 分析

### (1) 통계/ 과학 데이타베이스의 특징

상용 데이타베이스와는 달리 null 값이 많아 散在性(sparseness)를 띠는 방대한 자료를 취급하며 行列, 벡터, 排列, 多次元, 時系列(time series) 등 다양한 형태의 데이타를 갖는다[1,2,3,11].

통계적 분석에 관계하는 속성은 범주속성과 요약속성으로 분류된다. 또한 통계적 演算인 요약화가 계층적으로 여러 단계에 걸쳐 일어날 수 있으며 메타 데이타(meta-data) 와 履歷 데이타(temporal data) 가 큰 비중을 차지한다. 이러한 구조적, 내용적 특징을 가지는 통계/ 과학 데이타베이스를 표현하기 위한 모형 연구가 계속되고 있다[4,5,8,11].

### (2) 기존 統計 데이타베이스 模型의 문제점

SUBJECT, GRASS, SAM* 등의 모형은 통계 데이타베이스의 내용적 특징인 질의사항의 意味(semantics) 를 잘 나타내나 몇 가지 문제점이 있다[11].

첫째로 논리적 데이타베이스 설계과정에서 모형화는 초기 段階이므로 위 3가지 모형들을 통계/ 과학 데이타베이스의 표현에 사용하면 각 모형에 적합한 데이타베이스 설계의 과정들을 다시 구상하여야 한다[6].

둘째로 통계적 모형은 集計化 演算 표현을 위주로 하므로 現實世界 客體(object)모습 보다는 屬性을 표현단위로 한다. 이로써 그러한 속성이 어떠한 객체에 속하는지를 알아내는 일이 복잡해지고 객체의 모습이 명확하지 못하면 데이타베이스의 설계에 혼란이 생긴다.

統計的 個體-關聯性 模型은 개체를 표현단위로 함으로써 관계성 형성을 명확히 하고 모형 확장 및 뷰(view)統合을 원활히 하는 등 우수한 모형으로 평가되나 요약화의 다단계 표현을 위한 관련성키리의 계층적 구조가 갖는 限界性 등이 문제가 된다.

## 3. 統計/ 科學 個體-側面 模型

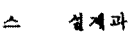
個體-側面 模型(Entity-Aspect Model) 은 개체(entity 또는 system entity) 노드와 측면(aspect)노드의 두가지 형을 갖는 노드들로 구성되며, 개체는 측면에 따라 서로 다른 계층적 트리구조를 갖게되므로 하나의 개체를 여러가지 측면에서 관찰할 수 있음은 물론 개체의 細部的 계층구조를 나타낼수 있는 장점을 지니는 모형이다.

개체-측면 모형은 하나의 개체에 대하여 選擇된 측면에 따른 세부관계와 구조를 분석하는데 있어 stepwise refinement 와 specialization을 사용하는 접근방식을 취한다[10].

개체-측면 모형을 좀더 확장하여 이를 統計/ 科學 個體-側面 模型(Statistical and Scientific Entity-Aspect Model,SEAM) 이라 하고 이 모형을 이용하여 내용적 특징을 위주로 하는 통계/ 과학 데이타베이스의 描寫를 어떻게 할 것인가를 연구한다.

### (1) 시스템 개체의 표현

模型化하고자 하는 통계/ 과학 자료의 분석대상 시스템 개체를 표현하기 위하여 다음과 같이 정의한다.

[ 定義 1]: 통계/ 과학 자료의 분석 대상이 되는 複合的 最上位 개체를 시스템 개체(system entity, 複合性 個體) 라 하고 "  " 으로 표현한다. 이는 根 노드(root node) 가 된다.

### (2) 범주속성의 표현

개체를 구성하는 속성 中 통계적 질의사항에서 범주속성으로 사용되는 속성을 표현하여야 한다. 여기에서 統計的 個體-關聯性 模型의 정의를 응용 [ 定義2]을 정의한다.

[ 定義2]: 질의사항에서 汎疇化를 위하여 사용하는 속성을 범주속성이라 하고 링크(link)를 "" 로 표현한다.

### (3) 요약속성의 표현

어떤 속성이 요약속성이 되는가를 표현하기 위해 [定義3]를 정의한다.

[ 定義3]: 범주속성에 의해 集團化되는 속성을 요약속성이라 하고 개체 노드에 연결하여 "

"으로 표현한다.

(4) 要約化의 多段階

통계적 분석을 거친 결과를 가지고 다시 새로운 통계적 분석을 하는 질의사항이 존재하는 경우 이러한 상황을 명확하고 간단하게 표현할 수 있게 하며 따라서 複合演算과 質疑事項의 계층적 상황에 대한 명확한 표현이 가능해진다. 통계/ 과학 개체- 측면 모형에서는 통계/ 과학 데이타베이스의 效率的 表現을 위해 통계적 분석 대상을 하나의 시스템 개체로 표현하므로 요약화의 다단계가 個體間의 階層性으로 나타나게 된다.

이로써 하나의 개체( 또는 시스템 개체) 를 여러 측면으로 分類, 觀察하여 세부적 계층구조를 나타낼 수 있는 통계/ 과학 개체- 측면 모형을 정의하였다. 표 1은 통계/ 과학 개체- 측면 모형의 표현 방법을 보여준다.

표 1. 통계/ 과학 개체- 측면 모형의 표현

Table 1. Description of SEAM.

| 명 칭 | 표 현 |
|---|---|
| 개 체 (entity) | |
| 측 면 (aspect) | |
| 속 성 (attribute) | |
| 범주속성(category attribute) | |
| 요약속성(summary attribute) | |
| 시스템 개체 (system entity) | |

## 4. 統計/ 科學 個體- 側面 模型化

먼저 개체를 정한 다음 그 개체에 대한 여러 측면의 결정이 필요하며 존재할 수 있는 범주속성 및 요약속성을 찾아내야 한다. 통계/ 과학 개체- 측면 모형을 이용한 모형화는 다음의 단계들을 循環的으로 適用하면서 完成된다.

[ 段階1]: 모형화 하고자 하는 統計/ 科學 資料의 분석 대상 시스템 개체 (system entity, 복합성 개체) 또는 단순 개체(simple entity) 를 선정한다. 이 노드는 根 노드(root node) 가 된다.

[ 段階2]: [ 段階 1] 의 개체에 대한 여러 가지 개체 E1,E2,.....,En 을 고려한다. 단,[段階1]에서 선정된 개체가 단순개체이면 [단계2]를 省略한다.

[ 段階3]: 각 개체에 존재하는 속성 $a_1,a_2,.....,a_n$ 을 표시(stepwise refinement) 하고 해당 개체가 요약속성을 필요로 하는 경우 요약속성 Si(i=1,n) 표현한다.

[ 段階4]: 속성 $a_1,a_2,.....,a_n$중 어떤속성 측면에서 통계적 분석을 요하는 경우 측면 A1,A2,.....,Am을 표현하고 속성별로 범주속성 Cj(j=1,m)를 표현한다.

[ 段階5]: 選擇된 側面 Ai 에 對하여

i) 동일한 측면이 다른 개체하에 존재하면 종결한다.

ii) 계속적인 specialization 이 고려되는 경우 필요한 개체를 표현한다. 이에 대한 속성이 계속 요구되면 [段階3]으로부터 다시 적용하고 어떤속성도 요구되지 않으면 해당 개체에 대한 요약속성(Si) 및 데이타 客體(data object) $O_1,O_2,.....,O_k$ 를 표시하고 종결한다.

iii) 계속적인 specialization 이 고려되지 않는 경우 側面에 대한 데이타 객체 $O_1,O_2,.....,O_k$를 표시하고 종결한다.

이처럼 통계/ 과학 개체- 측면 모형화는 循環的이며 stepwise refinement 와 specialization 을 하여 행하여진다.

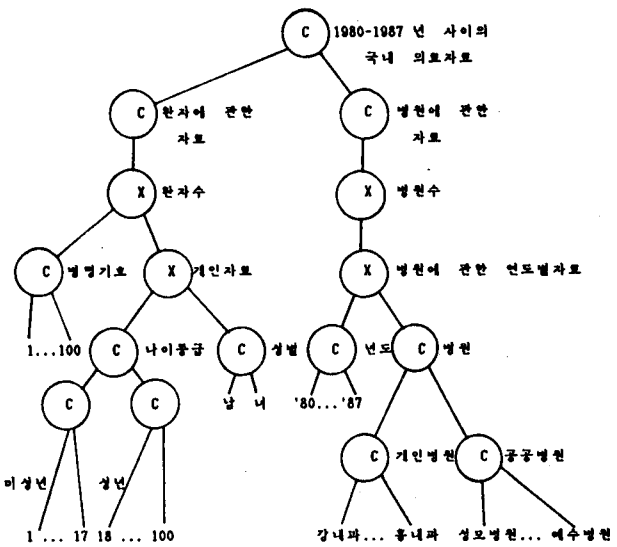



그림 1. SUBJECT 모형을 이용한 의료자료 모형

Fig.1. Information model for hospital using SUBJECT model.

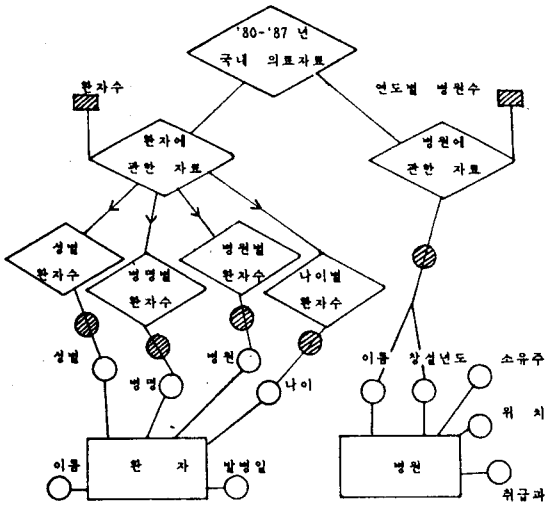



그림 2. 통계적 개체 - 관련성 모형을 이용한 의료자료 모형

Fig.2. Information model for hospital using statistical entity-relationship model.

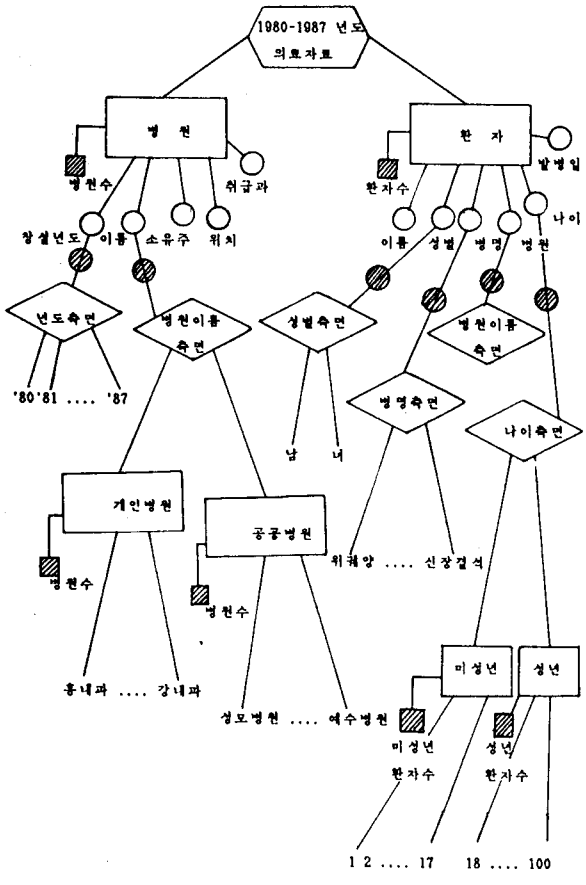



그림 3. 통계/ 과학 개체- 측면 모형을 이용한 의료자료 모형

Fig.3. Information model for hospital using SSEM.

그림 1, 그림 2 ,그림 3은 각 모형에 의한 醫療資料 模型化를 보여준다. 그림 1은 SUBJECT모형을 이용한 의료자료 모형이고 그림 2는 통계적 개체-관련성 모형을 이용한 의료자료 모형[11]이다. 동일한 의료자료에 대해 통계/ 과학 개체- 측면 모형에 의한 모형화 결과는 그림 3과 같다.

그림 3을 설명하면 우선 [段階 1] 에 의하여 통계 분석 대상 시스템 개체( 복합적 개체) 로 "1980-87년 의료자료" 가 선정되었고 [段階 2] 에 의해서 "병원" 과 "환자" 개체가 고려 되었다. [定義 3] 에 근거하여 "병원" 과 "환자" 개체에 요약속성 "병원수", "환자수" 를 표현하였고 [定義 2] 에 의하여 통계 질의사항에서 사용되는 속성에 대해 범주속성을 표시하였다. " 년도 측면", " 병원이름측면", " 병명측면" 등의 측면노드와 연결되는 속성들 즉, " 창설년도", "이름", "병명" 등은 범주속성이 된다. " 환자" 개체의 " 병원이름" 측면은 개체 " 병원" 하에 존재하므로 더이상 표현하지 않는다. 하나의 측면에 어떤 개체가 다시 존재할 수 있는 경우 이를 표시하고 그 개체에도 요약속성을 표시할 수 있다.

마지막으로 SUBJECT, GRASS등과 같은 모형처럼 最下位 계층인 端 노드(terminal node) 를 이용, 데이타의 구체적인 값을 개체 또는 側面노드에 연결하여 표시하였다.

## 5. 結 論

本 論文에서는 統計/ 科學 데이타베이스 模型化에 사용될 수 있는 統計/ 科學 個體- 側面 模型을 定義하였고 모형화의 例를 보였으며 SUBJECT, 통계적 개체- 측면 모형등과 比較, 圖式하였다.

통계/ 과학 개체- 측면 모형은 하나의 개체를 여러가지 측면에서 모형화할 수 있고 그 개체에 대한 階層的 細部構組를 把握할 수 있으며 일반 모델과 달리 個體, 側面노드를 사용함으로써 設計道具로서 長點을 갖는다. 또한 통계적 분석에 관련된 개체의 범위 및 각 노드의 레벨(level) 을 명확하게 나타낼수 있다. 계층구조에 있어서 각 個體間의 關聯性 표현이 어려운 점을 감안, 측면노드를 사용하여 어느 정도 관련성을 나타내었으며 시스템 個體를 通하여 個體間의 連結을 시도하였다.

통계/ 과학 개체- 측면 모형에 대한 앞으로의

研究 方向은 각 質疑事項을 표현한 部分的 모형을 統合하는 구체적이고 정형적인 連結 規則(connection rule)과 全體 統合 模型에서 특정 質疑事項에 해당되는 사항만을 選擇하는 分技 規則(branch rule)의 연구가 必要하며 계층구조로 因하여 발생되는 실제 具現上의 문제와 各 개체간의 關聯性 표현을 좀더 명확히 하는 方向의 研究가 이루어져야 한다.

## 참 고 문 헌


1. A. Shoshani, "Statistical Databases : Characteristics, Problems and Some Solution," Proc. VLDB, pp. 208-222, 1982.
2. A. Shoshani and H. K. T. Wong."Statistical and Scientific Database Issues," IEEE Trans. on Software Engineering, Vol. SE-11, No. 11, pp. 1040-1047, 1985.
3. A. Shoshani, et al., "Characteristic of Scientific Databases," Proc. VLDB, pp. 147-160, 1984.
4. M. Rafanelli and F. Ricci, "Proposal of a Logical Model for Statistical Database," Proc.2nd Int. Workshop on SDM, pp. 264-272, 1983.
5. P. Chan and A. Shoshani, "SUBJECT : A Directory Driven System for Organizing and Accessing Large Statistical Database," Proc. VLDB, pp. 553-563, 1981.
6. S. B. Yao, et al., "An Integrated Approach to Database Design," Lecture Notes in Computer Science : Database Design Techniques I,Spring-Verlag, pp. 1-30, 1982.
7. S. M. Dintelman, "Data Models for Statistical Database Applications,"IEEE Database Engineering, Vol. 7, No. 1, pp. 38-42, 1984.
8. S. Y. W. Su, "SAM* : A Semantic Association Model for Corporate and Scientific- Statistical Database," Information Sciences 29, pp. 151-199 1983.
9. S. Y. W. Su, et al.,"Logical and Physical Modeling of SSDB," Proc. 2nd Int. Workshop on SDM, pp. 251-263, 1983.
10. 김창화, 백두권, 황종선, " 데이타 베이스 설계를 위한 E-A(Entity-Aspect) 모델의 연구," '87년 봄 학술발표 논문집, 한국 정보 과학회, pp. 90-93, 1987.
11. 박석, 이애영, " 통계 데이타베이스를 위한 개체-관련성 모형의 확장," 정보 과학회 논문지, 제14권, 제1 호, pp. 55-64, 1987.
12. 박석, 이춘희, "LOGS : 통계 데이타베이스를 위한 논리적 그래프 모델," '85년 가을 학술발표 논문집, 한국 정보 과학회, pp. 181-191, 1985.