

복합생성 자료검색의 모형화

이 춘 열¹⁾

A Modelling of Multi-derived Data and Its Retrieval Scheme

Current database systems are based on the assumption that a datum denotes the same meaning; however, in reality, the violation of this assumption is not unusual. Some data are created in such a way that they represent different sets of attribute values. The current research formulates this phenomenon as dissimilarities of derivation rules and defines multi-derived data as ones that are derived by multiple rules. For multi-derived data, this research proposes a new retrieval scheme and analyze its implication with relation to data retrieval.

1) 국민대학교 정보관리학과

I. 서론

본 논문은 기존의 전통적 자료검색 모형이 자료의 생성특성을 충분히 반영하지 못하고 있음에 기초하고 있다. 일반적으로 자료 생성이라 함은 데이터베이스내에 저장된 속성 값들이 부여되는 절차를 의미한다. 예를 들면 '신장'이라는 자료의 생성 또는 생성법칙은 신장을 측정, 기록하는 과정을 의미하며, '업무 평점'의 생성법칙은 해당 사원의 업무수행능력의 평가 절차이다.

자료의 생성법칙은 정보의 공유 및 자료 검색과 연관하여 중요한 의미를 지닌다. 예를 들면 인력 데이터베이스에 특정인의 신장이 170cm 로 기록되어 있다고 가정하자. 이 자료는 해당 사람의 키가 신장이 170cm 로 기록된 다른 사람과 같으며, 신장이 165cm로 기록된 사람보다는 크다는 것을 의미한다. 그러나 만약 데이터베이스에 기록된 신장이 동일한 기준이 아닌 서로 다른 높이의 구두를 신은 상태에서 측정되었다고 가정하면, 이와 같은 해석은 항상 성립하지 않는다. 즉, 데이터베이스내에 저장된 같은 자료값들이 같은 의미를 나타냄은, 이들이 동일한 생성법칙에 의하여 생성되었음을 가정한다. 부연하면, 같은 속성(예: 신장)을 나타내는 자료값들은 동질의 생성법칙(예: 신장의 측정 기록 과정)에 의하여 만들어졌으며, 따라서 같은 자료값은 (예: 170cm) 같은 의미를 지니는 것으로(즉, 해당인들의 신장이 같다는 것으로) 해석된다.

전통적으로 데이터베이스내 저장 자료들은 동일한 생성법칙에 의하여 만들어졌으며, 이에 따라 같은 자료값은 동일한 의미를 지님을 전제로 검색 공유되어 왔다. 그러나 자료 생성의 동질성이 모든 경우에 대하여 항상 만족되는 것은 아니다. 예를 들면 회사원의 업무평점의 평가기준은 근무부서에 따라 서로 상이할 수 있다. 또한 사람의 신장도, 예에서 제시된 바와 같이, 경우에 따라서는 다른 방법에 의하여 측정된 값들이 혼합 저장될 수 있다.

생성법칙의 비동질성이 정보의 공유 및 자료 검색과 연관하여 중요한 의미를 지니는 이유는 자료의 의미의 비동질성을 야기하기 때문이다. 예를 들어 데이터베이스내 저장된 회사원의 업무평점이 근무부서에 따라 서로 상이한 평가기준에 의하여 평가되었다면, 비록 동일 업무평점 (예: "A, B, C, D" 또는 "우수, 불량")이라 하더라도 서로 상이한 의미를 지니게 된다. 바꾸어 이야기하면, 판매부서의 우수 사원이 반드시 연구부서의 우수사원이 되리라는 보장이 없으며, 역 또한 성립하지 않는다. 이와 같이 이질적인 생성법칙에 의하여 생성된 자료의 의미는 자료값과 더불어 자료의 생성법칙을 동시에 고려하여 해석되어야 한다.

예를 들면, 생산부서와 판매부서로 부터 연구부서 요원을 충원한다고 가정할 경우, 선발 대상자들은 각자의 업무 평가치와 더불어 해당 근무부서의 평가기준을 동시에 고려하여 선발되어야 한다. 부연하면 사원

의 업무수행능력은 업무평점과 더불어 근무부서의 평가기준을 동시에 고려하여 해석되어야 한다.

기존의 전통적 자료검색 모형은 이와 같은 복합적인 검색기준을 직접 지원할 수 있는 방안을 제시하지 못하고 있다. 이러한 사실에 기초하여, 본 논문은 다음의 질문에 대한 해답을 제시하고자 한다.

데이터베이스내 저장 자료들이 서로 상이한 복합적인 생성 법칙에 의하여 기록되었을 경우, 이러한 특성이 자료 공유를 위한 검색구조에 어떠한 영향을 미치며, 이들 자료의 가장 효과적인 검색 방안은 무엇인가?

이를 위하여 제2장에서는 본 논문의 주 연구 대상인 복합생성자료의 특성을 정의하고 이들 특성이 자료검색에 미치는 영향을 분석한다. 제3장에서는 복합생성자료의 검색모형을 제시하고, 제시된 모형의 효과성을 분석 제시한다. 끝으로 제4장에서 본 연구의 제한점 및 추후 연구방향을 요약 제시한다.

II. 복합생성 자료

사원의 업무수행능력 평가치와 같이 동일 속성의 자료값을 나타낸다고 하더라도 서로

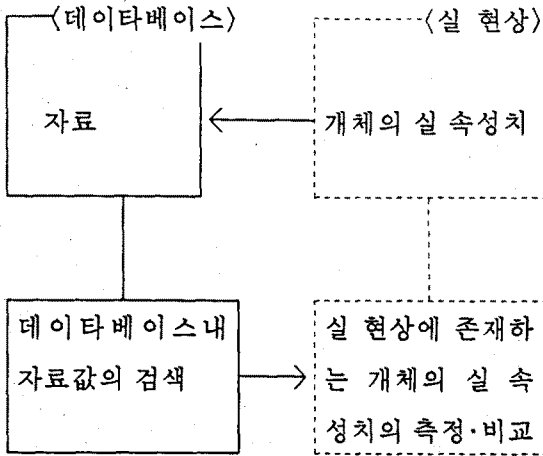
상이한 의미를 지니는 자료들을 본 연구에서는 복합생성자료라 한다²⁾. 제1장의 예에서 제시된 바와 같이 복합생성자료의 검색은 기존의 전통적 자료검색과 달리 자료값뿐만 아니라 자료의 생성법칙 또한 동시에 고려하여 수행되어야 한다. 이러한 복합생성자료가 기존의 전통적 자료와 비교하여 지니는 특성 분석을 위하여, 본 연구는 자료검색 구조를 개념화하고, 이를 기초로 하여 전통적 자료검색의 특성 및 복합생성자료의 특성을 정의한다.

1. 자료검색의 개념화

본 연구는 자료검색을, 〈그림 1〉에 제시된 바와 같이, 데이터베이스내 저장 자료값을 이용한 실 속성치의 비교과정으로 정의한다. 〈그림 1〉에서 제시된 바와 같이 자료검색의 궁극적인 목적은 실 현상에 존재하는 개체의 속성치를 측정 또는 비교하는 것이다. 따라서 자료검색이란 이들 속성치들을 직접 측정 또는 비교하는 대신에 데이터베이스내의 자료값들을 이용하여 대행하는 과정이다. 즉, 자료값들은 실 속성치를 부호화한 것이며, 자료값을 이용하여 실 속성치를 간접적으로 측정 비교하는 과정이 자료검색이다.

자료검색의 개념도에서 실 속성치에 대응하는 자료값을 부여하는 과정이 자료의 생

2) 복합생성자료의 특성 및 구체적인 정의는 본 장의 후반부에 제시됨.



〈그림 1〉 자료검색의 개념도

성이다. 부연하면 자료의 생성이란 실 속성치별로 대응하는 자료값을 부여하는 과정이며, 자료값을 부여하는 법칙이 자료의 생성법칙이다. 이러한 관점에서 볼 때 자료의 생성법칙이란 개체의 속성치에 대응하는 자료값을 부여하는 사상(mapping)으로 표시된다. 즉, 자료의 생성법칙을 f 라 하면, f 는 다음과 같이 속성치 영역으로 부터 자료값 영역으로의 사상으로 정의된다.

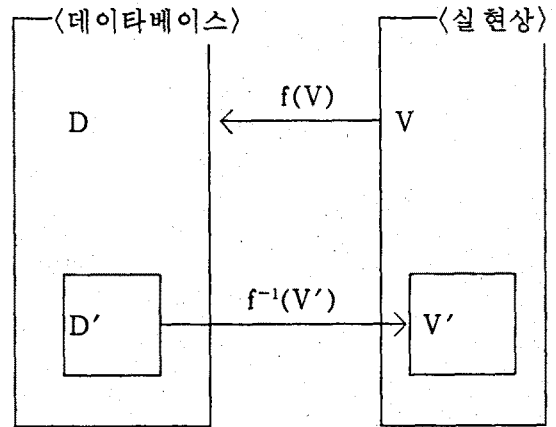
$$f : V \longrightarrow D$$

V : 속성치 영역 (value domain)

D : 자료값 영역 (data domain)

이상에서 정의된 자료 생성법칙을 이용하여 자료검색 과정을 구체화하면 다음과 같다. 자료검색은 실 현상에 존재하는 개체의 실 속성치 (즉, 속성치 영역 V 의 구성 원소

또는 부분 집합을) 또는 이들 속성치를 지니는 개체를 자료값 영역 D 의 구성원소를 이용하여 비교 추출하는 과정이다. 즉, V 의 부분집합인 V' 또는 V' 의 속성치를 지니는 개체들을 추출하기 위하여, 데이터베이스를 매개로 하여, 자료값 영역의 부분집합인 D' 을 추출하고 이에 대응하는 V' 을 유도하는 과정으로 설명된다. 이러한 관점에서 〈그림 1〉에 제시된 자료검색의 흐름을 생성법칙 f 를 매개로 하여 재 정리하면 〈그림 2〉와 같이 도시된다.



〈그림 2〉 자료검색의 개념도와 자료 생성법칙

2. 자료 특성의 분류 기준

전통적으로 데이터베이스내 각 자료값들은 대상개체의 속성(치)에 대하여 완전한 정보를 제공하는 것으로 가정된다. 이를 완전정보가정(complete information assump-

tion)이라 하며, 완전 정보가정을 만족시키는 자료를 완전자료(complete data)라 한다. 보다 구체적으로 완전정보가정이란 데이터베이스내의 자료값이 개체의 실 속성치를 비교 평가하기 위한 완전한 기준으로 사용될 수 있음을 의미한다. 본 연구는 완전자료를 가장 전형적인 자료유형으로 설정하고, 완전정보가정을 자료검색의 개념도를 이용하여 구체화한다. 이러한 과정을 통하여, 완전자료 및 불완전자료의 분류 기준을 도출하고, 본 연구의 주 대상인 복합생성자료의 특성을 정의 제시한다.

완전정보가정이 만족되기 위하여서는, 데이터베이스내의 자료값이 개체의 실 속성치를 비교 평가하기 위한 완전한 정보를 제공할 수 있어야 한다. 이를 전제에서 제시된 자료의 생성법칙 및 자료검색의 개념적 모형에 기초하여 분석할 때, 자료값 영역 및 생성법칙과 관련하여 다음의 두 조건이 충족되어야 한다.

- ① 자료값 영역에 대한 가정: 대상 개체의 속성(치)가 상호 독립적인 자료값에 의하여 기록되어야 한다.
- ② 생성법칙에 대한 가정: 자료의 생성법칙이 동일하여야 한다.

첫번째 조건인 개체의 속성치가 상호 독립적인 자료값에 의하여 기록되어야 함은 자료값영역(data domain)을 구성하는 모든 자료값 중에서 해당 개체의 속성치를 나타

낼 수 있는 자료값은 유일하게 하나뿐이어야 함을 의미한다. 바꾸어 이야기하면, 하나의 자료값이 개체의 속성치를 나타내는 진의 값이면 나머지 자료값들은 해당 속성치를 나타내는 진의 값이 될 수 없음을 의미한다. 예를 들면 3자리 정수값으로 구성된 자료영역상에서 특정인의 신장을 반영하는 자료값이 170(cm)라는 것은 여타의 모든 자료값, 즉 168(cm) 또는 190(cm)들은 특정인의 신장을 반영하는 진(眞)의 자료값이 될 수 없음을 의미한다. 이러한 가정을 만족시키는 자료값들을 크리스섬 데이터(crisp data)³⁾라 한다.

크리스섬 데이터의 가정을 자료 생성의 측면에서 살펴보면 대상 개체의 속성치에 대응하는 자료값이 유일하게 존재함을 말한다. 즉, 임의의 속성치 $v \in V$ 에 대하여 $f(v) = d$ 이면, d 와 상이한 모든 d' 에 대하여 $f(v) \neq d'$ 가 성립하여야 한다. 그러나 모든 자료값들이 다 크리스섬 데이터 가정을 만족시키는 것은 아니다. 예를 들면 신장을 나타내는 자료값의 예에서 특정인의 신장이 '약 170'이라는 자료값으로 표현되었다고 가정하자. 이 경우 '약 170(cm)'라는 자료값은 171cm 또는 169cm도 해당 사람의 신장을 반영하는 진(眞)의 자료값이 될 수 있음을 의미한다. 따라서 '약 170cm'라는 자료값은 171cm 또는 169cm가 해당 신장을 나타내는 자료값

3) 크리스섬 데이터는 Zadeh[1965]에 의하여 제시된 퍼지 집합(fuzzy set)에 대응하는 크리스섬 숫자(crisp number)를 데이터로 확대한 개념이다.

이 될 수 없음을 한정하지 못하며, 따라서 대상개체의 속성(치)에 대하여 완전한 정보를 제공하지 못한다. 이와 같이 모든 자료값이 여타 모든 자료값에 대하여 해당 개체의 속성치가 될 수 없음을 한정하지 못할 경우 이들 자료값들은 실 속성치를 비교 평가하기 위한 완전한 기준으로서 제한점을 지닌다.

크리셋 데이터 가정과 더불어 완전정보가정을 충족시키기 위한 추가적인 조건으로서 생성법칙의 동질성에 대한 가정이다. 생성법칙의 동질성에 대한 가정이란 자료값들이 동질의(자료생성) 법칙(homogeneous derivation rule)에 의하여 부여되어야 한다는 것이다. 여기서 동질의 법칙이라 함은 개체의 속성치가 서로 같을 경우 이들의 자료값이 동일하게 부여되어야 함을 의미한다. 예를 들면 신장이 같은 두 사람이 있을 경우, 신장을 나타내는 자료값이 동질의 자료생성법칙에 의하여 만들어졌음은, 이들이 같은 자료값으로 표시됨을 의미한다.

앞에서 제시된 크리셋 데이터 가정이 자료값 영역 자체의 완전성을 의미한다면, 생성법칙의 동질성에 대한 가정은 자료값이 부여되는 생성법칙의 완전성을 의미한다. 자료의 생성법칙이 동일하면 같은 속성치는 같은 자료값으로 표시되며 따라서 생성법칙 f 는 속성치 영역 V 로 부터 자료값 영역 D 로의 함수로 표시된다. 바꾸어 이야기하면

특정 속성치에 대하여 V 로 부터 D 로의 함수는 단 하나만이 존재한다.

이상에서 제시된 완전정보가정의 성립조건들을 요약하면, 완전정보의 제공이란 속성치에 대응하는 자료값이 유일하게 하나만 존재하여야 하며, 이를 위하여서는 크리셋 데이터 가정과 동질의 생성법칙 가정이 동시에 만족되어야 한다. 이들 두 조건이 동시에 충족될 경우 특정 자료값 d 가 나타내는 속성치 $f^{-1}(d)$ 가 속성치의 집합(crisp set)으로 정의되며, $f^{-1}(d)$ 에 포함된 모든 속성치는 자료값 d 에 의하여 기록된다. 이에 따라 자료값은 속성치를 비교 평가하기 위한 기준으로 사용될 수 있으며, 속성치에 대한 완전한 정보를 제공하게 된다.

모든 자료들이 이들 완전정보가정을 항상 충족시키는 것은 아니다. 완전정보가정을 만족시키지 않는 자료들을 총칭하여 불완전자료라 한다. 불완전자료의 대표적인 예로서는 이미 크리셋 데이터 가정에서 예시된 '약 170cm'와 같이 자료값이 속성치에 대한 부분정보만을 제공하는 경우이다. 부분정보를 제공하는 자료값들을 부분자료(partial data)⁴⁾라 칭하며 대표적인 불완전 자료 유형으로 연구되어 왔다[Lipski, 1979 : Wong, 1982].

부분자료가 크리셋 데이터 가정을 위반한 자료 유형이라면 또 다른 불완전 자료의 유형으로서는 동질의 생성법칙을 위반한 자료를 생각할 수 있다. 여기서 동질의 생성법

4) 대표적인 부분자료로서는 퍼지 데이터, 누락 자료값(null value) 등이 있다.

칙을 위반한 자료 유형이란 자료들이 이질적(즉, 복수의) 생성법칙에 의하여 생성되었음을 칭한다. 즉, 자료의 생성법칙이 둘 이상 존재함으로써 개체의 속성치가 서로 같은 경우에도 다른 자료값이 부여될 수 있음을 의미한다. 이질생성 법칙에 의하여 만들어진 자료의 대표적인 예로서는 본 논문의 서두에 예시된 사원의 업무평점 또는 문헌 검색을 위한 검색어 등이 있다. 이와 같이 복수의 생성법칙에 의하여 생성된 자료를 복합생성자료(multi-derived data)라 칭한다.

복합생성자료는 속성치에 대하여 완전한 정보를 제공하지 못한다는 점에서 부분자료와 유사성을 지닌다. 그러나 복합생성자료는 불완전 정보의 제공 요인면에서 부분자료와 달리 자료값 자체의 불완전성이 아닌 생성법칙의 상이성에 기인한다. 즉, 특정 개체의 속성치가 이질적인 생성법칙들에 의하여 기록될 경우⁵⁾, 자료값에 대응하는 속성치들이 명확하게 정의될 수 없으며, 따라서 자료값은 속성치에 대한 완전정보를 제공하지 못하게 된다. 이러한 복합생성자료는 대부분 자료값이 자연어(예: 검색어) 또는 평가치(예: 업무평가치)와 같이 부분자료일 경우 많이 발생한다. 이에 따라 정통적으로 부분자료의 일 유형으로 취급되어 왔다 [van Rijsbergen, 1979 : Blair, 1984].

본 연구는 복합생성자료를 부분자료와 분리하여 독자적인 데이터 유형으로 제시한

다. 이를 위하여 본 연구는 복합생성자료를 업무평점과 같이 자료값(예: 평가치 자체)과 자료의 생성특성(예: 근무부서)을 결합함으로써 명확히 할 수 있는 경우로 한정한다. 즉, 본 연구에서 고려하는 복합생성자료는 자료의 생성특성을 명확히 할 수 있는 경우로 한정하며, 자료의 생성특성을 한정하는 속성을 조건속성(condition attribute)이라 한다. 업무평점 예에서 근무부서는 평가치의 생성특성을 한정하며, 따라서 근무평점의 조건속성이다. 이에 반하여 조건속성의 식별이 곤란한 복합생성자료 또한 존재한다. 이의 대표적인 예가 문헌 검색을 위한 검색어와 같은 자연어들이다. 일반적으로 자연어는 개인별로 다른 의미를 부여할 수 있음으로서 이의 생성특성을 제한하는 조건속성의 식별이 용이하지 않으며, 식별한다고 하여도 무한히 많은 조건속성치를 지님으로서 조작이 용이하지 않다.

이상에서 논술된 완전정보가정의 세부조건 및 복합생성자료의 분류 특성을 요약하면 다음과 같다.

일반적으로 존재할 수 있는 자료들의 제 특성은 자료영역과 자료생성특성을 결합함으로써 명확시 된다. 이 중 완전자료란 자료값들이 크리섭 데이터로 구성되며, 자료의 생성이 동질의 생성법칙에 의하여 이루어진 자료를 칭한다.

5) 의사결정과 관련하여 Mitroff, et.al [1982, 1983]은 이질적인, 즉 상호 모순적인 법칙들이 존재함을 제시하고 있다.

〈표 1〉 자료의 분류 유형

		생 성 법 칙		
		동질생성법칙 가정 충족	동질생성법칙 가정 위반	
			조건속성 존재	조건속성 식별 곤란
자료값 영역	크리섭 데이터 가정 충족	완전자료	복합생성자료	n.a ¹⁾
	크리섭 데이터 가정 위반	부분자료	복합생성 부분자료	자연어 (natural information)

이들 기준에 따라 자료 유형을 분류하면 〈표 1〉에 제시된 바와 같다.

〈표 1〉에 제시된 바와 같이, 부분 자료는 크리섭 데이터 가정을 위반한 자료들이며, 복합생성자료는 동질 생성법칙 가정을 위반한 자료들을 칭한다. 이들 자료 유형 중에서 본 논문은 복수의 생성법칙에 의하여 생성된 자료로서 조건 속성이 존재하는 자료를 연구대상으로 하며, 이후 이를 복합생성 자료로 한정하여 칭한다.

3. 복합 생성자료 검색의 특성

본절은 복합생성자료의 검색이 지닌 고유 특성을 전통적 자료검색 모형의 틀하에서 분석 제시한다. 전통적 자료검색 모형은 완전대응(exact matching)에 근거하고 있다. 따라서 복합생성자료를 전통적 자료검색에 기초하여 검색할 경우 자료의 검색질의(query formula)는 검색요구를 충분히 반영하지 못하게 되며, 이 결과 찾고자 하는 유

용한 자료(relevant data)를 누락하거나 유용하지 않은 자료(non-relevant data)를 검색하게 된다. 예를 들면 ‘연구부서’, ‘생산부서’ 및 ‘판매부서’의 사원중에서 신설될 새로운 연구부서에 적합한 ‘우수’ 사원들을 검색하고자 한다고 가정하자. 이 경우 업무평점이 ‘우수’로 평가된 모든 부서 사원을 검색한다고 하면, 연구부서에 ‘우수’하지 않은 사원들도 포함하게 된다. 이와 반대로

		연구부서	
		우수	불량
생산부서	우수	0.5	0.5
	불량	0.5	0.5

		연구부서	
		우수	불량
판매부서	우수	0.5	0.5
	불량	0.5	0.5

연구부서의 근무사원중 '우수'로 평가된 사원으로 한정한다고 하면, 현재 타 부서에 근무하지만 '우수'한 사원들을 누락하게 되는 오류를 범하게 된다.

복합생성 자료의 검색오류는 본질적으로 자료 생성법칙의 이질성에 기인한다. 복합생성 자료검색의 특성 분석을 위하여, 먼저 자료 생성법칙의 이질성을 정형화하고 이를 이용하여 생성법칙의 이질성에 따른 자료검색의 오류를 분석한다.

법칙들간의 이질성이라 함은 동일한 속성치가 상이한 법칙에 의하여 자료화됨으로서 상이한 자료값으로 표현될 확률로서 정형화된다. 예를 들면 연구부서의 평가법칙과 생산 및 판매부서의 평가법칙간의 이질성이, 다음과 같이 조건부 확률로 표현될 수 있다고 가정하자.

여기서, 각 격자는 판매부서 또는 생산부서에 의하여 우수 또는 불량으로 평가된 사원이 연구부서에 의하여 우수 또는 불량으로 평가될 조건부 확률을 나타낸다.

조건부 확률분포표를 이용하여, 복합생성 자료검색의 오류는 각 검색질의의 Type I 오류와 Type II오류를 이들 조건부 확률로부터 유도함으로써 측정된다. 일 예로서 연구부서의 부서장이 <표 2>에 제시된 데이터베이스를 검색하여 연구부서에 적합한 우수 사원을 충원한다고 가정하자. 이 경우, 전통적 자료검색 구조하에서, 연구부서장의 검

색요구는 다음과 같은 검색질의로 표현된다.

검색질의-1: 업무수행능력 평가치가 '우수'로 기록된 모든 사원을 선발

검색질의-2: 업무수행능력 평가치가 '우수'로 기록되었으며 근무부서의 종류가 '연구부서'인 모든 사원을 선발

이들 검색질의에 대한 검색 오류는 업무평점 생성법칙간의 상이성이 상기 제시된 조건부 확률분포표와 같다고 가정할 경우 다음과 같이 유도된다⁶⁾. 이상에서 예시된 바와 같이 완전대응(exact matching) 자료검색모형하에서 복합생성자료의 검색은 어떠한 검색질의를 구성하더라도 TypeI오류와 TypeII오류를 수반하게 된다. 즉, 검색질의-2에서와 같이 충원조건을 평점이 '우수'로 기록되었으며 현 근무부서가 연구부서인 사원으로 제한할 경우, 현재 타 부서에 근무하고 있으나 연구부서에서 우수한 수행능력을 발휘할 수 있는 사원을 누락하는 제한점을 지닌다. 이에 반하여 검색질의-1과 같이 충원조건을 평점이 '우수'로 기록된 모든 사원으로 확대할 경우는 타 부서에서 비록 우수한 수행능력을 발휘하였으나 연구부서에서는 우수할 가능성이 낮은 사원을 충원하는 위험을 부담하게 되며, 또한 현재 타 부서에서 '불량'한 수행능력을 발휘

6) 여기서 $(e_2, e_3)/2$ 는, 평균적으로, e_2 와 e_3 의 두 사원중 한명이 해당 cell에 포함됨을 의미한다.

〈표 2〉 인력 데이터베이스

사 원	업 무 평 점	근 무 부 서
e1	우수	연구부서
e2	우수	생산부서
e3	우수	판매부서
e4	불량	연구부서
e5	불량	생산부서
e6	불량	판매부서

〈검색질의-1〉

	우수한 사원	우수하지 않은 사원	
충원된 사원	e1, (e2, e3)/2	(e2, e3)/2	e1, e2, e3
누락된 사원	(e5, e6)/2	e4, (e5, e6)/2	e5, e5, e4

〈검색질의-2〉

	우수한 사원	우수하지 않은 사원	
충원된 사원	e1	-	e1
누락된 사원	(e2, e3, e5, e6)/2	e4, (e2, e3, e5, e6)/2	e2, e3, e5, e6, e4

하였으나 연구부서에서 우수한 수행능력을 발휘할 수 있는 사원을 누락하는 제한점을 지닌다.

요약하면, 전통적 자료검색 구조에 의한 복합자료의 검색은 항상 TypeI 오류 또는 TypeII 오류를 수반하게 된다. 이러한 현상은 생성법칙들간의 이질성으로 인하여 자료값과 속성치가 일치하지 않으며, 이에 따라 검색된 자료와 찾고자 하는 속성치를 지닌 개체가 불일치하기 때문이다.

Ⅲ. 복합생성 자료검색 모형

복합생성자료검색의 어려움은 생성법칙의 상이성에 기인한다. 복합생성 자료의 의미는 자료값뿐만이 아니라 생성법칙에 의하여 결정된다. 따라서 복합생성자료검색의 오류를 제거하기 위하여서는 자료값과 더불어 자료의 생성특성을 동시에 결합하여 자료검색을 수행하여야 한다. 예를 들면, 사원의 업무수행능력은 부여된 평점과 더불어 사원

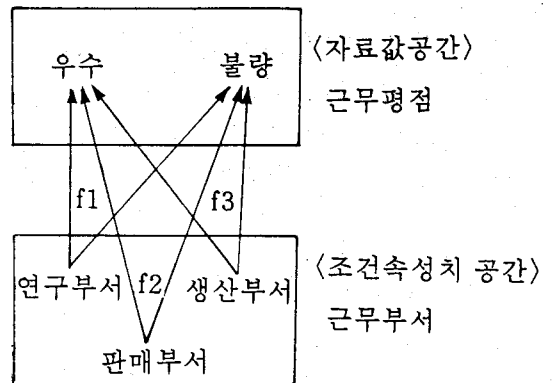
이 근무하는 근무부서의 평가기준을 동시에 고려하여 해석되어야 한다. 본장에서는 생성법칙의 복합성을 자료검색에 포함시키기 위한 기본 개념을 제시하고 이를 바탕으로 복합생성자료를 위한 검색모형을 제시한다.

1. 기본 개념

복합생성 자료검색을 위한 기본 개념을 제시하기 위하여 먼저 본 논문의 서두에서 예시된 업무수행능력의 평가 예를 고려하여 보자. 본 예는 연구부서를 위하여 우수한 업무수행 능력을 보유한 사원 검색을 목적으로 한다. 따라서, (이전 평가자의 평가가 올바르다는 가정하에서), 연구부서를 위한 우수한 사원검색을 위하여서는 자료값인 업무평점과 생성법칙인 평가기준을 동시에 고려하여야 한다. 자료의 생성법칙인 평가기준이 조건속성값인 근무부서에 의하여 한정됨을 고려할 때, 우수 사원의 검색공간은, (그림 3)에 제시된 바와 같이, 근무평점을 나타내는 자료값 공간과 근무부서를 나타내는 조건속성치 공간으로 구성된다. 즉, 복합생성자료의 검색공간은 자료값 공간과 조건속성치 공간으로 구성되며, 이들 공간은 복합생성 자료의 생성법칙에 의하여 연계된다.

연구부서를 위한 우수 사원의 검색 공간에서, 주어진 검색 목적을 가장 잘 충족시키는 집단은 현 근무부서가 연구부서인 동시에 업무수행능력이 '우수'로 평가된 사원들이다. 즉, 조건속성치 공간에 정의된 조건

속성을 만족시키면서 속성치 공간에 정의된 검색조건을 만족시키는 집단이다. 그러나 이와같이 한정된 검색은, 전장에서 제시된 복합생성 자료검색의 특성에서 살펴본 바와 같이, 우수한 수행능력을 발휘할 수 있는 사원을 누락하는 제한점을 지닌다. 이러한 현상에 기초하여 볼 때, 연구부서를 위한 우수 사원의 충원방안은 현 근무부서가 연구부서인 동시에 업무평점이 '우수'인 사원들을 가장 먼저 충원하고, 그 다음에는 타 부서에 근무하면서 연구부서에 근무하였을 경우 '우수' 평점을 받았을 가능성이 높은 사원들을 충원하는 방안이다. 이때 충원대상에서 반드시 제외되어야 할 사원들은 현



- f1 : 연구부서의 근무평점 생성법칙 (평가기준)
- f2 : 판매부서의 근무평점 생성법칙 (평가기준)
- f3 : 생산부서의 근무평점 생성법칙 (평가기준)

(그림 3) 연구부서를 위한 우수 사원의 검색 공간

근무부서가 연구부서이면서 평가치가 ‘불량’인 사원들이다. 그 이유는 현 근무부서가 연구부서일 경우 업무수행능력은 기존의 평가치에 의하여 명시된 바와 동일하기 때문이다.

상기 제시된 우수 사원 검색의 기본 개념을 정리하면 다음과 같다.

첫째, 충원조건을 가장 잘 만족시키는 집단은 현 근무부서가 연구부서인 동시에 업무평점이 ‘우수’인 사원들이다. 즉, 자료값의 검색조건과 조건속성의 조건을 모두 만족시키는 집단이다.

둘째, 충원 조건을 가장 만족시키지 않는 집단은 현 근무부서가 연구부서인 동시에 업무평점이 ‘불량’인 사원들이다. 즉, 조건속성의 조건은 만족시키나 자료값의 검색조건을 만족시키지 않는 집단이다.

셋째, 첫째와 둘째 조건에 포함되지 않는 집단은 검색조건을 부분적으로 만족시키는 집단이다. 바꾸어 이야기하면 이들은 조건속성의 조건을 만족시키지 않음으로서 검색요구의 충족도를 이분적으로 단언할 수 없는 집단을 칭한다.

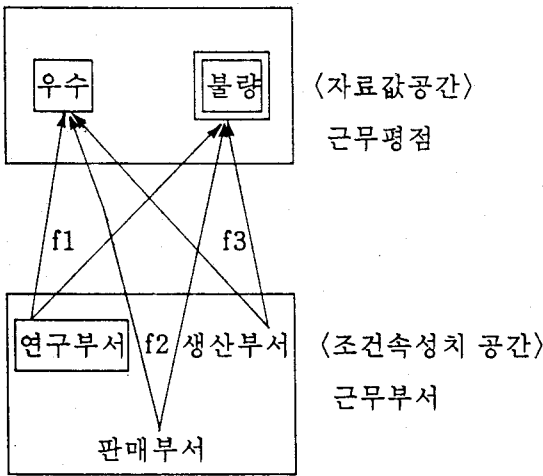
이상 제시된 복합생성 자료검색의 기본개념을 Lee[1990, 1992]에서 제시된 양기준 선호모형을 이용하여 일반화 하면 다음과 같다. 첫번째 집단인 충원조건을 가장 잘 만족시키는 집단은 복합생성자료검색의 정

근거점이다. 정근거점은 검색된 대안이 가능한 한 유사하기를(가까이 위치하기를) 바라는 상상 가능한 최적의 이상적 대안을 칭한다⁷⁾. 이에 반하여 충원조건을 가장 만족시키지 않는 집단은 복합생성자료검색의 부근거점이다. 부근거점은 가능한 한 유사하지 않기를(멀리 떨어져 위치하기를) 바라는 최악의 대안이다. 제시된 예에서, 정근거점과 부근거점은 모두 평가기준이 연구부서이어야 한다는 조건속성의 검색조건을 충족시키는 집단이다. 이중 정근거점은 업무수행능력의 평가치가 ‘우수’하여야 한다는 자료값의 검색조건을 충족시키는 집단이며, 부근거점은 자료값의 검색조건을 충족시키지 않는 집단이다. 따라서 이들 정근거점과 부근거점은 이분적 판단기준으로 명확히 식별될 수 있는 집단이다 (<그림 4> 참조).

이들 정근거점과 부근거점을 기준으로하여 판단하여 볼 때, 연구부서를 위한 우수 사원의 충원은 해당 사원들이 정근거점과 얼마나 유사하며 부근거점과 얼마나 상이하나의 기준에 의하여 평가될 수 있다. 즉, 정근거점과 부근거점에 포함되지 않는 검색조건을 부분적으로 만족시키는 집단은 이들이 정근거점에 상대적으로 얼마나 가까이 위치하는가에 의하여 검색질의의 부분적 충족도가 결정된다. 이를 양기준 선호모형에서는 상대거리(relative distance)로 측정

7) 정근거점은 Zeleny[1976]의 선택공리에서 제시된 근거점과 동일한 개념이다. 따라서 Lee[1990, 1992]에서 제시된 양기준선호모형은 Zeleny의 선택이론에 부근거점을 추가하여 확대 보완한 것으로 해석될 수 있다.

한다.



〈자료값공간〉

근무평점

〈조건속성치 공간〉

근무부서

□ : 정 근거점

□□ : 부 근거점

〈그림 4〉 복합생성 자료검색 공간에서의 정 근거점과 부 근거점

(연구부서를 위한 우수 사원의 검색 예)

상대거리란 특정 자료값이 정 근거점과 부 근거점을 기준으로 하여 상대적으로 정 근거점에 가까운 정도를 나타내며, 다음과 같이 측정된다. 정 근거점을 E^* 라 하고 부근거점을 E^o 로 하며, 특정 자료값의 정근거점과 부 근거점으로 부터의 거리를 각각 $d(e, E^*)$

와 $d(e, E^o)$ 로 표현하자. 이 경우 상대거리 $rd(e; E^*, E^o)$ 는 다음과 같이 표시된다.

$$rd(e; E^*, E^o) = d(e, E^*) / (d(e, E^*) + d(e, E^o))$$

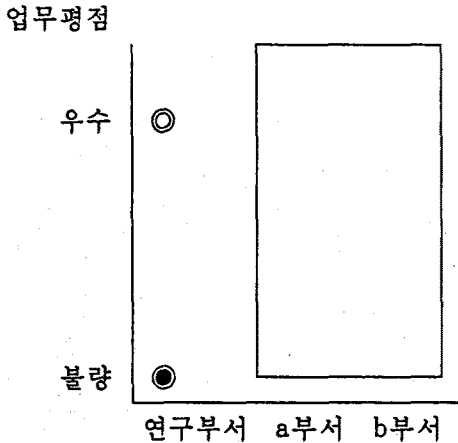
정 근거점과 부 근거점들로 부터 복합생성자료의 상대거리를 측정하기 위하여서는, 〈그림 4〉를 통하여 유추할 수 있듯이, 첫째, 자료값 영역 및 조건 속성 영역 각각에 대하여 거리함수를 정의할 수 있어야 하며⁸⁾; 둘째, 이들 거리를 자료값 공간과 조건속성공간을 통합하여 결합할 수 있어야 한다.

자료값 영역 및 조건 속성치 영역 자체에 대한 거리함수의 정의는 각 값들간의 유사성을 나타내는 함수의 정의란 점에서 영역의 특성에 따라 정의될 수 있다. 그러나 이들을 이용한, 복합생성 자료검색 공간에서 각 근거점들로 부터의 거리 측정은 다른 의미를 지닌다. 복합생성자료의 자료값과 조건속성치는 서로 다른 공간에 존재하며, 따라서 조건속성을 만족시키지 않는 자료값들의 정근거점과 부근거점으로 부터의 거리를 직접 측정함은 용이하지 않다. 본 연구는 이와 같은 이원적인 공간에서의 거리 측정을 위한 일 대안으로서 조건속성치를 자료

8) 본 연구에서 제시되는 복합생성자료검색의 기본 개념은 자료생성법칙간의 이질성을 조건 속성들간의 유사성을 통하여 간접적으로 조작함에 기초한다. 이와 같은 개념화는 조건 속성값들간의 유사성(또는 이질성)이 본질적으로 무엇을 의미하는지에 대한 객관적인 해석 기준을 제시하지 않는다는 점에서 한계성을 지닌다. 즉, 조건 속성값들간의 유사성(또는 이질성)은 자료생성법칙간의 유사성(또는 이질성)을 나타내기 위한 proxy에 불과하다는 한계점을 지닌다. 그러나 본 대안의 가장 큰 장점은 생성법칙간의 이질성을 조건속성치라는 데이터베이스에 저장된 자료값으로 측정함으로써 데이터베이스내 저장 자료값을 이용한 복합생성자료의 검색을 가능하게 한다는 점이다.

값 공간의 또 다른 차원으로 가정함으로써, 복합생성자료검색 공간에서의 거리를 확장된 자료값 공간에서의 유클리디언 거리로서 합성 유도한다.

이상에서 제시된 근거점들과 평가 기준을 일반화하여 자료값 공간상에서 기하학적으로 표시하면 <그림 5>와 같다. <그림 5>에 예시된 바와 같이, 복합생성자료의 검색공간은 자료값을 나타내는 축과 조건속성치를 나타내는 축으로 구성된다. 이들 조건속성치들과 자료값들을 상호 유사성에 따라 각각 x축과 y축에 배열하면, 복합생성자



- ◎ : 정근거점
(충원 조건을 가장 잘 만족시키는 집단)
- : 부근거점
(충원 조건을 가장 만족시키지 않는 집단)
- : 충원 조건을 부분적으로 만족시키는 집단

<그림 5> 복합생성자료 검색을 위한 기본 개념
(연구부서를 위한 우수 사원의 검색 예)

+료들은 2차원 공간상의 점으로 표현된다. 이와 같이 형성된 공간상에서 각 자료들이 정 근거점에 유사할수록 (또는 부 근거점과 상이할수록) 검색기준을 만족시킨다고 해석된다.

<그림 5>에서 제시된 바와 같이 연구부서를 위한 우수사원의 검색은, 현재 근무부서가 연구부서인 사원들의 경우, 업무평점이 '우수'인 사원은 가장 선호되어야 하며, 업무평점이 '불량'인 사원은 가장 기피되어야 한다. 도형을 통하여 유추할 수 있듯이, 연구부서가 아닌 타 부서에 근무하는 사원일 경우, 연구부서와 유사한 부서에서 근무하면서 업무평점이 '우수'인 사원이 상대적으로 정 근거점에 가까이 위치함으로써 선호될 것으로 판단된다.

2. 조건부 검색질의의 해석

본 절에서는 복합생성 자료검색을 위한 양기준 선호모형의 적용방안을 제시한다. 이를 위하여 먼저 복합생성자료의 검색기준을 명시하기 위한 질의 형태인 조건부 검색질의를 제시한다.

조건부 검색질의란 자료값의 검색조건만을 명시하는 기존의 전통적인 검색질의와 달리 자료값의 생성법칙을 결정하는 조건속성의 값도 함께 명시하는 검색질의를 말한다. 이를 위하여 본 연구는 전통적 접속어인 \neg (not), \wedge (and) 및 \vee (or)에 추가하여 조건부 접속어로서 $|$ (cond)을 제시한

다. 조건부 접속어는 조건 속성의 값에 대한 검색조건을 자료값의 검색조건과 분리하기 위하여 사용된다. 조건부 검색질의의 일례로서, 상기 제시된 연구부서를 위한 층원예를 고려하자. 본 검색요구는 다음과 같은 조건부 검색질의로 표현된다.

(업무평점='우수') | (근무부서='연구부서')

조건부 검색질의의 해석은 다음과 같이 2단계로 구성된다.

(1) 자료값에 대한 검색 조건의 해석

(2) 조건속성치에 대한 검색조건의 해석

여기서 전자는 전통적 검색질의의 해석과 동일하다. 즉, 자료값의 검색조건을 만족시키면 진(眞)의 값을 지니며, 만족시키지 않으면 부(否)의 값을 지닌다. 그러나 후자는 조건부 검색질의에 있어서만 존재하는 고유한 것으로서, 조건 속성값에 대응하는 자료 생성법칙간의 이질성을 평가함으로써 자료값의 해석 결과를 보완하는 단계이다.

자료생성법칙간의 이질성은, 전 절에서 기술된 바와 같이 조건 속성들간의 유사성으로 조작된다. 조건속성치간의 유사성은 정규 거리함수에 의하여 측정된다⁹⁾. 임의의 자료영역 $dom(A)$ 상에 정의된 거리함수 $d(v_i, v_j)$ 는 다음과 같이 정의된다.

$$(1) 0 \leq d(v_i, v_j) \leq 1,$$

$$(2) d(v_i, v_j) = 0 \text{ if } v_i = v_j,$$

$$(3) d(v_i, v_j) = d(v_j, v_i)$$

정규거리함수를 이용하여 조건부 검색질의 해석은 다음과 같이 정형화된다.

첫째, 자료값에 대한 검색 조건의 해석 결과 및 조건 속성간의 거리값을 이용하여, 데이터베이스에 존재할 수 있는 모든 자료실체의 속성치는 2차원 공간상의 점 $\langle x, y \rangle$ 로 표시된다. 여기서,

① y값은 자료값의 해석 결과를 나타낸다. 즉, 자료값이 검색조건을 만족시키면 1의 값을 지니며, 만족시키지 않으면 0의 값을 지닌다.

② x값은 조건속성간의 정규거리를 표시한다. 정규거리는 0부터 1까지의 값을 지니며, 조건 속성의 값이 동일한 경우에는 0의 값을 지니며, 매우 상이한 경우에는 1의 값을 지닌다.

이상의 표현방안에 의하여 실 현상에 존재하는 모든 가능 개체들은 아래의 2차원 공간상의 점 $\langle x, y \rangle$ 로 표시되다.

$$\{0,1\} \times [0,1]$$

둘째, 조건부 검색질의의 해석을 위하여 정근거점과 부근거점을 복합생성자료의 검색공간상에 좌표로 표현한다. 정근거점과 부근거점은 생성법칙들간의 상이성이 음의

9) 엄밀히 이야기하여 거리는 유사성의 반대 개념이다. 따라서 정규 거리함수에 의하여 측정된 거리 d 와 유사성 s 는, $s=1-d$,의 관계를 지닌다.

상관관계를 가질 수 있는 경우와 없는 경우에 따라 서로 달리 설정된다. 생성법칙의 상이성에 따른 근거점의 설정 방법 및 이에 기초한 조건부 검색질의 해석 방법은 이하 세부적으로 기술된다.

셋째, 정 근거점과 부 근거점을 이용하여 각 자료 실체들이 검색조건을 충족시키는 정도를 측정한다. 검색질의 충족도는 2차원 공간상에서 각 점으로 표시된 자료실체가 부 근거점을 기준하여 상대적으로 정 근거점에 가까운 정도를 측정함으로써 도출된다. 정 근거점을 E^* 라 하고 부근거점을 E^o 로 표현하자. 이때, 각 자료 실체의 근거점으로 부터의 거리는 각각 2차원 정규 공간상에서의 유클리디언 거리로 측정된다. 이들 거리를 각각 $d(e, E^*)$ 와 $d(e, E^o)$ 로 표현하면, $d(e, E^*)$ 와 $d(e, E^o)$ 은 다음과 같이 유도된다.

$$d(e, E^*)^2 = x^2 + (1-y)^2$$

$$d(e, E^o)^2 = x^2 + y^2$$

각 자료실체 d 가 조건부 검색질의, $g | h$, 를 만족시키는 정도는 2차원 공간상에서 부근거점을 기준하여 상대적으로 정근거점에 가까운 정도를 측정함으로써 도출된다. 이를 수식으로 표현하면 다음과 같다.

$$(g | h)(e) = d(e, E^o) / (d(e, E^*) + d(e, E^o))$$

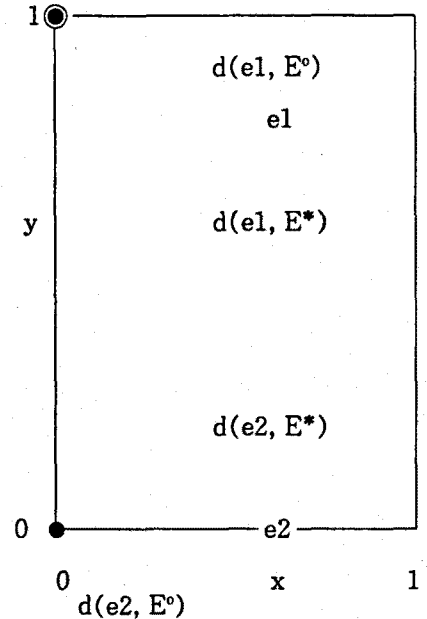
이상에서 제시된바와 같이 복합생성자료의

검색은 정 근거점과 부근거점이 설정되면, 이들에 대한 상대거리 측정으로 귀결된다.

2.1 근거점의 설정 및 상대거리의 측정 - 양의 상관관계

복합생성자료검색의 기본개념에서 제시된 바와 같이, 복합생성자료의 정 근거점은 자료값이 요구하는 생성법칙과 동일한 법칙에

자료값의 해석치



검색조건 유사성

⊙ : 정근거점 ● : 부근거점

$d(e1, E^o)$: 자료값이 e1의 정근거점으로부터의 거리

$d(e1, E^*)$: 자료값이 e1의 부근거점으로부터의 거리

$d(e2, E^o)$: 자료값이 e2의 정근거점으로부터의 거리

$d(e2, E^*)$: 자료값이 e2의 부근거점으로부터의 거리

<그림 6> 조건부 검색질의 해석

의하여 생성되었으며 또한 자료값이 검색조건을 만족시키는 실체들이며; 부근거점은 자료값이 요구하는 생성법칙과 동일한 법칙에 의하여 생성되었으나 자료값이 검색조건을 만족시키지 않는 실체들이다. 전자는 복합생성자료검색을 위한 2차원 공간상에서 $\langle 0, 1 \rangle$ 로 표시되며, 후자는 $\langle 0, 0 \rangle$ 로 표시된다. 이에 따라 복합생성자료검색의 2차원 공간에서 정근거점과 부근거점 및 복합생성자료 e_1 과 e_2 의 이들 근거점으로 부터의 거리는 〈그림 6〉에 도시된 바와 같이 측정된다.

〈그림 6〉은 조건부 검색질의의 해석을 도시한 것이다. 그림에서 제시된 바와 같이 조건부 검색질의는 각 자료값들을 조건속성치간의 거리 (x 좌표)와 자료값의 해석 (y 좌표) 결과를 이용하여 이차원 공간의 좌표로 표시하고 이들 2차원 좌표로 부터 근거점까지의 거리를 측정하여, 상대거리를 계산함으로써 해석된다.

그러나 이와 같은 근거점의 설정은 생성법칙들간의 유사성에 대한 양의 상관관계를 묵시적으로 가정한다. 여기서 양의 상관관계란 요구하는 생성법칙에 의하여 생성된 자료와 같은 의미를 지니는 자료값이 타 생성법칙에 의하여 생성될 수 없음을 말한다. 보다 구체적으로 설명하면, 복합생성자료검색을 위한 2차원 공간상에서 조건속성간의 거리와 자료생성법칙간의 유사성은 반대개념이다. 즉, 조건속성간의 거리가 0일 경우

는 두 생성법칙이 서로 동일한 경우로서 유사성이 가장 높은 경우이며, 조건 속성간의 거리가 1인 경우는 유사성이 가장 낮은 경우이다. 이 경우 만약 유사성이 가장 낮은 생성법칙들이 서로 상반된 자료값을 생성하는 법칙들로 정의될 경우, 정근거점과 부근거점은 위에서 제시된 바와 같이 조건속성을 만족시키는 자료값들만이 아니라 조건속성의 거리가 1인 서로 상반되는 생성법칙에 의하여 생성된 자료값 또한 포함하여야 한다. 묵시적인 양의 상관관계의 가정이란 이와 같은 경우가 발생할 수 없음을 의미한다¹⁰⁾.

전장에서 예시된 연구부서 요원의 총원예에서 연구부서와 생산부서간의 유사성이 0.5(거리가 0.5)이며, 연구부서와 판매부서간의 유사성이 0(거리가 1)이라고 하자. 이 경우 연구부서 총원을 위한 조건부 검색질의의 해석 결과는 〈표 3〉에 제시된 바와 같이 산출된다.

〈표 3〉에 제시된 각 사원들에 대한 해석 결과는 〈그림 7〉과 같이 도시된다. 여기서, $e_i \leftarrow e_j$ 또는 $e_i \rightarrow e_j$ 는 e_i 가 e_j 보다 검색질의를 더욱 잘 충족시킴을 의미한다.

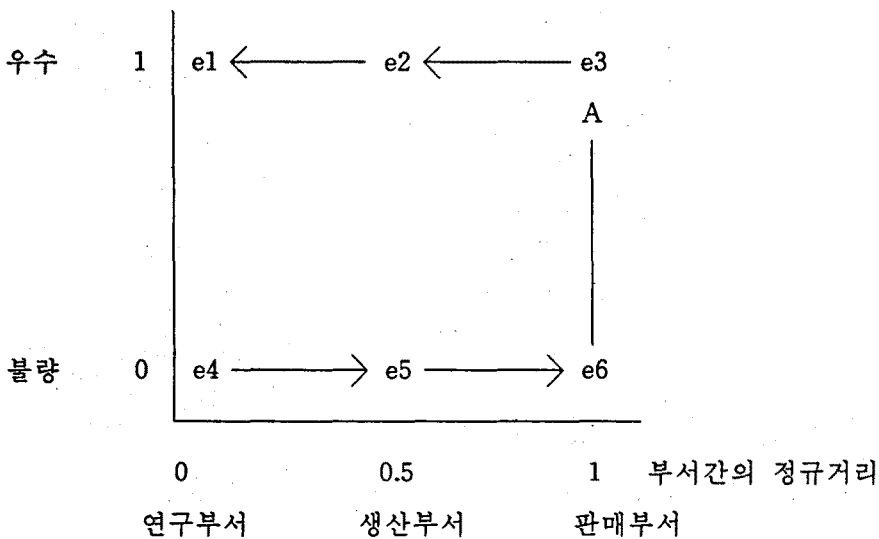
〈그림 7〉에 예시된 조건부 검색질의의 해석 결과는 일반적 예상과 일치한다. 즉, 특정 사원이 우수한 업무수행능력을 보유하는 것으로 평가 기록되어 있을 경우에는 평가부서의 종류가 연구부서과 유사할 수록

10) 복합생성법칙의 자료들이 음의 상관관계를 지니는 경우는 2.2절에서 제시된다.

〈표 3〉 조건부 검색질의 해석(양의 상관관계)

사 원	$\langle x, y \rangle$	$\langle d(e, E^*), d(e, E^0) \rangle$	$(g h)(e)$
e1	$\langle 0, 1 \rangle$	$\langle 0, 1 \rangle$	1
e2	$\langle 0.5, 1 \rangle$	$\langle 0.5, 1.118 \rangle$	0.691
e3	$\langle 1, 1 \rangle$	$\langle 1, 1.414 \rangle$	0.586
e4	$\langle 1, 1 \rangle$	$\langle 1, 0 \rangle$	0
e5	$\langle 0, 1 \rangle$	$\langle 1.118, 0.5 \rangle$	0.309
e6	$\langle 0.5, 1 \rangle$	$\langle 1.414, 0 \rangle$	0.414

업무평점의 해석치



〈그림 7〉 조건부 검색 질의 해석 결과의 도시화

(연구부서를 위한) 우수한 수행능력을 보유할 가능성이 높으며, 업무수행능력 평가치가 불량한 것으로 평가 기록되어 있을 경우에는 평가부서의 종류가 연구부서과 상이할 수록 우수한 수행능력을 보유할 가능성이 높을 것으로 예상된다.

이상에서 제시된 조건부 검색질의는 기하학적 특성상 각 자료실체가 조건부 검색질의를 충족시키는 상대적인 순서는 조건속성치간 상대적 유사성이 변하지않는 한 상대적 측정 단위에 대하여 불변적이라는 특성을 지닌다. 즉, 생산부서와 연구부서의 유사

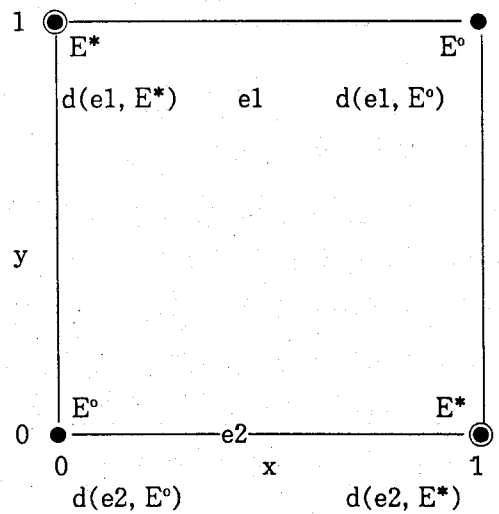
성이 0.5에서 0.2로 변하고, 판매부서와 연구부서의 유사성이 0에서 0.1로 변화하여도 생산부서에 근무하는 e2는 판매부서에 근무하는 e3보다 검색질의 충족도가 항상 높다. 바꾸어 이야기하면 연구부서에 대한 생산부서와 판매부서간의 유사성의 순서가 바뀌지 않는 한 이들 검색질의 해석 순서 또한 변화하지 않는다. 조건부 검색질의 해석이 지니는 이러한 특성은 복합생성 자료 검색시 측정단위가 지니는 자의성을 최소화한다. 즉, 조건속성간의 유사성 측정 단위 변화로 인한 검색 결과의 변동가능성을 최소화 한다.

2.2 근거점의 설정 및 상대거리의 측정 - 음의 상관관계

전절에서 제시된 조건부 검색질의 해석 모형은 서로 상이한 생성법칙에 의하여 부여된 자료값들이 적어도 음의 상관관계를 지니지 않음을 묵시적으로 가정하고 있다. 여기서 음의 상관관계를 지니지 않음이란, 복합생성자료의 검색요구를 완전히 충족시키는 집단이 오직 요구하는 생성법칙과 동일한 법칙에 의하여 생성되었으며 또한 자료값이 검색조건을 만족시키는 실체들과 같은 집단뿐이며, 이질적인 생성법칙에 의하여 생성된 자료값중에는 존재하지 않음을 의미한다. 그러나 상대 법칙과 자료값은 다르나 그 의미는 항상 같은 상호 정반대의 자료값을 생성하는 상호 이질적인 법칙들이 존재할 수 있다. 이 경우, 정근거점은 요구

하는 생성법칙과 동일한 법칙에 의하여 생성되었으며 또한 자료값이 검색조건을 만족시키는 실체들에 추가하여 가장 이질적인 생성법칙에 의하여 생성된 자료값중 검색조건을 만족시키지 않는 실체들 또한 포함하여야 한다.

자료값의 해석치



검색조건 유사성

● : 정근거점 ● : 부근거점

$d(e1, E^{\circ})$: 자료값 e1의 부근거점으로부터의 거리

$d(e1, E^*)$: 자료값 e1의 정근거점으로부터의 거리

$d(e2, E^{\circ})$: 자료값 e2의 부근거점으로부터의 거리

$d(e2, E^*)$: 자료값 e2의 정근거점으로부터의 거리

(그림 8) 조건부 검색질의 해석(음의 상관관계 포함)

〈표 4〉 조건부 검색질의의 해석(음의 상관관계)

사 원	$\langle x, y \rangle$	$\langle d(e, E^*), d(e, E^0) \rangle$	$(g h)(e)$
e1	$\langle 0, 1 \rangle$	$\langle 0, 1 \rangle$	1
e2	$\langle 0.5, 1 \rangle$	$\langle 0.5, 0.5 \rangle$	0.5
e3	$\langle 1, 1 \rangle$	$\langle 1, 0 \rangle$	0
e4	$\langle 0, 1 \rangle$	$\langle 1, 0 \rangle$	0
e5	$\langle 0.5, 1 \rangle$	$\langle 0.5, 0.5 \rangle$	0.5
e6	$\langle 1, 0 \rangle$	$\langle 0, 1 \rangle$	1

음의 상관관계의 가정하에서 복합생성자료의 검색요구를 완전히 충족시키는 집단은 요구하는 생성법칙과 동일한 법칙에 의하여 생성되었으며 또한 자료값이 검색조건을 만족시키는 실체들 뿐만 아니라, 이질적인 생성법칙에 의하여 생성되었으며 자료값이 검색조건을 만족시키지 않는 실체들로 구성된다. 이에 따라 복합생성자료검색의 정 근거점과 부근거점을 2차원 공간에서 좌표로 표시하면, 〈그림 8〉에 나타난 바와 같이 정 근거점을 $\langle 0, 1 \rangle$ 에 추가하여 $\langle 1, 0 \rangle$ 도 포함하도록 수정되어야 한다. 같은 논리하에서 부 근거점 또한 $\langle 0, 0 \rangle$ 에 추가하여 $\langle 1, 1 \rangle$ 을 포함하도록 수정된다.

연구부서를 우한 우수사원의 검색 예에서, 판매부서와 연구부서의 이질성이 매우 높아 판매부서의 불량한 사원이 항상 연구부서의 우수한 사원이 될 경우를 가정하여 보자. 이러한 생성법칙간의 음의 상관관계가 존재할 경우 연구부서 충원을 위한 조건부 검색질의의 해석 결과는 〈표 4〉에 제시

된 바와 같이 산출된다.

음의 상관관계를 지니는 복합생성자료의 검색에서 각 자료값의 근거점으로 부터의 거리는 최단거리로 정의된다. 즉, e2의 정근거점으로 부터의 거리 $d(e2, E^0)$ 는 〈그림 8〉에 제시된 바와 같이 $\langle 0, 0 \rangle$ 으로 부터의 거리로 측정되며, $d(e2, E^*)$ 는 $\langle 1, 0 \rangle$ 으로 부터의 거리로 측정된다. 즉, 임의의 자료값 e에 대하여 정근거점으로 부터의 거리 $d(e, E^*)$ 및 부근거점으로 부터의 거리 $d(e, E^0)$ 는 다음과 같이 수정 정의된다.

$$d(e, E^*)^2 = \min\{d(e, \langle 0, 1 \rangle), d(e, \langle 1, 0 \rangle)\}$$

$$= \min\{x^2 + (1-y)^2, (1-x)^2 + y^2\}$$

$$d(e, E^0)^2 = \min\{d(e, \langle 0, 0 \rangle), d(e, \langle 1, 1 \rangle)\}$$

$$= \min\{x^2 + y^2, (1-x)^2 + (1-y)^2\}$$

〈표 4〉의 해석 결과에서 제시된 바와 같이, 생성법칙간의 음의 상관관계를 가정할 경우, e1과 e6는 모두 정 근거점에 포함되며, e3와 e4는 부 근거점에 포함된다. 따라

서 이들 복합생성자료들은 동일한 검색질의 해석치를 지닌다.

2.3 근거점의 설정 및 상대거리의 측정—비이분적 자료값의 검색

본 연구에서 제시된 조건부 검색질의 해석은 자료값의 검색조건이 이분적임을 가정하고 있다. 그러나 많은 경우 자료값의 검색조건 또한 비이분적이다. 예를 들면 연구부서의 총원 예에서 업무수행능력의 평가치가 {우수, 양호, 보통, 불량}의 다단계 값을 지닐 경우 이들 자료값의 분류는 당연히 이분적이라기 보다는 '우수'를 가장 선호하

고 '불량'을 가장 선호하지 않는 비 이분적인 관계를 지닌다. 이러한 비 이분적인 자료값의 검색조건은 복합생성자료검색의 공간을 $\{0,1\} \times [0,1]$ 의 2차원 공간에서 $[0, 1] \times [0,1]$ 의 2차원 공간으로 확장함으로써 실현된다. 예를 들어 <표 2>의 인력 데이터베이스가 <표 4>와 같이 수정되었다고 가정하자. 이 경우, 자료값의 해석결과는 비 이분적 값을 갖도록 수정되어야 한다. 편의상 본 논문에서는 자료값이 검색조건을 만족시키면 (즉, 우수이면) 1의 값을 지니며, 완전히 만족시키지 않으면 (즉 불량이면 0), 중간값인 양호이면 0.66, 보통이면 0.33

<표 4> 인력 데이터베이스(수정)

사 원	업무수행능력 평가치	근 무 부 서
e1	우수	연구부서
e2	우수	생산부서
e3	우수	판매부서
e4	양호	연구부서
e5	양호	생산부서
e6	양호	판매부서
e7	보통	연구부서
e8	보통	생산부서
e9	보통	판매부서
e10	불량	연구부서
e11	불량	생산부서
e12	불량	판매부서

의 값을 지나도록 해석된다고 가정하자¹¹⁾. 그러면 연구부서를 위한 우수 인력 검색을 위한 조건부 검색질의의 해석은 <표 5>와 같이 유도된다.

<표 5>에서 자료값의 해석결과와 조건속 성치의 거리함수는 두번째 열에 제시된 바

와 같이 $\langle x, y \rangle$ 좌표로 제시된다. 따라서 조건부 검색질의의 해석은 이들 좌표와 근거 점들과의 거리를 이용하여 자료값의 해석결과가 0과 1의 이분적 값을 가질경우와 동일한 방법으로 해석된다.

<표 5> 조건부 검색질의의 해석(비 이분적 자료값의 해석)

사 원	$\langle x, y \rangle$	$\langle d(e, E^*), d(e, E^0) \rangle$	$(g h)(e)$
e1	$\langle 0, 1 \rangle$	$\langle 0, 1 \rangle$	1
e2	$\langle 0.5, 1 \rangle$	$\langle 0.5, 1.148 \rangle$	0.691
e3	$\langle 1, 1 \rangle$	$\langle 1, 1.414 \rangle$	0.586
e4	$\langle 0, 0.66 \rangle$	$\langle 0.34, 0.66 \rangle$	0.66
e5	$\langle 0.5, 0.66 \rangle$	$\langle 0.604, 0.836 \rangle$	0.581
e6	$\langle 1, 0.66 \rangle$	$\langle 1.056, 1.198 \rangle$	0.531
e7	$\langle 0, 0.33 \rangle$	$\langle 0.67, 0.33 \rangle$	0.33
e8	$\langle 0.5, 0.33 \rangle$	$\langle 0.836, 0.604 \rangle$	0.419
e9	$\langle 1, 0.33 \rangle$	$\langle 1.198, 1.056 \rangle$	0.469
e10	$\langle 0, 0 \rangle$	$\langle 1, 0 \rangle$	0
e11	$\langle 0.5, 0 \rangle$	$\langle 1.1180, 0.5 \rangle$	0.309
e12	$\langle 1, 0 \rangle$	$\langle 1.414, 1 \rangle$	0.414

VI. 결 론

본 연구는 복합생성 자료라는 새로운 자료 유형을 정의하였으며, 복합생성 자료의 특성을 충분히 반영하기 위한 새로운 검색

구조로서 복합생성 자료검색 구조를 설계, 제시하였다. 복합생성 자료의 검색요구는 조건부 검색질의에 의하여 정형화되며, 이는 전통적 검색질의와 달리 조건부 접속어를 포함한다. 조건부 검색질의의 해석은 양

11) 바꾸어 이야기하면, 자료값 '우수', '양호', '보통', '불량'간의 정규거리함수가 이들 자료값들을 각각 자연수 1, 2, 3, 4에 대응 표현함으로써 다음과 같이 정의되었다고 가정하자.

$$d(i, j) = |i - j| / 3$$

기준 선호모형에서 제시된 정 근거점과 부 근거점이라는 두 근거점에 기초하여 실현되며, 검색 결과는 직관적 기대와 일치하는 것으로 평가되었다. 이러한 연구 결과, 본 논문에서 제시된 복합생성 자료검색 구조는 현 전통적 검색구조가 지니는 문제점을 해결할 수 있는 일 방안으로 사료된다.

본 연구에서 제시된 복합생성 자료 검색 모형의 현실적 타당성은 제시된 조건부 검색질의의 해석결과에 대한 실증적 분석을 필요로 한다. 그러나 이에 대한 분석은 본 연구의 중점 사항으로 취급되지 않았으며, 따라서 추가적인 연구를 필요로 한다.

참 고 문 헌

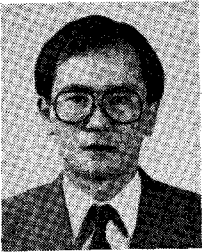
- Blair, D.C. "The Data-Document Distinction in Document Retrieval," *Communications of the ACM*, 27(4), April 1984, pp. 369-374.
- Lee, C.Y. *A Multi-criteria Data Retrieval Model: An Application of Multi-attribute Preference Model to Data Retrieval*, unpublished PH.D. Dissertation, University of Michigan, Ann Arbor, MI, 1990.
- Lee, C.Y. "Double Anchors Preference Model (DAPM): A Decision Model for Non-binary Data Retrieval (in Korean)," *The Journal of MIS Research*, 2(1), June 1992, PP. 3-16.
- Lipski, Jr., W. "On Semantic Issues Connected with Incomplete Information Databases", *ACM Transactions on Database Systems*, 4(3), September 1979, pp. 262-296.
- Mitroff, I.I., R.O. Mason, and V.P. Rarabba." Policy as Argument-A Logic for Ill-structured Decision Problems," *Management Science*, 28(2), December 1982. pp. 1391-1401.
- Mitroff, I.I. and Mason, R.O. "Can We Design Systems for Managing Messes? Or, Why So many Management Information Systems Are Uninformative?," *Accounting, Organization and Society*, 1983, pp. 195-203.
- van Rijsbergen, C.J. *Information Retrieval*, Butterworths, London, 1979.
- Wong, E. "A Statistical Approach to Incomplete Information in Database Systems," *ACM Transactions on Database Systems*, 7(3), September 1982, pp. 470-488.

Zadeh, L.A. "Fuzzy Sets," *Information and Control*, 8, January 1965, pp. 338-353.

tude Model (ADAM)," *Management Science*, 23(1), March 1976, pp. 12-26.

Zeleny, M. "The Attribute Dynamic Atti-

저 자 소 개



저자 이춘열은 현재 국민대학교 정보관리학과에 재직 중이다. 그는 서울대학교 산업공학과를 졸업하고, 동 대학원 경영학과에서 경영학 석사학위를 취득하였다. 또한 국방관리연구소 전산과학연구위원회 (현 국방정보체계연구소)에서 국방정보체계의 설계 및 분석, 평가에 참여하였다. 그 후 미시간 대학교 (앤아버, 미시간)에서 Computer and Information Systems를 전공하여 경영학 박사학위를 취득하였다. 그의 주요관심분야는 데이터베이스 응용, 자료검색, 정보공학등이며, 현재 데이터 모델링 및 정보공학적 기법을 응용 시스템에의 효과적 적용 방안에 대한 연구를 수행중이다.