韓國 OA學會 論文誌
第7卷 第4號, 2002. 12.

2002-7-4-1-12

참조 스키마 생성을 위한 개념적 스키마 분석

김 흥 수*

Conceptual Schema Analysis for Creation of Reference Sche

Heung-soo, Kim*

요 약

데이터베이스 설계를 위해 구축된 수많은 개념적 스키마들을 재사용하기 위한 분석수단이 요구된다. 본 논문에서는 개념적 스키마를 분석하는 방법을 제시하고 참조 스키마를 추출하기 위한 스키마 분석을 실험을 하였다. 스키마 통합은 유사함 값이 0.6 이상인 경우에 적용하는 것이 바람직하다. 분석 방법을 통해 생성되는 참조 스키마는 개념을 포괄적으로 표현할 수 있고 스키마 재사용을 위한 수단이 된다. 그리고, 참조 스키마 추출에 필요한 상세한 분석자료를 구하기 위해서는 피쳐를 근거로 스키마를 분석하는 것이 효과적인 수단이 된다.

Abstract

Large sets of conceptual schema have been constructed for database design. In recent times, the need of analytic aid for schema reuse is increasing. In this paper, it is presented analysis technique of conceptual schema, and experimented schema analysis for extraction of reference schema. It is desirable for integration of related schema to have been applied in case of similarity value above 0.6. Reference schema which is created through the analysis technique enable to describe concepts of them and can be the way of schema reuse. And a feature analysis can be effective measure to get details of analytic data which is necessary for extraction of reference schema.

* 동주대학 컴퓨터정보통신계열 부교수

논문접수 : 2002. 10.10
심사완료 : 2002. 12. 7

Ⅰ. 서론

데이터베이스 스키마를 분석하는 목적을 구분하면 첫째, 기존 시스템을 재구성하거나 구조 전환 시에 참조하는 정보를 추출하기 위하여 둘째, 재사용과 관련된 참조 스키마들의 라이브러리 구축을 위하여 셋째, 기존 데이터베이스 시스템 내의 정보 흐름과 데이터 중복을 확인하기 위해 넷째, 웹서버에 저장된 개념적 정보의 분석이나 서로 다른 시스템의 상호 동작 가능성을 점검하기 위하여 등이다. 데이터베이스 설계에서 개념적 모델은 적절한 역할을 하고 필수적이기 때문에 그 동안 수많은 개념적 스키마들이 구축되었으며, 이러한 스키마들을 재사용하기 위한 분석수단이 요구되고있다.[1] 따라서 개념적 스키마의 분석방법에 대한 연구가 필요하게 되었다.

따라서, 본 논문에서는 개념적 스키마를 분석하는 방법을 제시하고 8개 스키마를 대상으로 하여 참조 스키마를 추출하는 분석 실험을 하였다. 스키마 분석 방법을 통하여 생성되는 참조 스키마가 도메인의 주요 개념을 포괄적으로 표현할 수 있고 스키마 재사용을 위한 분석 수단으로써 도움이 될 수 있도록 과정을 설명한다. 즉 스키마의 디스크립터(descriptor)를 추출하여 추상적 표현으로 하고 스키마 유사함에 근거, 통합하여 참조 개념적 스키마를 생성시키는데 중점을 둔다. 본 논문은 다음과 같이 구성된다. 2장에서는 스키마 디스크립터의 추출과정을 실험하고, 3장에서는 스키마의 추상화를 설명하고 4장에서는 스키마 유사함에 대한 평가과정과 참조 스키마를 생성하는 실험을 하고 결과를 분석한다. 그리고, 5장에서는 본 논문에 대한 결론을 내리고 향후 연구에 대해 기술한다.

Ⅱ. 스키마 디스크립터 추출

2.1 용어 피쳐

스키마에 디스크립터를 관련시키는 색인을 하는데 디스크립터는 추상화, 비교 분석을 위해 내용과 주제를 의미론적으로 표현한다. 한 디스크립터는 가중치가 있는 피쳐(feature)들 리스트이며 [그림 1]과 같이 표현한다. 피쳐는 단일 혹은 한 쌍의 용어로 구성되며, 용어는 스키마 요소나 속성 명칭이다.

스키마 디스크립터를 구성하는 단일 용어 피쳐는 각 요소들의 기여도를 구해 평균기여도를 초과하는 대표 요소들을 선택하여 (대표요소, -)로 정하고 디스크립터에 추가한다. 요소 관련성은 E-R 스키마에 대해 고려한다.[2] 기여도는 요소 속성에 기여도를 0.9를 주고, 요소와 특수화 요소간에 강한 연결을 가지면 0.7을 주고, 그 외 연결은 기여도 값을 0.4로 한다.

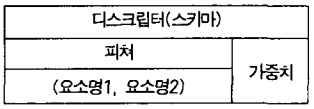

그림 1. 디스크립터의 표현
Fig 1. Representation of descriptor

$$\text{관련성 척도} = \sum_{j=1}^{n} (\text{기여도}) \times (\text{i번째 요소의 관계 링크 수})$$

[그림 2] 스키마1에 적용하여 관련성 척도를 구한다.

예를 들면

스키마1(고객)=속성 × 기여도+링크수 × 기여도

$$= 3 \times 0.9 + 3 \times 0.4 = 3.9$$

스키마1(고용인)=$3 \times 0.9 + 2 \times 0.4 = 3.5$

각 요소들에 대한 평균 기여도를 구하면 3.9임으로 평균기여도 이상인 요소만 디스크립터에 포함된다. 한 쌍의 용어 피쳐를 추출하는 방법은 대표요소에 대해 구조적 특성을 추출하여 (요소명, 구조적 특성) 피쳐와 대표요소에 링크되어 있는 요소들을 (요소명, 관련요소) 피쳐로 한다. 스키마1에 적용하여 추출한 2개 용어 피쳐들은 〈표 1〉에서 보인다.

2.2 피쳐 가중치

분석대상인 스키마들의 전체 디스크립터에서 피쳐의 관련성을 실수 가중치로 각 피쳐와 관련시킨다. repository의

뚜렷한 피쳐 집합 p을 대상으로 하여, 도메인을 처리하는 스키마들이 저장된 한 부분 즉 문맥에서 피쳐의 가중치를 구한다. 단일 문맥에서 j번째 피쳐의 가중치는 아래 식으로 구할 수 있다. 가중치를 [0, 1]범위의 실수 값으로 도출시키기 위하여 $\sqrt{s}$로 나누었고, s를 구하기 위해 피쳐 집합 p을 대상으로 하여 전체 피쳐 발생빈도와 각 피쳐를 포함하는 디스크립터 수를 조사한다.

$$\text{가중치} = \ln(\text{총 스키마수}/\ j\text{번째 피쳐가 있는 디스크립터 수})/\sqrt{s}$$

$$s = \sum_{i=1}^{p} (\text{i번째 피쳐 발생빈도})2 \times \ln(\text{총 스키마수}/\ \text{i번째 피쳐가 있는 디스크립터 수})2$$

스키마1 디스크립터의 가중치를 계산하기 위해서 8개 스키마를 실험 대상으로 하였으므로 N=8이며, 8개 스키마의 전체 피쳐 70개중에서 대표적인 피쳐는 52개이므로 p=52이다. 3개의 스키마는 주문과 상세히 관련되고, 제안은 2개 스키마와 관련되며, 공급자는 3개 스키마와 관련되며, 고용인은 2개 스키마에서 기술되고 있다.(고객,-),(주문,-),(제안,-),(공급자,-)피쳐들에 대한 가중치는 repository내의 고객, 제안, 주문, 공급자라는 용어 빈도인 0.943이고, (송장, -)피쳐는 가중치가 0.75인데 이유는 송장이란 용어는 빈도를 더 적게 사용하며 더 적은 수의 스키마와 관련하기 때문이다. (고객, 송장)피쳐는 송장이란 적은 빈도의 영향으로 가중치가 0.32 이며, 나머지 피쳐인 (송장, 금액),(주문, 송장)의 가중치는 0.25로 스키마1에서 작은 값을 가진다.

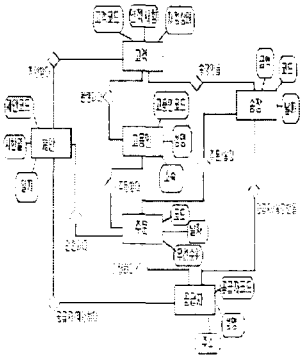

그림2. 주문 관리 스키마 1
Fig 2. Order management schema 1

표 1. 스키마 1의 디스크립터
Table 1. Descriptor of schema 1

	디스크립터(스키마1)	
	피쳐	가중치
단일 용어	(고객, -)	0.93
	(제안, -)	0.93
	(공급자, -)	0.93
	(주문, -)	0.93
	(송장, -)	0.75
한쌍 용어	(고객, 상태)	0.784
	(제안, 시한일)	0.32
	(공급자, 코드)	0.653
	(송장, 금액)	0.25
	(주문, 우선 순위)	0.784
	(공급자, 주문)	0.943
	(제안, 주문)	0.75
	(제안, 고객)	0.32
	(제안, 공급자)	0.653
	(고객, 송장)	0.32
	(주문, 송장)	0.25

Ⅲ. 스키마 추상화

3.1 클러스트(cluster) 기준

스키마 요소들의 클러스트에 근거해서 추상화하는 목적은 기본 개념을 표현하는 높은 응집을 가진 서버 스키마를 얻기 위함이다.[3] 서버 스키마의 높은 응집(cohesion)은 낮은 연결성(coupling)을 의미한다. 요소간의 친밀함(closeness)과 용어의 유사성(affinity)을 스키마 클러스트 기준으로써 사용한다.

스키마 내 요소간의 친밀함은 총 링크형태의 수 t 만큼 누적 함수로써 구할 수 있다.

$$\text{친밀함}(\text{요소1},\text{요소2}) = \sum_{i=1}^{t} (\text{기여도}) \times (\text{링크 수})$$

유사성 값은 용어의 관계 형태에 미리 정의된 값을 사용한다. 스키마 디스크립터들의 피쳐들에 사용된 용어를 비교하여 동일한 용어이면 1을, 동의어 관계이면 0.9, 광의나 협의의 용어 관계이면 0.6, 관련성 있는 용어이면 0.4를 유사성 값으로 사용하고 그 외 관련이 없으면 0을 사용한다.

연결성은 클러스터 내 요소의 유사성과 친밀함을 근거

로 다음 식으로써 평가한다. 가중치는 유사성에 0.4, 친밀함에 0.6을 주었는데, 이는 다소 주관적이며 스키마 구조의 의미에 따라 다를수 있다.

연결성(클러스터1,클러스터2)
={0.4(용어 유사성 값의 누적)+ 0.6 (요소 친밀함 값의 누적)}/총 클러스터요소 수

3.2 스키마 클러스트과 추상화

스키마의 클러스트은 단계적으로 다음 과정을 수행한다.

1) 대상 스키마의 n개 요소를 한 클러스터로 정한다
2) 요소들을 2개씩 연결 계수를 계산해서 행렬의 요소를 c_{ij}=연결성(c_i, c_j), c_{ii}=0 되게 연결 행렬을 채운다.
3) 초기 n개 요소들을 클러스터들로 그룹화 하는 동안에 다음을 반복한다.
   ① 높은 연결 계수를 가진 클러스터는 다른 클러스터 간에는 연결성이 약하므로 가장 연결 계수가 높은 클러스터들을 선택한다.
   ② c_i, c_j을 c_{i+j} 클러스터로 병합한다.
   ③ c_i, c_j 의 행 열 요소는 제거한다.
   ④ 새로운 클러스터 c_{i+j}의 연결 계수를 계산해서 요소를 추가한다.

요소의 집합과 링크의 집합으로 구성된 스키마에 클러스트와 추상화를 해서 대표요소 집합과 추상적 링크의 집합인 추상적 스키마를 정의하여 고수준으로 표현한다. 이러한 추상적 스키마는 적절히 선택된 추상적 요소를 가지는데, 추상적 요소의 유용성은 다른 스키마에 의해 공유된 요소들을 확인하고 스키마를 이해하고 검색하는 것을 용이하게 해준다. 추상적 스키마의 정의는 클러스터 스키마에서 시작하며 추상적 요소를 정의함은 대표적 요소를 구하는 것이고, 링크의 정의는 링크 재 구조화를 해야 한다. 재 구조화에서는 의미가 적은 링크는 제외시키며 포괄적으로 만들어 재 명명한다. 특수화 링크관계는 cycle을 허용하지 않으나 추상적 요소간 링크로서 클러스터 스키마 요소들 간의 원래의 링크는 추상적 스키마 내에 cycle이 존재 할 수 있다. 이런 경우 적어도 한 개의 링크를 탈락하거나 재 명명함은 필요하다.

링크 추상화는 개념적으로 관련된 링크 그룹을 한 개 링크로 변형하는 과정이므로 기준은 이름이 유사하거나 클러스터 스키마 내의 스키마 집단간에 같은 경우 적용한다. 추상적 링크의 정의는 원래 링크의 이름에 주로 근거한다.

Ⅳ. 스키마 유사함 평가

4.1 디스크립터의 피쳐

스키마의 유사함(similarity)은 디스크립터의 피쳐를 근거로 유사성 기준을 적용해서 평가한다. 유사성을 나타내는 피쳐 수가 많을수록 스키마간에 유사함은 더욱 높아진다. 유사성 값은 단일 용어 피쳐인 경우 용어간의 유사성 값이 되고, 2개 용어 피쳐이면 용어간 유사성 평균값을 구하여 적용한다. 유사함 계수는 유사함(스키마1, 스키마2)로 표시할 수 있으며 스키마의 유사함 계수는 유사성값 합계함수와 Dice함수를 각각 독립적으로 적용해서 구한다.

(1) 피쳐 유사성 합계함수를 적용한다.
유사함(스키마1, 스키마2)= 2 ×(피쳐간 유사성 값의 합계)/ (스키마의 총 피쳐수)

(2) Dice함수를 적용한다.
유사함(스키마1, 스키마2)= 2 ×(유사성 값이 평균이상인 피쳐수)/ (스키마의 총 피쳐수)

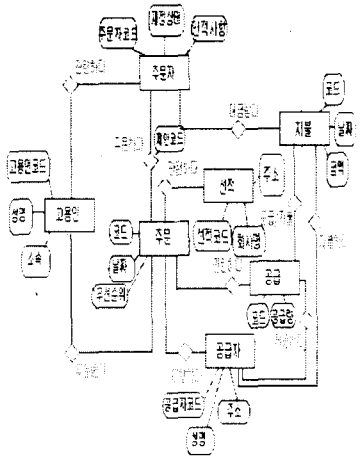

그림3. 주문 관리 스키마 2
Fig 3. Order management schema 2

예를 들면, 스키마1과 스키마2의 유사함을 조사하기 위해 가중치를 가진 피쳐들 리스트로 디스크립터를 표현하였다. 스키마1과 스키마2의 피쳐 수는 각각 16, 12개로 구성되어진다. 가중치를 계산하기 위해 스키마의 수는 N=8을 고려했는데, 8개 스키마의 대표적인 피쳐 수는 p=52이다. 주문은 3개 스키마와 관련되며, repository 3개의 스키마는 공급자와 관련되고, 지불은 2개 스키마와 관련되며, 고용인도 2개 스키마와 관련되고 있다.

스키마의 유사함을 평가하기 위해서는 먼저 피쳐들의 용어들을 비교해서 피쳐 유사성 값을 구해야 된다. 예를 들어 스키마1의 피쳐인 (송장, 금액)과 스키마2의 피쳐인 (지불, 금액)에 대한 유사성 값을 구하는 과정을 보면

유사성(송장, 지불) = 0.4

유사성(금액, 금액) = 1 임으로 유사성 평균값 0.7이다. 유사함 계수는 유사함(스키마1, 스키마2)으로 표시할 수 있으며 계산함수는 피쳐 유사성 합계함수와 Dice함수를 각각 적용해서 구했다.

* 유사성 값 합계함수를 적용하면

유사함(스키마1, 스키마2) = 2 × 8.45/ (12+16) = 0.6

* Dice함수를 적용하면

유사함(스키마1, 스키마2) = 2 × 9/ (12+16) = 0.64

Dice함수를 적용한 경우 더 큰 결과 값을 얻었다.

표 2. 스키마 2의 디스크립터
Table 2. Descriptor of schema 2

	디스크립터(스키마2)	
	피쳐	가중치
단일 용어	(주문자, -)	0.93
	(지불, -)	0.75
	(공급자, -)	0.93
	(주문, -)	0.93
한 쌍 용어	(주문자, 상태)	0.784
	(주문, 우선 순위)	0.784
	(지불, 금액)	0.653
	(공급자, 코드)	0.8
	(주문자, 지불)	0.402
	(주문, 공급자)	0.78
	(지불, 공급자)	0.65
	(주문, 주문자)	0.8

4.2 실험과 결과 분석

스키마들을 다중 정의하거나 부분적으로 겹치는 것을 피하고 한 스키마로 표현되는 참조 스키마를 추출하기 위해서 다른 추상적 스키마 내의 유사한 요소들을 식별 분석해서 통합하였다. 유사한 추상적 요소를 식별하기 위해서 유사함 기법을 사용하여 추상적 스키마들과 관련하고 있는 클러스터를 분석하고 유사한 개념을 다르게 표현한 스키마를 대상으로 하였다. 유사함 값이 0.6이상이 되어야 통합하는 것이 바람직하다고 판단하여 수행하였다.

[그림 2]와 [그림 3]에 있는 스키마를 겹쳐서 [그림 4] 통합한 참조 스키마를 생성하는 과정을 보면, 스키마1과 스키마2 모두 고객이라는 추상적 요소와 관련하는 클러스터가 있으므로 고객이라는 개념을 표현하는 단일 참조 스키마로 통합되어진다. 스키마1과 스키마2 의 모든 클러스터에 있는 요소들의 속성과 링크를 모두 고려하였고 용어표현이 일치하지 않는 문제를 해결하는데는 예를 들면, 송장과 지불이라는 요소명칭이 각 스키마에서 동일한 요소 형태의 서로 다른 표현이므로 지불이라는 명칭으로 선정하였다. 스키마1에는 고객과 주문요소간에 제안요소가 있고 스키마2에는 주문자와 주문요소간에 주문한다는 관계에 속성으로 제안코드가 있다. 그러므로 어떤 주문자로부터 받은 주문은 그 주문자가 받은 제안과 관련되기 때문에 스키마1의 고객과 제안이란 관계를 통합된 참조스키마에 포함시키기 위해 주문자와 주문 요소간의 제안이라는 요소를 통합하고 고객과 주문자 명칭 중에서 주문자 명칭으로 결정하였다.

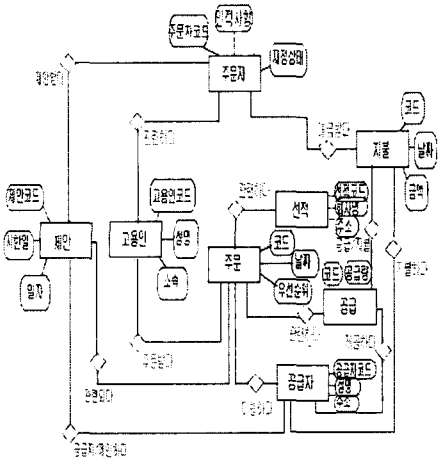

그림 4. 통합된 참조 스키마
Fig 4. Reference schema after integration

V. 결론

본 연구에서는 개념적 스키마를 분석하는 방법을 제시하고 참조 스키마를 추출하기 위해서 8개 스키마를 실험 대상으로 하여 스키마를 분석하는 실험을 하였다. 분석 방법을 통해 생성되는 참조 스키마는 개념을 포괄적으로 표현하고 스키마 재사용을 위한 분석 수단으로써 충분히 도움이 될 수 있다. 디스크립터 추출에서는 스키마 색인 작업과 대표적 피쳐들로 정제시킨 다음 모든 스키마의 디스크립터 관점에서 피쳐 관련성을 가중치로 주었다. 그리고 클러스터 기준에 따라 스키마를 추상화하였으며, 친밀함과 유사성 가중치는 의미의 중요성에 따라 변할 수 있다. 스키마의 유사함 계수는 Dice함수보다 피쳐 유사성 합계함수가 더 작은 값을 주었으며, 평균 이상인 피쳐의 수를 대입하는 부분과 비교하면 피쳐 유사성 합계가 더 결과 값이 정확하다고 본다. 그리고, 유사한 개념을 다르게 표현한 스키마를 대상으로 하여 유사함 값이 0.6 이상은 되어야 스키마 통합이 바람직하다고 본다. 연구 결과 피쳐를 근거로 스키마를 분석한다면 상세한 분석자료를 얻을 수 있고 자료를 바탕으로 참조 개념적 스키마를 생성할 수 있으므로 분석을 필요로 하는 환경에서 효과적인 수단이 된다고 본다. 그리고 향후 스키마 분석과정을 위한 도구개발, 스키마 라이브러리 구축 등의 자동화 연구가 진행되기를 기대한다.

참고 문헌

[1] Bellizona,R. ,Fugini,M.G., "Reusing specipications in OO application" IEEE Softw. 12,2(March) 1995.

[2] Batini,C., Di Battista, "Structuring primitives for a dictionary of entity relationship data schemas. IEEE Trans. Softw. Eng. 19,4(Apr.), 1993.

[3] Feldman, P., Miller, D., "Entity Model clustering: Structuring a data model by abstraction. Comput. J. 29,4, 1986.

[4] Sheth, A, P., Navathe, S. B., "On automatic reasoning for schema integration" Int. J. Intell. Coop. Inf. Syst. 2,1(june), 1993.

[5] Jarke, M. et AL. "DAIDA: An environment for evolving information systems" ACM Trans. Inf. Syst. 10,1(Jan.) 1992.

[6] Chen, M,. Han, J., Yu, "Data mining: An overview from a database perspective", IEEE, Dec.,1996.

[7] Elmasri, R., Navathe, S., "Fundamentals of database system" third edition, 2000.

저자소개

김 홍 수

1979 울산대학교 전자계산학과 공학사

1984 숭실대학교 전자계산학과 공학석사

1998 경상대학교 전자계산학과 박사수료

현재 동주대학 컴퓨터정보통신계열 부교수

연구분야 :

데이터베이스관리, CAD

멀티미디어 콘텐츠