

□ 기술해설 □

의미구조 기반의 자기조직형 정보베이스

한국문화예술진흥원 이원규* · 김명철* · 김성훈** · 강윤희** · 김해수** · 김문호

● 목

차 ●

- 1. 서 론
 - 1.1 정보의 특성
 - 1.2 관계와 표현
- 2. HINS
 - 2.1 시스템의 요건
 - 2.2 시스템의 구성
- 3. 정보검색
 - 3.1 HINS의 정보검색
 - 3.2 HAIR

- 4. 의미구조 모델
 - 4.1 확장된 하이퍼그래프
 - 4.2 의미정보 표현
 - 4.3 의미구조 관리시스템
- 5. 개념구조의 생성
 - 5.1 개념구조로의 시소러스
 - 5.2 문제점분석
 - 5.3 시소러스 자동구조화
- 6. 결 론

1. 서 론

1.1 정보의 특성

모든 정보처리 시스템은 일반적으로 정확한 대량의 데이터(Data)를 신속하게 처리·가공하여 정보(Information)의 형태로 이용자에게 제공한다는 기본 구조를 갖는다. 데이터란 인위적으로 창출된 원시자료이고, 정보란 원시자료가 부가기능에 의해 처리되어 이용자의 요구(Needs)에 맞도록 가공된 결과물이라 하겠다. 생성되는 방법에 따라 정보를 분류하면 다음과 같다.

1차 정보 : 창조력을 바탕으로 인간에 의해 생성된 결과물

예) 문서, 그림, 영상, 음악, 악보, 조각 등

2차 정보 : 1차정보간의 관계에 대한 조합으로 인간에 의해 생성된 순차적 결과물

예) 백과사전, 하이퍼미디어 생산물, 서지목록 등

3차 정보 : 1, 2차정보에 누락된 혹은 잘못된 관계를 학습[44], 유추, 추론 등의 고급기능에 의해 기계적으로 발견·재구성된 결과물

4차 정보 : 1, 2, 3차 정보를 근거로 기계에 의해 창출된 결과물. 즉, 인간에 의해서만 창출되던 1차 정보의 생성

위에 분류된 정보의 형태는 정보생성의 사이클을 이루며(그림 1-1), 관리시스템 및 고급기능[1,2]의 개발을 수반한다. 특히 1차 정보는 기술의 발달에 따라 텍스트 정보에 국한되지 않고 멀티미디어 데이터를 포함한다. 멀티미디어 데이터를 수용할 수 있다는 것은 다양한 형태의 데이터를 취급할 수 있다는 것 외에도, **정보표현수준(Description Level)의 향상과 함께 정보상징화의 단계를 줄여 정보전달의 정확도를 높인다.** 예를 들어, 물체를 지칭하는 용어로 실체에 대한 개념을 전달할 뿐 아니라, 2차원의 사진이나

*정회원

**준회원

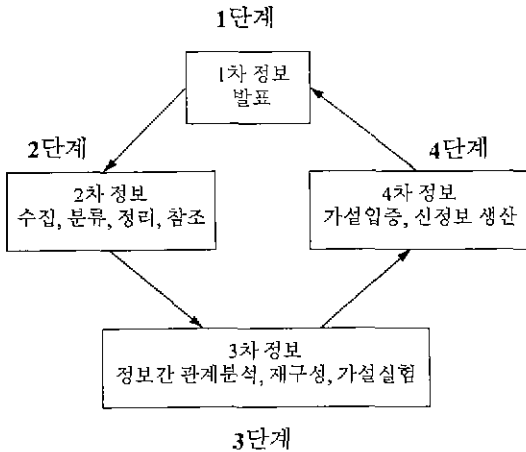


그림 1-1 정보생성 사이클

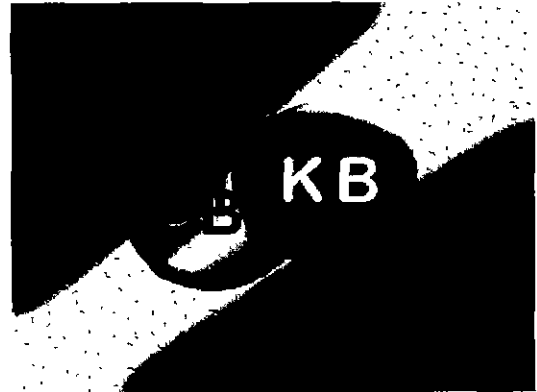


그림 1-2 실세계의 정보공간과 도메인

설계도로 표현하여 개념의 이해를 돕는다는 것이다.

이와같은 정보의 분류 및 생성과정은 실세계의 정보에 대한 면밀한 분석과 적합한 모델설정을 전제로 한다. 실세계의 정보공간은 시·공간에 따라 유동적이며 열린 공간(Open Space)이다. 유동적이란 관점이나 목적에 따라 설정되는 도메인(Domain)이 변형된다는 것을 암시하며, 열린 공간이란 새로운 정보의 추가에 따라 부가되는 관계정보가 무수히 많다는 것을 의미한다[3]. 그림 1-2는 이러한 정보공간에 관한 개념을 추상적으로 표현한 것이다. 정보공간 내에 무수히 많은 정보입자들이 존재한다고 할 때, 각종 시스템의 대상 데이터의 도메인은 빛줄기 속에 포함된 정보입자들에 비유될 수 있다.

이상에서 나타난 정보의 특성을 정리하면 다음과 같다.

- 중복성(Overlapping)
- 상대성(Relativity)
- 재귀성(Recursion)

정보공간에서의 거리는 수학에서의 마찬가지로 두 입자간의 최단 길이이다. 그러나 정보입자가 유동적이므로 정보입자간의 거리는 복수개 존재할 수 있다. 따라서 정보공간은 3차원공간이 아니라, n 차원의 유클리드 공간으로 정의된다[4].

1.2 관계(Relationship)와 표현(Representation)

정보의 특성은 시스템 구축에 있어서 문제점으로 대두되기도 하지만, 정보간의 정확한 관계 정의에 의해 해결될 수 있다. 대부분의 정보모델 혹은 데이터모델은 정보와 그들간의 관계표현 및 관리방법에 초점을 맞추고 고안되어 왔다. 초기에는 관계형 모델(Relational Model)[1], 네트워크 모델(Network Model), 계층 모델(Hierarchical Model) 등의 모델이 있었다. 특히 관계형 모델은 확장 관계형 모델(Extended Relational Model) [5]로 확장되거나 변형되어 상용화에 이르렀다. 단위정보의 내부구조(Internal Structure)를 표현하고자 내포된 관계형 모델(Nested Relational Model)이 고안되었다. 또한, 규칙(Rule)에 의해 관계정보를 구체화시키는 규칙기반 데이터 모델(Rule-Based Data Model) 등이 등장했다. 그 외에도 지식베이스 시스템(Knowledge Base System), 전문가시스템(Expert System) 등이 있었으나, 여전히 정보표현의 제약, 정보의 특성반영 부족 등의 문제가 남아 있다. 객체지향형 데이터 모델(Object Oriented Data Model)[6-15]은 비교적 유연한 정보표현이 가능하지만 정보의 특성을 충분히 반영하지 못한다. 텍스트정보에 국한시켜 고려하더라도 이들 모델은 제한된 영역(Closed Space)의 데이터만을 가공하여 처리함으로써 실세계의 정보에 대해 의미적인 처리를 할 수 없다는 단점을 갖는다. 실세계의 정보 처

리를 위해서는 정보표현의 융통성을 수용하고 정보의 의미를 분석하여 정보의 특성을 반영할 수 있는 정보모델이 필요하다.

이상에서 알 수 있듯이 실세계의 정보를 수용할 정보모델은 단위 정보 및 정보간의 관계에 대한 분석이 필수적이다. 하드웨어의 발달에 따라 다양한 형태의 정보간의 유기적관계를 중심으로 생성되는 방대한 정보량을 활용할 수 있다. 그러나 학습, 유추, 추론, 문제해결 등의 고급기능은 분야별 전문지식을 보유하고 있는 전문가에 의존하고, 혼재되어 있는 전문분야간의 명확한 관계정의의 부재로 활용되지 못한다. 따라서 대량정보 처리를 전제로한 정보간의 관계정의에 있어서의 문제점은 다음의 두 가지로 요약될 수 있다.

1. 분야별 지식간의 정확한 관계정의
2. 정보공간에서의 지식간의 관계정의

다시 말하면 전자를 「구분(Identification)」, 후자를 「통합(Classification)」의 문제라 하겠다. 이것은 앞에서도 언급한 정보의 특성에 기인한다.

관계정보의 처리는 구조화된 정보의 처리라 할 수 있다. 그림 1-3은 단위정보와 정보간의 관계를 도식적으로 표현한 것이다. 단위정보는 그들간의 관계정의로써 또는 관계정의 내의 구성요소로서 작용한다. 정보공간내의 단위정보가 유동적이란 것은 실세계의 정보가 이러한 정보 특성을 반영하고 있기 때문이다. 언어로 표현되는 지식에 있어서 용어의 다의성(多義性) 및 표현의 다양성(多樣性)은 용어의 상징화에 따라 용어의 표현 및 용어에 의해 구조화된 관계정의에 있어서 역할을 달리하기 때문이다. 구조정보를 처리하기 위해서는 관계의 역할표현, 관계조작의 시스템화, 동적인 구조표현 등의 문제로 요약될 수 있다.

이상에서 논의된 사실을 정리하면 다음과 같다.

1. 실세계의 정보공간은 방대하다.
2. 전문가에 의해 생산되는 유연한 정보표현 방법이 요구된다.
3. 정보의 특성을 반영한 정보모델이 필요하다.

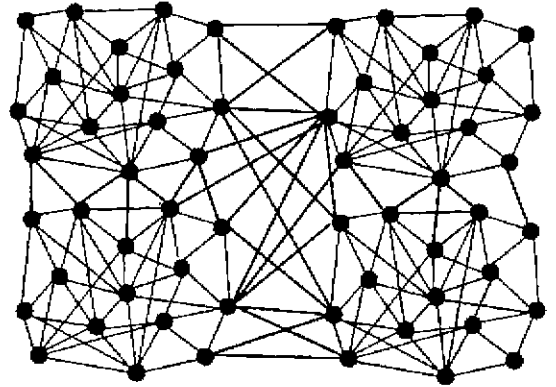


그림 1-3 단위정보와 관계

4. 관계정의에 따른 의미의 구조화가 필요하다.
5. DB, KB를 포함하는 정보는 고급기능에 의해 자기조직화(Self-Organizing) 되어야 한다.

제 2장에서는 기존의 정보처리기술을 응용해 구축한 시스템 HINS에 대한 설명과 문제점 등에 대해 논의하고, 제 3장에서는 HINS의 정보검색 방법 및 문제점, 시스템의 진화에 따라 정보검색엔진이 갖추어야 할 기능과 그 구성 및 문제점에 대해 논의한다. 제 4장에서는 확장 하이퍼그래프 형태로 고안된 의미구조기반의 정보모델과 유연성 있는 정보표현 방법, 의미구조 관리 시스템의 구성을 제시한다. 제 5장에서는 고급기능에 의해 자기조직화될 의미구조의 구축에 중요한 역할을 할 개념구조의 용이한 구축방법에 대해 설명하고, 제 6장에서는 결론 및 제반 문제점, 향후계획 등에 대해 기술한다.

2. HINS

문화체육부와 한국문화예술진흥원이 협력하여 추진중에 있는 문화정보망 구축사업의 일환으로 개발되고 있는 정보망운영의 모델시스템 HINS(Hypermedia Information Network System)는 1993년 1월부터 문화·예술분야의 데이터를 중심으로 설계·구현되고 있다. 문화정보망은 초고속 정보통신망 구축사업, 문화체육부와 과기처의 협력하에 설립된 국어공학센터의 한글정보처리를 위한 기반사업, CATV와 관련된 각종 영

표 2-1 문화정보망 구축의 단계별, 연도별 계획(안)

단계별목표	문화정보망	연도별목표	문화에세데이터베이스
제 1 단계 (준비기)	기반 환경 구축 시험 시스템 개발 추진 체제 정비 법적, 제도적 정비 관련단체 지원 정보활용 마인드 형성 정보 분석 및 확보	'93년도(준비기)	연구개발 조직 구성 환경정비 중장기 사업 설계 기본 연구 용역 수행 요소기술 습득
		'94년도(개발기)	모델 시스템 개발 관련기관 기술 자문 및 지원 연구 및 구현 체제 확립
		'95년도(지원기)	관련기관 지원확대 및 홍보 문화정보망에서의 역할 정의 연구용역 축소 및 자체연구 확대 문예진흥기금 확보 방안 모색
제 2 단계 (지원기)	세부정보망 활성화 분야별 기술의 고도화 문화정보 보급의 활성화 "문화정보센터" 정립 세부정보망 조직화 국제화, 세계화 정보 구조화	'96년도	
		'97년도	
		'98년도	
제 3 단계 (자립기)	문화과학 정착기 문화정보망 확대 초고속 통신망과 접속 준비 고도의 안방 문예 어카이브 "문화정보센터" 안정화 기초연구 체계 돌입 정보 조직화	'99년도	
		'00년도	
		'01년도	
		'02년도	

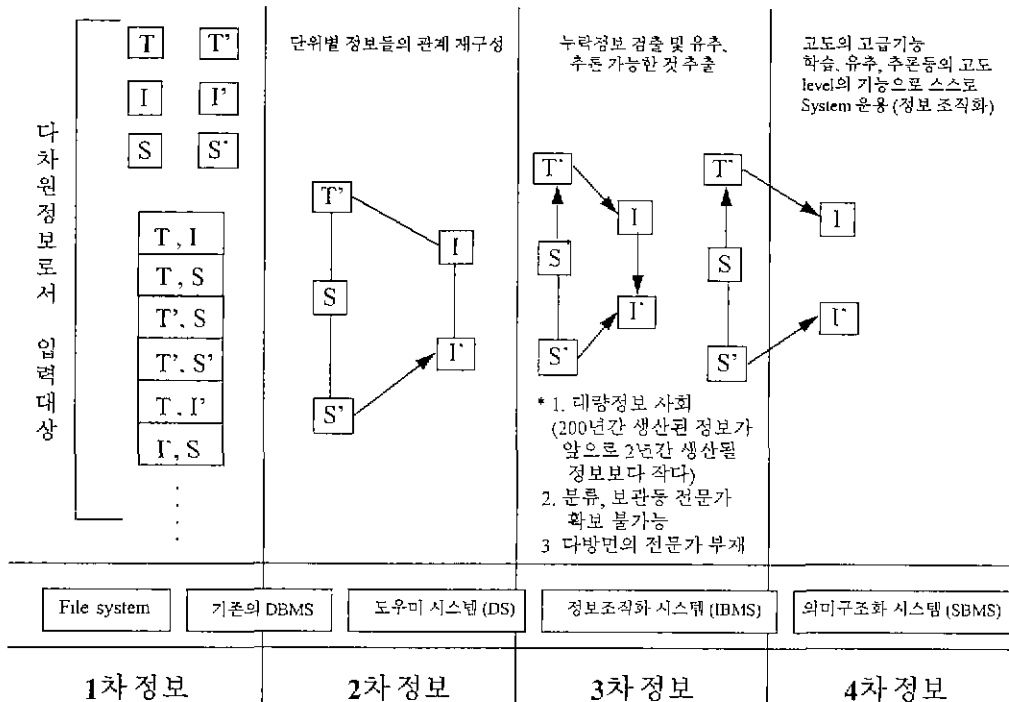


그림 2-1 정보생성에 의한 분류와 관리시스템의 관계

상사업, 향후 정보통신부의 역할 등과 연계된 일련의 국가 정보망 구축계획에 있어서 중요한 역할을 담당한다[16]. 개괄적인 향후 추진계획안은 표 2-1과 같다.

2.1 시스템의 요건

시스템 구축의 기본방침은 사용자 중심의 정보시스템(User Oriented Information System)의 구현이다. 따라서 다음과 같은 기본 요건이 만족되어야 한다.

1. 멀티미디어 데이터를 수용해야 한다.
2. 연산, 검색, 추론 등의 어플리케이션이 일체화되어야 한다.
3. 대량정보의 통신을 전제로 해야 한다.
4. 실세계 정보를 수용할 진화적인 모델이 제시되어야 한다.

다시 말해서 다양한 형태의 데이터, 정보, 지

식이 고도의 기능에 의해 통합운영되며, 정보모델의 발전에 따라 능동적으로 전환되어야 한다. 1.1절에서 논의된 정보의 분류와 모델의 진화에 따른 관리시스템의 발전, 단계별계획을 정리하면 그림 2-1과 같다.

2.2 시스템의 구성

시스템 구성을 위한 연구·개발부문은 크게 7가지로 구분된다(그림 2-2).

1. 사용자습성에 관한 조사 및 UI(User Interface)의 개발
2. 하이퍼미디어 엔진의 개발
3. 통합 스키마의 설계에 관한 연구
4. 관리시스템의 개발
5. 원시데이터의 축적 및 가공 지원기술 개발
6. 자기조직형 진화모델을 위한 지식축적에 관한 연구

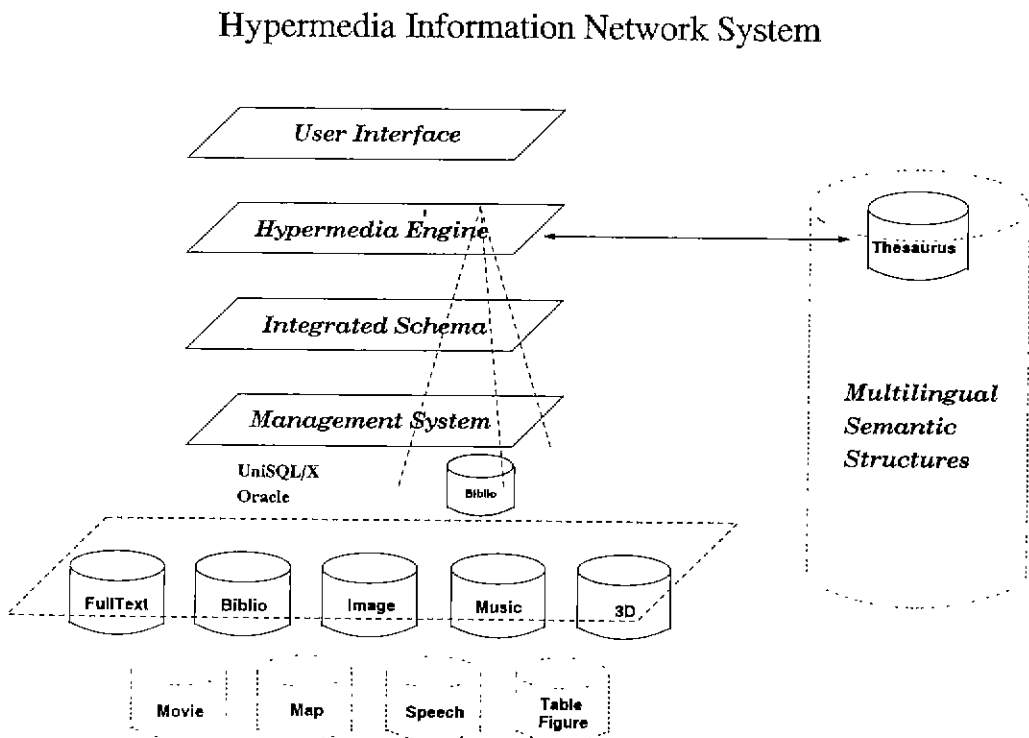


그림 2-2 HINS의 구성 개념도

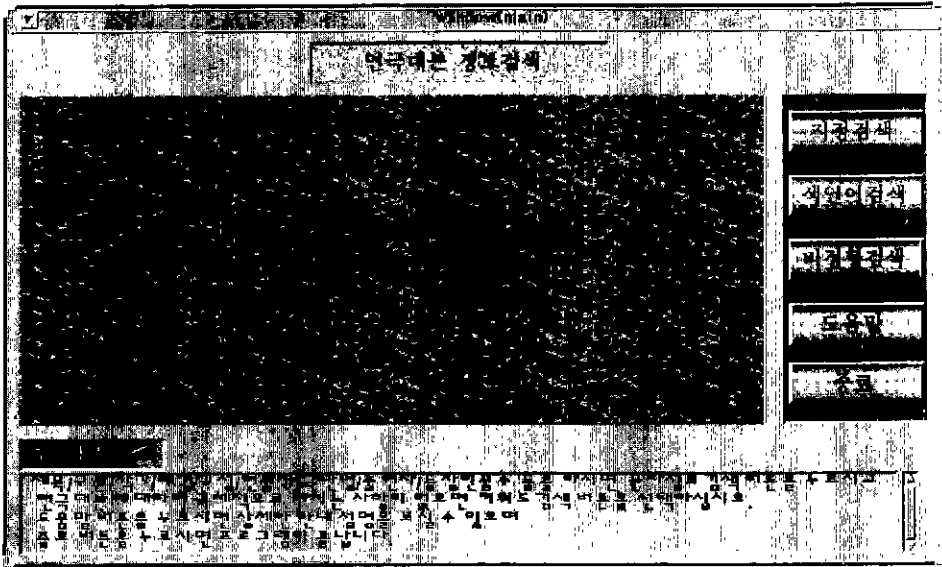


그림 2-3 HINS의 초기화면

7. 정보통신 체계의 정립 및 새로운 개념의 통신 프로토콜 개발

현재는 주로 연극대본을 주요 대상데이터로 선정하고 원시데이터 확보 및 가공에 주력하고, 관리시스템으로는 객체지향형의 UniSQL¹⁾을 도서정보의 관리·운영을 위해서만 별도의 상용시스템을 이용하고 있다. 따라서 스키마 설계는 확보된 기초 데이터를 중심으로 시행착오를 거듭하며 최적화하고 하이퍼미디어 엔진은 최소한의 기본기능의 제공을 목표로 개발해 왔다. 유저 인터페이스는 OpenInterface²⁾를 이용하여 기본화면을 제공하고 기종에 무관한 통신 소프트웨어로의 이용이 용이함을 검증했다. 인위적인 데이터 가공에 HTML을 이용하여 Mosaic을 통해 브라우징할 수 있도록 했다[17-21]. 그림 2-3은 HINS의 초기화면이다.

HINS는 모델시스템으로 적은 양의 데이터를 대상으로 구축된 실험적 시스템이며, 여전히 고려해야 할 커다란 문제로는 다음과 같은 것이 있다[22].

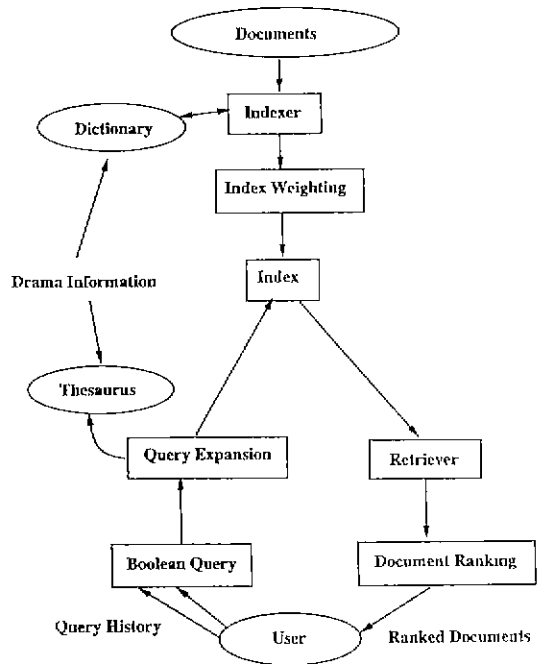


그림 3-1 HINS의 초기단계 정보검색 엔진의 구성개념도

1. 공개 시스템으로 전환시의 보완
2. 대량의 하이퍼미디어 제작물의 통합관리 및

¹⁾UniSQL사의 등록상표이다.
²⁾Neuron Data의 등록상표이다.

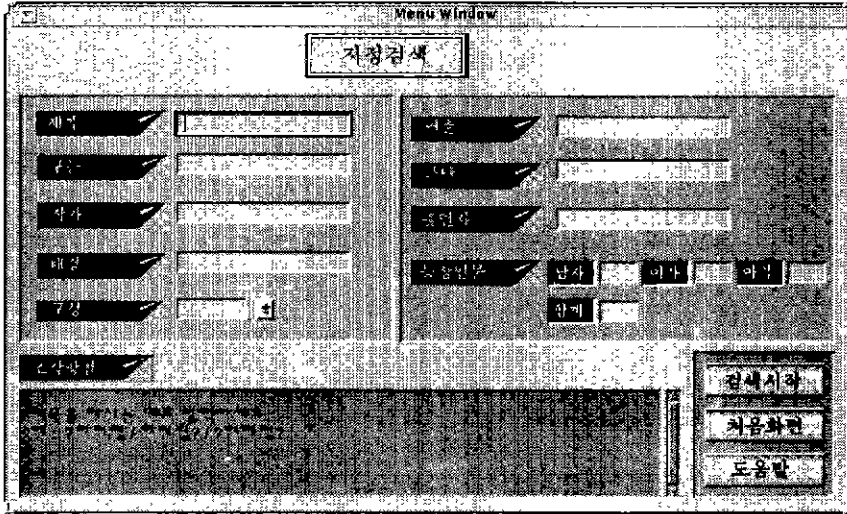


그림 3-2 문자검색을 위한 기초화면

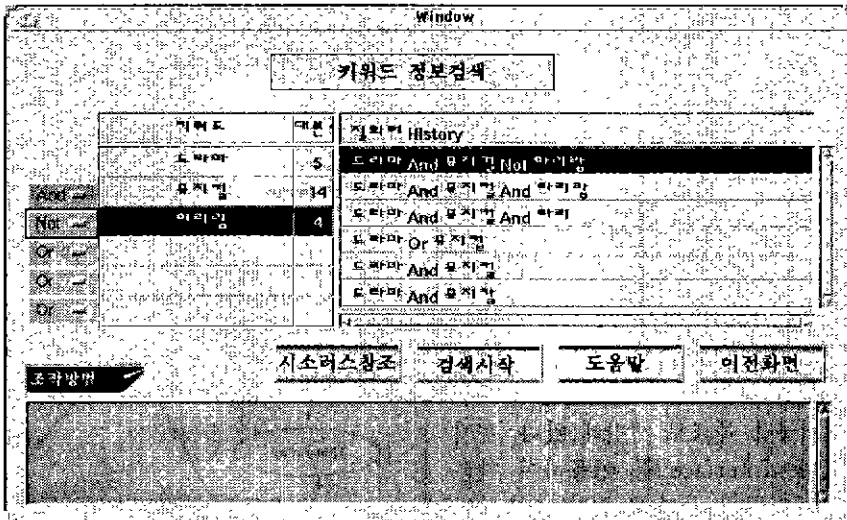


그림 3-3 질의 형성 화면

운영

3. 방대한 정보공간을 대상으로 한 UI

위 문제점은 정보특성의 반영 및 유연한 정보표현방법의 개발, 정보모델의 정립에도 밀접히 관련된 사항이다.

3. 정보검색

정보검색에 초점을 맞추고 축적된 검색대상 데이터의 형태는 사전,全文정보(Fulltext), 전문에 포함된 그림, 표 등, 이미지, 소리, 영상, 지도, 음성 등 다양하다. 하이퍼미디어 시스템에서 고려되어야 할 두 가지 중요한 난제는 대량정보의 처리와 구조적 정보의 취급이라 할 수 있다. 전자는 의미처리와 정보공간내의 구조 파악에 의해

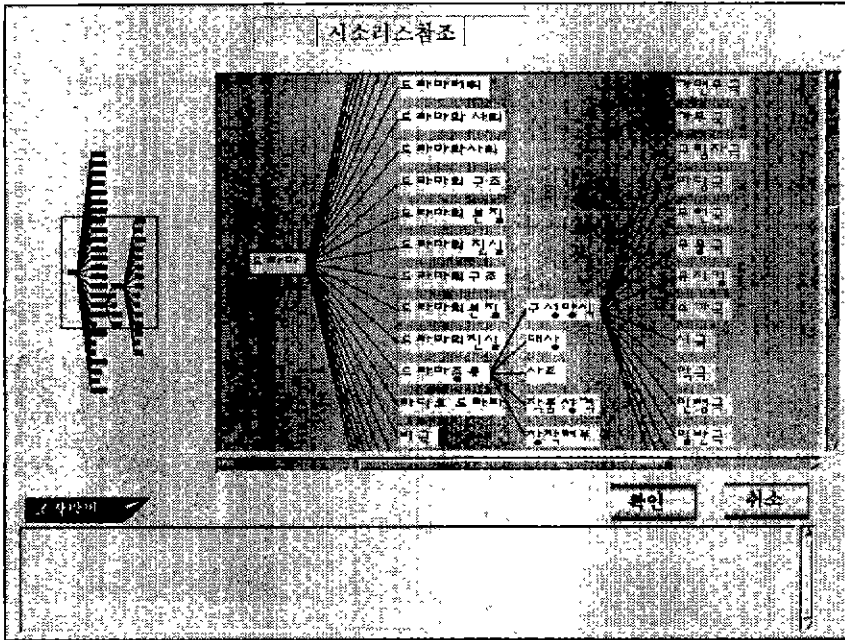


그림 3-4 시소러스 참조화면

제출번호	제출일자	제출인명	제출직위	제출부서	제출내용	제출결과
1991년 7월	1991년 7월	김영희	연구원	연구부	시소러스의 개념과 필요성	수용
1991년 8월	1991년 8월	김영희	연구원	연구부	시소러스의 구성과 활용	수용
1991년 9월	1991년 9월	김영희	연구원	연구부	시소러스의 개발과 평가	수용
1991년 10월	1991년 10월	김영희	연구원	연구부	시소러스의 미래와 결론	수용

그림 3-5 검색결과 디스플레이 화면

해결될 수 있으며, 후자는 자기조직적인 링크의 생성, 링크에 대한 직접역세스, 목적이나 관점에 따른 동적 구조의 브라우징 등에 의해 해결될 수 있다.

그림 2-1의 관리시스템의 발달에 따른 HINS의 초기단계 정보검색엔진의 구성은 그림 3-1과 같고, 그림 3-2~3-6은 정보검색과정을 나타내는 중간 화면구성이다. 검색엔진은 다음과 같은 특징을 갖는다.

3.1 HINS의 정보검색

1. 대규모의 사전, 시소러스의 구축을 전제로



그림 3-6 Mosaic을 이용한 가공데이터의 브라우징

한다.

2. 전문정보에 대한 키워드 색인 기법으로 자동색인을 이용한다.

3. 블리안 연산을 기반으로 자연어 검색으로 확장한다.[23-25]

4. 용어에 대한 Weigting을 근거로 도큐먼트의

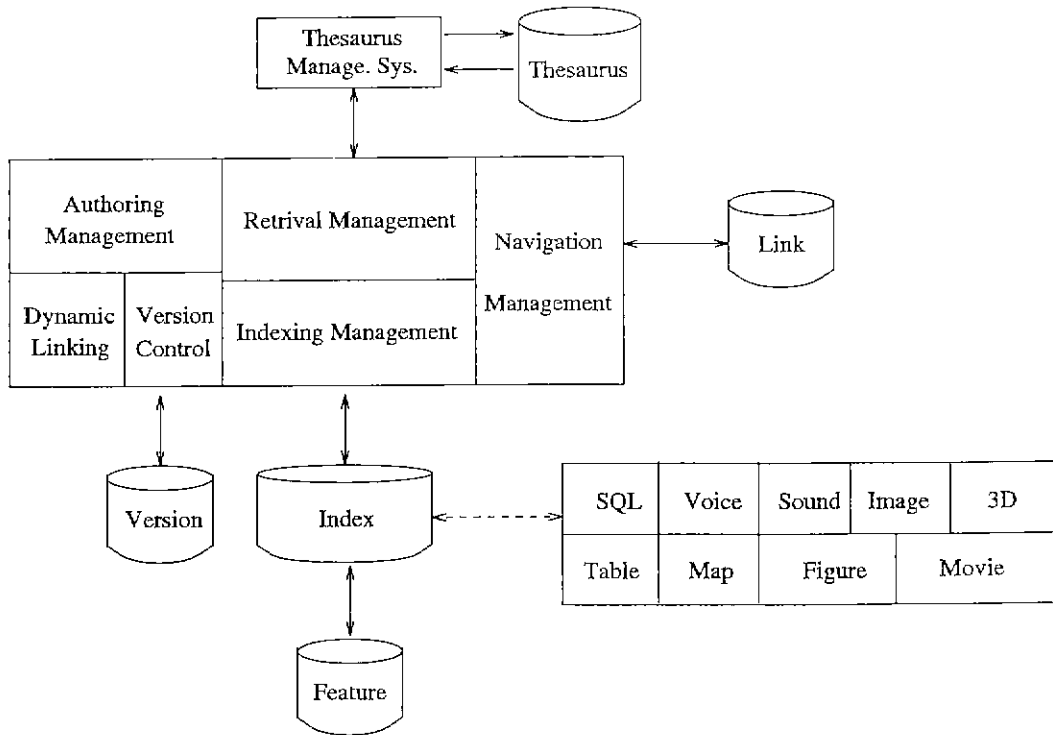


그림 3-7 HAIR의 구성

Rank를 책정한다[26-28,47].

3.2 HAIR

HAIR(Hypermedia Advanced Information Retrieval) 엔진은 대량의 하이퍼미디어 제작물 및 구조적 정보의 취급에 있어서의 문제점을 해결하기 위해 고안된 검색엔진이다(그림 3-7).

HAIR의 특징은 다음과 같다.

1. 대규모의 의미구조를 이용한다.
2. 정의된 관계를 기반으로 동적으로 링크를 생성한다.
3. 멀티미디어 데이터에 대한 특성추출의 자동화와 직접검색을 가능하게 한다.
4. 저작 및 검색의 유형을 버전 관리한다.

이상의 특징은 다음 장에서 논의될 의미구조 모델과 밀접한 관련이 있다.

4. 의미구조 모델

정보모델은 다음과 같은 기본조건을 만족해야 한다.

1. 정보객체에 대한 정의가 주어져야 한다.
2. 오퍼레이션 집합과 조작언어(Manipulation Language)가 정의되어야 한다.
3. 범위와 한계가 정해져야 한다.

그러나 이렇게 정의된 정보모델에 의해 구축된 데이터베이스는 약 5%, 지식베이스는 약 0.5%의 성공율을 나타낸다. 대부분의 실패요인은 정확한 데이터의 확보가 이루어지지 않았거나 정보모델의 기본조건이 적절하지 못했던 것에 기인한다. 또 실세계의 정보공간에 대한 인식부족에서 비롯되는 경우도 있다. 정보모델의 기초이론으로 가장 많이 이용된 것은 그래프 모델이다. ADT [29], 네트워크 모델, ER 모델[30], 의미망 모델, 프레임 이론, 하이퍼미디어, 객체지향 데이터모델 [46] 등이 그래프 모델에 근거한 정보모델이다.

4.1 확장된 하이퍼그래프

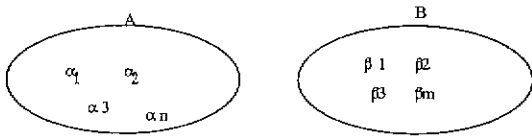


그림 4-1a 개념관계(계층관계)

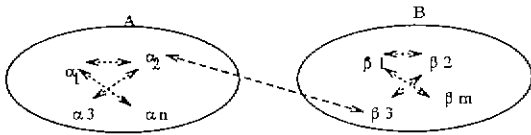


그림 4-1b 논리관계(인과관계)

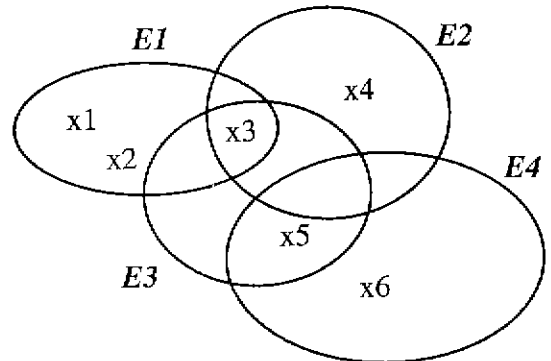


그림 4-2 하이퍼그래프 개념도

상대성, 중복성 등의 정보특성을 반영할 수 없기 때문에 적용된 모델이 하이퍼그래프 모델이다. 유한집합 X로 이루어지는 하이퍼그래프 H는 X의 부분집합의 집합으로 다음과 같이 정의된다[31].

정의 1 하이퍼그래프(Hypergraph)

x_i 를 노드, E_i 를 링크라 하고, $X = \{x_1, x_2, \dots, x_n\}$ 일 때,

X상의 하이퍼그래프 $H = \{E_1, E_2, \dots, E_m\}$ 는 다음의 조건을 만족해야 한다.

$$E_i \neq \emptyset (i=1, 2, \dots, m)$$

$$\bigcup_{i=1}^m E_i = X$$

그래프모델이나 하이퍼그래프모델을 기반으로 하는 많은 다루고자 하는 구조는 크게 개념구조와 논리구조를 포함하는 정보구조이다. 개념구조는 계층관계와 등가관계로 이루어지는 정보구조이며, 논리구조는 주로 인과관계에 의해 이루어진 정보구조이다(그림 4-1). 이러한 정보구조를 의미구조[32]라고 한다.

그림 4-1a는 개념 A, B 사이의 계층관계를 보인 것이다. 개념 A는 속성집합 α 로 이루어지고 α 는 $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ 의 속성들을 갖는다. $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ 은 개념 A를 구성하는 내부구조이며 개념 A의 하위개념이다. 개념 B는 속성집합 β 로 이루어지고 속성집합 β 는 $\{\beta_1, \beta_2, \dots, \beta_m\}$ 의 속성들을 갖는다. $\{\beta_1, \beta_2, \dots, \beta_m\}$ 는 개념 B를 구성하는 내부구조이며 개념 B의 하위개념이다. 그림 4-1b는 개념 A, B 사이의 연관관계를 보인 것

이다. 연관성은 단일개념내의 속성사이에 존재할 수 있으며 다른 개념간에 존재할 수 있다. 다음은 단일개념구조에서 속성사이의 연관성 및 방향성을 보인 것이다.

$$(\alpha_1 \rightarrow \alpha_2)$$

$$(\alpha_2 \rightarrow \alpha_1)$$

$$(\alpha_2 \rightarrow \alpha_3)$$

$$(\alpha_3 \rightarrow \alpha_2)$$

개념 B의 β_3 는 개념 A의 속성 α_2 와 연결됨으로써 개념 A와 B는 상호 관련성을 갖는다.

$$(\alpha_2 \rightarrow \beta_3)$$

$$(\beta_3 \rightarrow \alpha_2)$$

→는 개념의 확장방향을 표현하며, →의 왼쪽편 개념이 오른쪽편 개념으로 유도됨을 표시한다.

하이퍼그래프는 개념구조와 논리구조의 융합에 의한 표현을 이루고 있으며, 관계설정의 제약에서도 탈피하고 있다(그림 4-2).

하이퍼그래프에는 DRLH[33-35], Higraph, AHOM, GROOVY 등이 있다. 그러나 관계정의의 범위 및 한계에 대한 표현이 부족하다. 확장된 하이퍼그래프에서도 관계자체로 단위정보 또는 단위정보가 이루는 구조에 의해 표현된다는 점을 감안하여 실세계의 정보공간을 표현하는 정보모델을 정의한다. 확장된 하이퍼그래프는 다음의 특징을 갖는다.

표 4-1 구조화 정보의 표현수준

	이항관계	다대다 관계	상대성	중복성	내부구조	의미관계
그래프	◎	△	X	X	X	△
하이퍼그래프	◎	◎	◎	◎	X	△
확장된 하이퍼그래프	◎	◎	◎	◎	◎	◎

1. 단위정보들간의 관계는 단위정보이거나 단위정보들에 의해 이루어진 구조이다.
2. 단위정보는 그 자체로 또는 관계로 또는 구조의 구성원으로 작용한다.
3. 관계에 대한 직접접근이 가능하다.
4. 개념구조와 트리구조를 포함한 의미구조 표현이 용이하다.

정의 2 확장된 하이퍼그래프(Extended Hypergraph)

$$SS = \{HL_1, HL_2, \dots, HL_n\}$$

정의 2의 확장된 하이퍼그래프 SS는 하이퍼링크 HL의 집합으로 이루어지며, HL은 하나의 구성요소를 표현하는 단순 하이퍼 링크와 두개 이상의 구성요소를 갖는 복합 하이퍼링크로 구성된다[32]. 의미는 동류노드에 의해 표현하며, 하이퍼링크는 개념을 구성하는 구성요소들로 이루어진다. 정보공간은 노드와 링크로 구성되며, 정보공간을 구성하는 구성요소는 고정되지 않고 변화한다. 변화는 추가된 HL 집합에 의해 생성된다. 하이퍼링크의 중첩은 개념의 내부구조를 표현한다[36]. 이렇게 확장된 하이퍼그래프를 의미구조그래프(Semantic Structure Graph)라 한다[37-40,48].

정보구조를 표현하기 위한 정보모델로 의미구조 그래프를 사용한다. 정보표현을 위한 의미구조 그래프는 개념간을 연결하는 링크와 정보의 의미를 기술하기 위한 정보를 표현하는 노드로 구성되는 확장된 하이퍼그래프(extended hypergraph)를 사용한다. 의미구조 그래프상의 개념간의 연결은 의미구조 그래프내에 부그래프로

표 4-2 의미구조 기술문법

```

<term> ::= '<entry>'[('<descriptor>')]
<entry> ::= <string>
<descriptor> ::= '['<relation>']' ('<label>') ('<descriptor>')
<label> ::= (<entry>['<label>'] | <term>['<label>'])
<relation> ::= (<term> | <entry>)
    
```

기술되며, 그래프의 내부는 외부에 존재하는 관련된 외부 하이퍼 링크내에 연결한다. 그래프, 하이퍼그래프, 확장된 하이퍼그래프의 특징을 비교하면 표 4-1과 같다.

4.2 의미정보 표현

정보모델의 제약과 함께 지식, 의미, 속성 등으로 일컬어 지는 실세계정보의 표현에는 많은 제약이 따른다. 이러한 제약은 정보모델의 발달, 새로운 기능의 개발, 하드웨어의 성능 향상 등에 따라 이미 축적된 많은 정보를 재활용하지 못하고 사장시키는 경우도 많다. 따라서 정보모델의 진화 등에 상관없이 초기에 확보한 원시데이터의 효용가치가 유지될 수 있는 정보표현방법이 필요하다.

텍스트정보, 특히 지식이나 의미에 관한 실세계 정보의 기술을 위한 방법으로, 정보표현의 융통성을 제공하고 운영의 효율성을 위해 정보표현기법으로 표 4-2와 같이 EBNF를 이용하였다. 용어간의 관계는 관계명으로 설정되며 용어에 대한 설명은 구조적으로 표현될 수 있다. 이 정보표현기법을 의미구조 표현문법이라 하며 정보분석, 정보관계의 명확한 표현, 구성이 가능하며, 하이퍼링크를 갖는 확장된 하이퍼그래프 구

표 4-3 의미구조 기술문법에 의한 정보표현의 예

```
(유예지){
  [대별하여]{(현금자보){
    [한자명]{(絃琴字譜),
    [표기]{(육보){
      [한자명]{(肉譜)
    }
  }
},
(당금자보){
  [한자명]{(唐琴字譜)
}
}
```

성을 위한 입력형태이다[41].

현재는 *term*이나 *relation*에 구조를 허락하고 있지 않으나, 입력정보의 특성에 따라 확장가능하다. 따라서 위 기술문법과 같은 형태는 방향성, 재귀성, 중복성 등의 정보특성을 반영한 원시데이터 확보방안이라 할 수 있다. 용어기술(Term Description) TD는 *entry*, *relation*, *label*로 구성되어 확장된 하이퍼그래프 SS를 이룬다. 하이퍼링크인 TD는 의미를 표현하며 각 구성요소는 다음의 특징을 갖는다.

1. *entry*는 정보단위인 그래프의 식별자를 나타낸다.
2. *label*은 *entry*로부터 *relation*에 의해 연결되며, *term*에 대한 기술이다.
3. *label*은 *entry* 또는 다른 TD를 포함할 수 있고 *entry*와 *label*은 의미를 표현한다.
4. *relation*은 두개의 개념을 연결하는 관련성을 표현한다.

표 4-3은 표 4-2의 문법에 따라 작성된 의미구조의 예이다.

4.3 의미구조 관리시스템

의미구조기술 문법에 따라 쓰여진 정보표현은

정보모델을 사용하여 정형화할 수 있다. 정형화된 의미구조는 표 4-2의 표기법으로 기술된 의미구조 소스로 부터 하이퍼노드와 하이퍼링크를 갖는 확장된 하이퍼그래프를 구성한다. 의미구조 관리시스템의 하부의 물리구조는 내부구성으로서의 서지정보와 소스정보의 위치로 구성되며 저장의 인덱스와 키워드 인덱스에 의한 검색을 가능하게 한다. 의미구조 관리시스템은 전문가(human expert)로 부터 얻어진 지식간의 충돌 또는 자동구축된 개념구조간의 정보충돌을 검출할 수 있으며, 정보의 여러형태 표현을 단일 의미구조로 표현할 수 있다.

개념 연결은 2개의 개념이 직접연결되는 경우인 직접연결과 2개의 개념이 2번이상의 연결에 의해 이루어지는 간접연결로 구분된다. TD는 *relation*에 의해 *entry*에서 *label*로의 TD₀와 *label*에서 *entry*로의 방향성을 갖는 TD₁로 이루어진다. 즉, *relation*에 의해 연결된 2개의 개념구조를 표현하는 방향그래프이다. 표 4-3은 TD₁={유예지, 대별하여, TD₀}, TD₂={현금자보, 표기, TD₃} 등의 TD 집합을 갖는 확장된 하이퍼그래프 SS로 구성된다. SS의 특성을 정리하면 다음과 같다.

1. 단일관계는 2개의 개념을 연결한다.
2. 관계를 표현하는 동사/명사는 *term*으로 재구성이 가능하다.
3. 하이퍼노드의 포인터가 NIL인 경우 *term*으로 분리 가능하다.
4. 관계간에는 계층관계를 갖는다.
5. 개념간의 연결강도는 임의의 위치에서 위 또는 아래로 진행해 감에 따라 약해진다.
6. 최하위개념을 이루는 관계는 종단 연결된다.
7. 독립적으로 표현된 TD는 계층을 갖는 다른 TD의 서브그래프로 표현될수 있다.
8. 관계명과 용어명에 의해 그래프의 노드를 구성한다.

의미구조 표현문법으로부터 구성된 TD의 집합인 SS를 운영하기 위한 의미구조 관리시스템(semantic structure management system)은 입력소스로부터 관련정보를 추출하기 위한 전처리 단계를 필요로 한다. 그림 4-3은 시스템의 구성을 나타낸 것이다.

표 4-4 생성된 용어테이블의 예

번호	식별자	종류
7	육보	T
8	현금자보	TD
9	당금자보	TD
16	한자명	R
17	표기	R

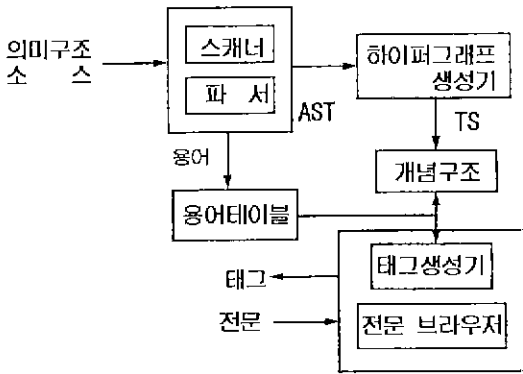


그림 4-3 입력소스의 전처리 및 SS 구성

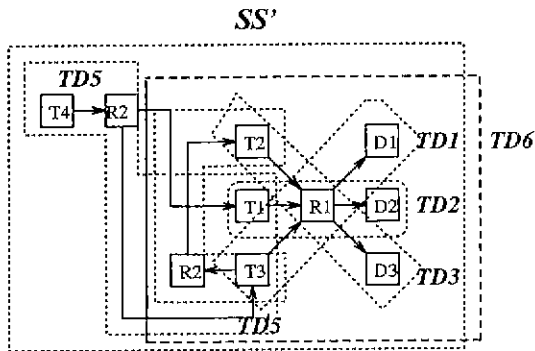


그림 4-4 단위 TD

확장된 하이퍼그래프의 구성을 위한 과정은 다음의 과정을 통해 이루어 진다.

1. EBNF표기로 기술된 의미구조를 인식하는 스캐너 및 파서 구성
2. 파서의 출력인 AST 구성
3. AST로부터 그래프 구성이 가능한 term 및 entry 추출
4. 추출된 term으로부터 TD 구성
5. TD로부터 TD_i를 구한 후 SS 구성

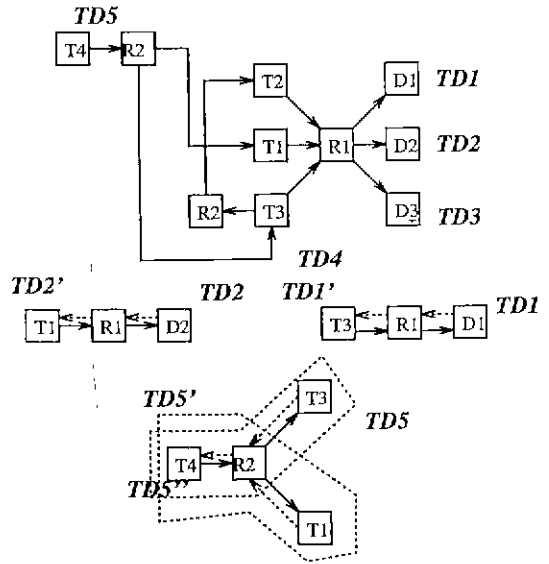


그림 4-5 재구성된 TD

하이퍼그래프 구성을 위한 전단계로 의미구조 소스를 처리하기 위한 스캐너와 파서를 구성한다. 스캐너는 식별자 토큰추출 및 기타 토큰의 식별을 위해 사용하며, EBNF표기로 기술된 의미구조를 TD로 구성하기 위해 컴파일러 자동화 도구인 YACC를 사용하여 파서를 구성한다.

파서의 출력형태인 AST의 운행과정에서 노드 식별자에 따라 관계명, 용어명. 기술자를 종류로 갖는 용어테이블을 생성한다. 표 4-4는 그림 4-3으로부터 생성된 용어테이블을 보인 것이다. 생성된 용어테이블은 전문(fulltext)의 태깅 및 전문의 검색과정에서 검색 키워드로 사용된다. 표 4-4의 종류값 D는 descriptor, T는 term, R은 relation을 나타내며, 하이퍼그래프에서 “당금자보”는 term과 descriptor로 서로 다른 TD내에서 존재함을 나타낸다.

의미구조내의 용어는 하이퍼그래프를 구성하는 서브그래프인 TD로 구성된다. TD의 구성을 위해 YACC 입력형태에서 각각의 용어를 용어 리스트에 추가한다. 용어 리스트는 의미구조의 하위의 개념에서 상위개념으로 연결되며, 연결 리스트를 운행하여 내포된 그래프(nested subgraph)를 검출한다.

검출된 그래프는 그림 4-4와 같으며 용어 그

래프는 방향성에 의해 새로운 TD인 TD_r를 생성하게 되며 그림 4-4로부터 생성된 SS는 그림 4-5와 같으며 11개의 TD로 이루어진다.

5. 개념구조의 생성

정보과학 분야에서 필요로 하는 의미해석 기술 개발에 중요한 역할을 하는 것이 시소러스이다. 정확한 관계 정의 및 대량정보 수용의 필요성은 시소러스 구축의 커다란 장애 요인이다. 시소러스 구축에는 다방면의 전문지식 활용과 막대한 비용 및 시간 투자가 요구된다. 기계적으로 관리 운영이 가능한 시소러스내의 관계는 BT, NT로 표현되는 계층관계와 USE, UF로 표현되는 등가관계로 이루어 진다.

개념적인 관계를 정의하는 두 관계를 기계적으로 추출하고, 기존의 평면적인 시소러스 구조를 실제계의 정보구조에 적합하게 조직화하여 시소러스 구축에 있어서의 문제점을 개선하는데 그 목적이 있다. 제시되는 알고리즘은 단일 언어내의 시소러스 구축 뿐만 아니라, 구축된 시소러스들의 융합 및 다국어 시소러스 구축에도 적용된다.

5.1 개념구조로의 시소러스

시소러스(Thesaurus)란 “어떤 두 단어간의 개념적인 계층관계를 정의한 관련어 사전”이라 정의된다. 시소러스의 시초라 일컬어 지는 P. M. Roget의 “Thesaurus of English words and phrases”는 문장 작성시에 표현하고 싶은 의미를 명확히 나타내는 단어를 선택하기 위해서 이용된 유사어 사전이었다. 도서관에서 방대한 도서를 정리하기 위해 사용된 주제표목표(subject headings list)도 시소러스의 일종이라 하겠다. 기계에 의한 자연언어처리에서는 정보검색시에 원하는 정보를 추출하기 위하여 사용되는 검색용어 집으로서의 역할을 수행한다.

기존의 시소러스로는 Thesaurus of Engineering and Scientific Terms(TEST), BSI³⁾ ROOT 시소러스, CAB 시소러스, Food : multilingual

thesaurus, INSPEC 시소러스, Thesaurus of ERIC descriptors, UNESCO:IBE 교육 시소러스, UNBIS 시소러스, IRRD⁴⁾ 시소러스, JICST 과학 기술용어 시소러스, 뉴스 시소러스⁵⁾ 등이 있다. 표현방법이나 범위 등에 따라 각각 특성을 갖고 있지만, 기본적으로는 등가관계, 계층관계, 연관관계와 같은 기본적인 관계들로 구성되어 있다.

1) 등가관계

BS 5723 및 ISO 2788에서는, 색인작업시 복수개의 단어가 같은 개념을 나타내고 있다고 인정되는 경우, 우선어 및 비우선어의 관계라고 정의하고 있다. 우선어란 그 개념을 표현하기 위하여 색인 작업시 이용되는 단어를 말하며, 비우선어란 이용되지 않는 용어를 말한다. 우선어에 대한 접두기호로는 USE, 비우선어에 대한 접두기호로는 UF(use for)가 사용된다.

2) 계층관계

개념상 상위어 및 하위어의 관계를 정의한 것이다. 상위어는 클래스 또는 전체를 나타내며, 하위어는 그 한 요소 또는 일부분을 나타낸다. 계층관계는 상위 및 하위 개념을 논리적으로 전개하는 순서로 위치시키기 위해 이용된다. 상위어 표시기호로는 BT(broader term), 하위어 표시기호로는 NT(narrower term)가 사용된다.

3) 연관관계

연관관계란 색인작성과 검색에 이용될 지도 모르는 대체용어를 제시하기 위해 용어간의 관계가 심리적으로 연상된다면 시소러스 내에 명시하는 것이다. 일반적으로 이 관계를 나타내기 위한 기호로는 RT(related term)를 이용한다.

개념간의 관계는 관계명으로 관계가 설정되며, 두가지 개념은 서로 관계명으로 링크된다. 이 관계명은 방향성(양방향성)을 가지고 있으며 하나의 관계류가 하나의 층(layer)을 형성한다.

5.2 문제점분석

⁴⁾International Road Research Documentation

⁵⁾일본 중일신문사

³⁾British Standard Institute

기존의 시소러스의 구축 및 유지에 있어서 중요한 문제점으로 고려되어야 할 사항으로는 다음과 같은 것이 있다.

1) 많은 시간과 비용을 필요로 한다.

분야에 따라 차이는 있겠지만 한 전문분야의 시소러스를 구축하는 데는 5~10년이 걸린다. 시소러스 구축에 선행되어야 할 것이 용어 선정이다. 전문용어 사전구축이 여기에 해당하는 선행 작업이라 하겠다. 일본 문부성이 주관한 “과학기술전문용어집”에 포함된 전문용어 하나당 40,000원 상당의 비용이 요구되었고, 5년간의 프로젝트로 진행된 EDR 사전구축에는 11조2천억원이라는 비용이 소요되었다. 더욱 어려운 점은 힘들여 수집한 용어가 시간의 흐름에 따라 변한다는 것이다.

결국 시소러스의 완성은 참여자들의 포기과 때를 같이 한다고 하겠다.

2) 대상 전문분야별로 많은 전문가를 필요로 한다.

구축된 시소러스의 관리 및 이용 전문가는 있을 수 있어도, 시소러스 구축 전문가란 존재할 수 없다. 시소러스를 구성하고 있는 용어의 의미는 단일 전문분야내에서도 반드시 일치하는 것은 아니어서, 용어의 개념정립 등에 전문가들의 지식이 필수적이다. 그러나 이들 전문가도 반복적인 시소러스 구축작업에는 의욕을 보이지 않는 경우를 흔히 접할 수 있다.

3) 여러 분야에 걸친 전문가를 필요로 한다.

지식 표현의 대표적인 수단이 언어이다. 언어는 생성, 성장, 소멸의 과정과 더불어 변화한다. 새로운 지식이나 수정된 지식을 표현하기 위해 기존의 용어를 이용하는 경우가 대부분이다. 그러나 기존의 용어는 타분야의 지식을 표현하기 위해 또는 다른 개념을 나타내기 위해 이미 이용되었던 것이다. 따라서 기존의 시소러스의 틀에 맞추기 위해서는 여러 분야의 전문가가 모여서 타협할 수 밖에 없다. 이 과정은 ‘표준화’와 자주 혼동되는 경우가 많으며 소집단의 주관적 관점에서 구축될 위험이 있다.

4) 다언어 시소러스 구축이 어렵다.

언어간 개념차이 및 어감차이로 인하여 용어의 의미가 일치하지 않는다. 외국의 시소러스를 번역한 사례에 자주 등장하듯이, 언어간에 일대일 대응이 이루어지지 않는 경우가 빈번하다. 통신기술의 발달에 따라 국제적인 정보 유통이 활발해 지고 서로의 정보를 공유하고자 하는 움직임이 대두되면서 번역이 불가능한 부분은 그대로 반영하면서 다언어 시소러스 융합에 관한 재의도 찾아 볼 수 있다[42].

이상의 문제점을 분석한 결과는 다음과 같다.

1) 실세계의 정보구조에 적합한 정보모델이 필요하다.

정보(용어)의 특성은 중의성, 방향성, 재귀성으로 대표될 수 있다. 그러나 기존의 모델은 이러한 정보의 특성을 수용할 수 없다. 따라서 n차원의 정보공간에 표현할 수 있는 정보모델과 판리시스템이 요구된다. 동시에 고려해야 할 사항이 유연한 지식표현방법이다. 전문(Fulltext)에서 인식하는 것이 가장 정확하지만, 전문인식을 위해서도 기계가독형의 지식표현이 유용하다. 시스템 구축과정도 순환적으로 정의된다(제 5장 참조).

2) 시소러스의 자동구조화가 필요하다.

다언어(Multilingual) 다층(Multilayer)의 시소러스 구축도 순환적으로 이루어 진다. 구축에서 뿐만 아니라 유지, 관리에 있어서도 자동화는 필수적이다. 자동구축된 것을 포함해서 모든 시소러스의 평가는 정보검색에 있어서 시소러스의 이용결과에 근거한다. 따라서 새로운 정보를 바탕으로 학습하고 전문가의 참여를 최소화하면서 일관성 있는 시소러스의 유지, 관리가 이루어져야 한다.

5.3 시소러스 자동구조화

정보검색 시스템에서의 시소러스는 용어들간의 관계를 설명하는 개념구조의 조직과 같은 것으로 간주되며, 실세계 정보공간에서의 관계는

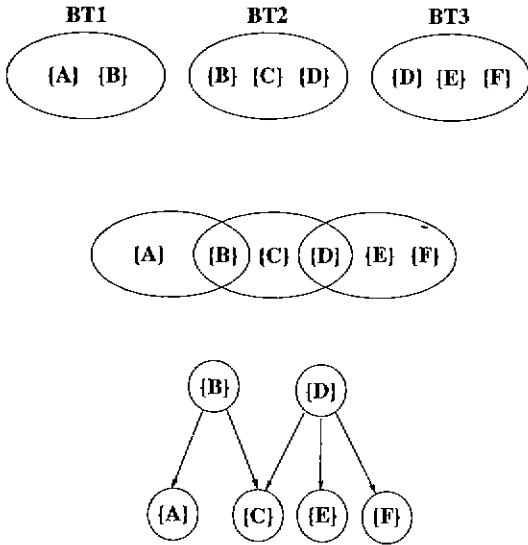


그림 5-1 시소러스 자동생성 기본개념

목적이나 관점에 따라 변화한다. 그러나 시소러스에서의 계층적 관계는, 용어가 고정되어 있기 때문에 여러가지 의미에 따라 표현되지 않는다. 예를 들어, CRCS(Computing Review Classification Structure) 분류표나 HRAF(Human Relations Area Files) 분류표 등을 보면 이들의 분류표는 오랜 세월 수정, 추가를 거듭해 오면서 많은 변화가 주어 졌다. 그럼에도 불구하고 분류표내의 중복성 및 모순은 여전히 남아 있다. 그 원인은 바로 목적이나 관점에 따라 분류표의 구성이 동적으로 변화해야 하는 것을 무리하게 고정적으로 2차원 공간에 표현하고자 했던 까닭이다. 실세계 정보에 대한 분류는 시소러스 구축 개념에 중요한 영향을 준다. 고정된 틀에 맞춰 정보를 구조화 하는 것이 아니라, 자유로운 구조와 틀을 허용하여 실세계의 정보 구조에 적합한 지식을 구축하는 것이다. 이렇게 구축된 지식은 고급의 정보처리 즉, 추론, 유추, 학습 등을 기계적으로 수용할 수 있는 기반이 된다. 그리고 상호관계를 나타내는 기호중의 하나인 RT는 여러가지의 의미를 가지지만 RT들간의 관계에 대한 의미는 상세히 설명되지 않는다. RT 자체가 상당히 주관적이기 때문에 이용자의 입장을 충분히 반영할 수 없다. 정보검색 측면에서만 보

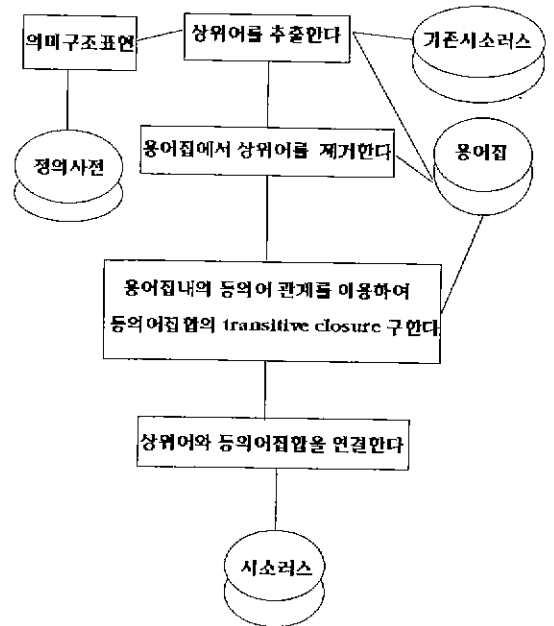


그림 5-2 시소러스 자동생성 작업흐름도

더라도 이용자의 요구가 반영되지 않거나 이용자의 요구가 반영된 검색 결과가 너무 방대해서 원하는 결과를 찾지 못하는 결과를 초래한다. 또한 기존의 시소러스는 다방면의 전문지식 활용과 막대한 비용 및 시간 투자를 필요로 하였고, 대부분 오래전에 만들어진 것이어서 새로운 용어에 대한 갱신이 이루어지지 않았을 뿐만 아니라 갱신을 할 수도 없다.

정보는 관점에 따라 다른 의미를 가지며, 정보의 조직화는 기존의 정보로부터 학습 및 추론을 통해 얻어지는 지식 획득(knowledge acquisition) 과정이다. 통제언어(Control Word) 시스템에서 유래한 전거통제(Authority Control)는 표준화라는 전제하에 인위적인 색인시스템을 구축한다. 다음과 같은 이유에서 전거통제는 시소러스에 흡수된다.

1. 표현의 다양성은 동의어 관계로 흡수된다.
2. 연상되는 모든 표현을 관리할 수 없다.
3. 이용자의 요구가 없는 한 새로운 표현은 관리, 검색되지 않는다.

동의어집합을 생성하여 상위어 또는 상위어집

표 5-2 동의어집합 생성 알고리즘

단어수	용어결과	비율(%)
1	36,337	31.0
2	66,085	56.4
3	11,861	10.1
4	2,320	2.0
5	405	0.3
6	105	0.1
7	26	0.0
8	8	0.0
9	4	0.0
10	6	0.0

\mathcal{M}, \mathcal{N} : 각 용어의 집합

\mathcal{E} : 동의어 관계의 집합

S_m, S_n : 각각 \mathcal{M}, \mathcal{N} 의 동의어 집합

동의어 집합의 초기 상태($m \in \mathcal{M}, n \in \mathcal{N}$) :

$$S_m^0 = \{m_0\}, S_n^0 = \{\}$$

동의어 집합 :

$$S'_n = \{n_k | (m_0, n_k) \in \mathcal{E}\}$$

동의어 관계를 이용한 \mathcal{M}, \mathcal{N} 의 동의어 집합 :

$$S'_n = \bigcup_{m_p \in S'_m} \{n_q | (m_p, n_q) \in \mathcal{E}\}$$

$$S_n^{i+1} = S_n^i \cup S'_n$$

$$S'_m = \bigcup_{n_s \in S_n^{i+1}} \{m_r | (m_r, n_s) \in \mathcal{E}\}$$

$$S_m^{i+1} = S_m^i \cup S'_m$$

$S_m^+ \subseteq \mathcal{M}, S_n^+ \subseteq \mathcal{N}$ 이 TC일 조건 :

$$S_n^+ = S_n^{i+1} = S_n^{i+2} = \dots = S_n^+$$

$$S_m^+ = S_m^{i+1} = S_m^{i+2} = \dots = S_m^+$$

'ord Element Index in Terminology Context)은 복합어의 특성을 이용하여 계층관계를 추출하는 새로운 타입의 인덱스이다.

▲SS-KWIC의 여섯가지 규칙[36]

1. 수정 원소들과 수정된 원소들의 위치
2. 원소들의 결합
3. 구성 원소들의 수
4. 접두사, 접미사, 수사 등과 같은 특수 원소
5. 명사 또는 복합어의 복수형
6. 그외(생략, 단순화 등등)

2) 용어의 정의(definition) 사용

합과 연결하여 전거처리의 과정은 시소러스 구축 및 참조과정 내에 포함된다. 동의어집합의 생성은 용어집 내에 표현된 동의어관계를 이용하여 기계적으로 구축할 수 있다. 그러나, 용어의 중의성에 의해 동의어집합이 무한대로 커지는 것이 단점이라 하겠다. 그림 5-1과 같이 구조화하기 위해서는 상위어를 추출해야 한다.

이상의 문제점을 바탕으로 시소러스를 자동구조화하기 위한 과정은 그림 5-2와 같다[37].

5.3.1 상위어 추출방법

기본적으로 시소러스에서의 상위어는 전문가에 의해 추출되지만, 이를 위해서는 많은 시간과 높은 비용을 요구하며, 또한 시소러스를 유지하기 위해서는 자동적으로 상위어를 추출해야만 하는 문제점을 안고 있다. 다음은 상위어 추출을 위한 4가지 방법이다.

1) 조어의 규칙(Coinage rule) 사용

일반적으로 용어는 대부분 복합어로 되어 있다. 통계분석에 의하면 전체 용어에서 복합어가 차지하는 비율이 약 70%가 된다(표 5-1). 조어의 규칙이란 복합어들 중에서 공통된 단어를 포함한 용어들의 집합을 구한 후 그 단어를 상위어로 추출하는 방법을 말한다.

조어의 규칙을 이용한 상위어 추출방법의 하나로 SS-KWIC(Semantically Structured Key W

용어의 정의는 관계된 많은 용어들을 수록한 사전에 정의되어 있다. 새로운 용어는 잘 정의된 (well-defined) 용어에 의해 새로이 정의됨으로써 새로운 용어와 기존에 있던 용어들 간의 관계가 정의된다. 예를 들어, “~의 일종이다.”라고 하면, 그 용어에 대한 계층관계를 나타내는 말로 사용되기 때문에 이러한 방법에 의해 계층관계를 생성할 수 있다.

3) 기존의 시소러스 사용

기존의 시소러스는 대부분 전문가들에 의해 수동으로 만들어 졌다. 비록 기존의 시소러스들이 구축된 오래되고, 새로운 용어들의 갱신이 이루어지지 않았다 하더라도, 분류용어에 가까운 기본적인 용어들간의 관계를 정의하고 있으므로 시소러스 자동구조화에 폭 넓게 활용된다. 타 시소러스와의 융합 및 세분화된 분야의 하위어 집합과의 연계에도 효율적이다.

4) 용어의 의미관계 표현에서 추출

만약 의미구조가 구축되었다면, 이 의미구조 안에서 표현된 의미관계에서 상위어를 추출할 수 있다. 즉, 의미구조 모델이 새로운 용어를 정의하는데 사용되면, 새로운 용어와 기존의 용어 사이의 관계는 의미구조안에 묘사된다.

상위의 세가지 방법은 전문가에 의해 결과를 최적화해야 할 필요가 있지만, 이 방법은 시소러스를 자동 구조화시키는데 가장 좋은 방법이라 할 수 있다.

5.3.2 동의어집합 생성방법

용어들 사이의 동의어 관계를 이용하여 자동으로 동의어 집합을 생성하여 구조화시키는 방법으로는 C-TRAN(Constrained Transitive Closure)를 이용한다. Transitive Closure(TC)는 용어에 대한 동의어 관계를 이용하여 동의어 집합을 구하는 방법으로서, 시작 용어를 선택하여 그 용어에 대한 동의어를 구하고, 시작용어와 동의어들을 포함한 동의어 집합에 대해 더 이상 동의어 요소(element)가 증가하지 않을 때까지 동의어 관계를 이용하여 동의어를 찾아 동의어 집

합에 추가시킨다. 동의어 집합의 TC를 구하기 위한 알고리즘은 표 5-2와 같다[43].

6. 결 론

실세계의 정보공간과 정보의 특성을 분석하고 정보의 효율적 표현 및 처리를 위해 의미구조 그래프를 사용한 정보모델에 의한 정보정형화 방법을 제안하였다. 의미구조그래프의 형태로는 확장된 하이퍼그래프를 채용하였으며, 확장된 하이퍼그래프를 사용한 정보검색 및 운영은 저장된 전문의 의미구조를 관리함으로써 정보의 변경과 추가에 따른 효율적인 정보운영이 가능하며, 그래프의 운행과정에서 새로운 정보의 생성이 가능하였다.

본 연구는 기존의 시소러스 구축 및 관리에 수반되는 제반 문제점들을 파악하여, n 차원 공간에서 표현되는 시소러스 자동구축 방법을 제시하였다. 제시된 방법은 기존의 시소러스의 융합 및 타언어로 구축된 시소러스의 흡수와 언어의 장벽을 초월한 다국어 시소러스 구축의 가능성을 보인다. 따라서 본 연구는 앞으로 방대한 데이터를 연구 결과에 적용시켜 구축 가능성을 확인한 후, 본 연구 결과가 적용된 다국어 시소러스 자동 구축 및 멀티미디어 데이터를 수용한 다국어 다매체 시소러스 자동구축 방법 및 그 활용에 대한 연구를 추진할 예정이다.

참고문헌

- [1] D. Gentner. The Mechanisms of Analogical Reasoning. "Similarity and Analogical Reasoning". Cambridge University Press 1989.
- [2] R. S. Michalski. "A Theory and Methodology of Inductive Learning". *Artificial Intelligence*, pp. 111-116, Vol. 20, 1983.
- [3] Y. Fujiwara. "The Model for Self Structured Semantic Relationships of Information and Its Advanced Utilization". In *47th FID Congress, Tokyo, 1994*.
- [4] W. G. Lee. "Construction of Semantic Structures in the Self-Organizing Information-Base System". PhD thesis, The University of Tsukuba,

- February 1993.
- [5] E. F. Codd. "Extending the Database Relational Model to Capture More Meaning". *ACM Transactions on Database Systems*. Vol. 4, No. 4, pp. 397-434, 1979.
- [6] C. Zaniolo. "The Database Language GEM". *Proceedings of SIGMOD'83*. 1983.
- [7] F. Manola and U. Dayal. "PDM: An Object-Oriented Data Model". *Proceedings of IEEE International Workshop on Object-Oriented Database Systems*. 1986.
- [8] A. Snyder. "Encapsulation and Inheritance in Object-Oriented Programming Languages". *Proceedings of OOPSLA'86*. 1986
- [9] D. Maier, J. Stein, A. Otis, and A. Purdy. "Development of an Object-Oriented DBMS". *Proceeding of the 1st ACM Conference on Object-Orient Programming Systems Languages and Applications*. pp. 472-482. 1986.
- [10] D. Maier and J. Stein. "Development and Implementation of an Object-Oriented DBMS". *Research Directions in Object-Oriented Programming*. 1987.
- [11] J. Banerjee, H. T. Chou, J. F. Garza, W. Kim, D. Woelk, N. Ballou, and H. J. Kim. "Data Model Issue for Object-Oriented Applications". *ACM TOIS*. Vol. 5, No. 1, 1987.
- [12] J. Banerjee, W. Kim, H. J. Kim, and Henry F. Korth. "Semantics and Implementation of Schema Evolution in Object-Oriented Databases". *Proceedings of ACM SIGMOD*. pp. 311-322, 1987.
- [13] W. Kim, J. Banerjee, H. T. Chou, J. F. Garza, and D. Woelk. "Composite Object Support in an Object-Oriented Database System". *Proceedings of OOPSLA'87*. pp. 118-125. October, 1987.
- [14] D. H. Fishman, D. Beech, H. P. Cate, E. C. Chow, T. Connors, J. W. Davis, N. Derrett, C. G. Hoch, W. Kent, P. Lyngbaek, B. Mahbod, M. A. Neimat, T. A. Ryan, and M. C. Shan. "Iris: An Object-Oriented Database management systems". *ACM TOIS*. Vol. 5, No. 1, 1987.
- [15] C. Lecluse, P. Richard, and F. Velez. "O2, an Object-Oriented Data Model". *Proceedings of SIGMOD* 1988.
- [16] 이원규, 이상헌, 김명철, 김성훈. "문화정보망 구축계획(안)". 한국문화예술진흥원. 1993.
- [17] 김명철 외. "하이퍼미디어를 이용한 문화예술 정보검색". 제1회 한국정보관리학회 전국논문대회. 1994.
- [18] T. H. Nelson. "The Hypertext". *Proceedings of the World Documentation Federation*. 1965.
- [19] F. G. Halasz, T. P. Moran, and R. H. Trigg. "NoteCards in A Nutshell". *Proceedings of CHI%GI*. 1987.
- [20] M. H. Chignell and P. A. Hancock. "Intelligent Interfaces". *Handbook of Human Computer Interaction*. 1988.
- [21] K. Greonbaek and Randall H. Trigg. "Hypermedia System Design Applying the Dexter Model". *Communications of the ACM*. Vol. 37, No. 2. 1994.
- [22] F. G. Halasz. "Reflections on NoteCards: Seven Issues for the Next Generation of Hypermedia Systems". *Communications of the ACM*. Vol. 31, No. 7, 1988.
- [23] G. Salton, E. A. Fox, and H. Wu. "Extended Boolean Information Retrieval". *Communications of the ACM*. pp. 1022-1036, Vol. 26, No. 12, 1983.
- [24] C. P. Paice. "Soft Evolution of Boolean Search Queries in Inforation Retrieval Systems". *Information Technology, Res. Dev. Applications*. pp. 33-42. Vol. 3, No. 1, 1984.
- [25] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong. "Extended Boolean Query Processing in the Generalized Vector Space Model". *Information Systems*. pp. 47-63, Vol. 14, No. 1, 1988.
- [26] W. B. Croft and P. Savino. "Implementing Ranking Strategies Using Text Signatures". *ACM Transactions on Office Information Systems*. pp. 42-62, Vol. 6, No. 1, 1988.
- [27] D. Harman and G. Candela. "Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking". *J. American Society for Information Science*. 1990.
- [28] G. Salton and C. Buckley. "Term-Weighting Approaches in Automatic Text Retrieval". *Information Processing and Management*. pp. 513-523, Vol 24, No. 5. 1988.

- [29] M. Stonebraker, B. Rubenstein, and A. Guttmann. "Application of Abstract Data Types and Abstract Indices to CAD Data". In Proc. of Ann. Meeting Database Week. pp. 107-115, 1983.
- [30] P. P. S. Chen. "The Entity-Relationship Model: Toward a Unified View of Data". ACM Transactions on Database System. pp. 9-36, Vol. 1, No. 1, 1986.
- [31] C. Berge. "Hypergraphs". North-Holland, 1989.
- [32] N. Uda, W. G. Lee, Y. Fujiwara. "Construction of Semantic Structures in the Self-Organizing Information-Base Systems". J. of Society of Information and Knowledge. Vol. 3, No. 1, 1993.
- [33] H. Boley. "Directed Recursive Labelnode Hypergraphs and their use as Representations for Knowledge". free session contribution to the Fourth Int. Joint Conf. Artificial Intelligence. 1975.
- [34] H. Boely. "A Theory of Representation (-Language, -Constructions, and -Relations)". Technical Report IFI-HH-M-38 Inst. fuer Informatik 1976.
- [35] H. Boley. "Directed Recursive Labelnode Hypergraphs: A New Representation-Language". Artificial Intelligence. pp. 49-85, Vol. 9, No. 1, 1977.
- [36] N. Uda. "Information Analysis for Modeling and Representation of Meaning". PhD thesis, The University of Tsukuba. February 1994.
- [37] Y. Fujiwara, W.G. Lee, Y. Ishikawa, T. Yamaguchi, A. Nishioka, K. Katada, N. Ohbo, and S. Fujiwara. "A Dynamic Thesaurus for Intelligent Access to Research Databases", 44th FID Congress. pp. 173-181, Helsinki, 1988.
- [38] Y. Fujiwara and T. Makino. "Management and Advanced Utilization of Semantically Organized Terminology and Knowledge". Proc. of TKE'93. pp. 141-151, 1993.
- [39] Y. Fujiwara, N. Uda, and X. Zhang. "Analogical Reasoning in Polymer Information Base Systems". 13th CODATA. 1992.
- [40] Y. Fujiwara, J. He, G. Chang, N. Ohbo, H. Kitagawa, and K. Yamaguchi. "Self Organizing Information Systems for Material Design". In Proc. of CAMSE'90. Tokyo, 1990.
- [41] 강윤희, 조성호, 이원규. "의미구조를 기반으로 한 정보모델". 제1회 한국정보관리학회 전국논문대회. 1994.
- [42] G. Rahmstorf. "A New Thesaurus Structure for Semantic Retrieval". In S. Fujiwara editor, 47th FID Congress. pp. 114-121, Tokyo, 1994.
- [43] 김해수, 이남경, 이원규. "시소러스 자동구조화". 제1회 한국정보관리학회 전국논문대회. 1994.
- [44] M. S. Olivier, S.H. Von Solms. "A Taxonomy for Secure Object-Oriented Databases". ACM Transactions on Database Systems. Vol. 19, No. 1, pp. 3-46, 1994.
- [45] Q. Li and D. McLeod. "Conceptual Database Evolution Through Learning in Object Databases". IEEE Transactions on Knowledge and Data Engineering. Vol. 6, No. 2, pp. 205-224, 1994.
- [46] A. Poulouvassilis and M. Levene. "A Nested-Graph Model for the Representation and Manipulation of Complex Objects". ACM Transactions on Information Systems. Vol. 12, No. 1, pp. 35-68, 1994.
- [47] N. Fuhr and U. Pfeifer. "Probabilistic Information Retrieval as a Combination of Abstraction, Inductive Learning, and Probabilistic Assumptions". ACM Transactions on Information Systems. Vol. 12, No. 1, pp. 92-115, 1994.
- [48] H. Yasukawa and K. Yokota. "Labeled Graphs as Semantics of Objects". Technical Report TR-600, ICOT. 1990.

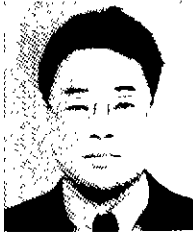
이 원 규



1985 고려대학교 영어영문학과(문학사)
 1987 (일본) 츠크바대학 이공학연구과(연구생)
 1989 (일본) 츠크바대학 이공학연구과(공학석사)
 1993 (일본) 츠크바대학 공학연구과 전자정보공학전공(공학박사)
 1993 ~ 현재 한국문화예술진흥원 책임연구원
 관심 분야: 정보표현 및 모델, 하이퍼미디어, 멀티미디어 정보검색

데이터베이스, 의미구조, 하이퍼미디어, 멀티미디어 정보검색

김 명 철



1986 서울대학교 전기공학과(공학사)
 1988 한국과학기술원 전산학과(석사)
 1988 ~ 1992 금성소프트웨어(주) 연구원
 1993 ~ 현재 한국문화예술진흥원 선임연구원
 관심 분야: 정보검색, 한글처리, 하이퍼미디어

김 성 훈



1987 고려대학교 이과대학 수학과(학사)
 1989 고려대학교 이과대학 전산학과(이학석사)
 1991 ~ 고려대학교 이과대학 전산학과(박사과정)
 1993 ~ 현재 한국문화예술진흥원 선임연구원
 관심 분야: 객체지향형 데이터베이스, 멀티미디어 데이터베이스

강 윤 희



1989 동국대학교 전산학과(공학사)
 1991 동국대학교 전산학과(공학석사)
 1991 ~ 1994 시스템공학연구소 연구원
 1994 ~ 현재 한국문화예술진흥원 선임연구원
 관심 분야: 프로그래밍 언어 및 컴파일러, 정보검색, 하이퍼미디어 시스템

김 해 수



1985 경희대학교 물리학과(학사)
 1988 ~ 1993 한국표준과학연구원 연구원
 1994 ~ 현재 한국문화예술진흥원 선임연구원
 관심 분야: 의미구조, 정보검색, 한글처리

김 문 호



1982 중앙대학교 수학과(학사)
 1982 ~ 1985 금성제전
 1985 ~ 1992 시그마 시스템
 1992 ~ 1994 오롬컴퓨터
 1994 ~ 현재 한국문화예술진흥원 책임연구원
 관심 분야: 하이퍼미디어 시스템, 데이터모델, 데이터통신

● '95 Teach-The-Teachers ●

- 일 자 : 1995년 2월 16일~17일
- 장 소 : 수안보 상록호텔
- 주 관 : 전산교육연구회
- 문 의 : 숙명여대 전산학과 최종원 교수
 T. 02-710-9445, F. 02-710-9296