94-b-2

의미구조를 기반으로 한 정보모델

The Information Model Based on Semantic Structures

○강윤희 조성호 이원규
한국문화예술진흥원

YunHi Kang SungHo Cho WonGyu Lee
{yhkang,shcho,lee}@caibs.kcaf.or.kr
The Korean Culture & Arts Foundation, Seoul, KOREA

요 약

과거 실세계 정보를 처리하기 위한 방법으로는 관계형데이타베이스, 객체지향데이타베이스, 지식베이스 시스템 등이 연구되었다. 이들 방법은 제한된 정보표현 및 정보의 운영 및 접근방법 등의 문제점을 갖는다. 정보의 구조화는 정보의 의미를 분석하고 정보의 특성에 적합한 융통성 있는 정보모델을 필요로 한다. 본 논문에서는 방대한 양의 정보처리 및 다양한 형태의 표현, 동적 변환 등의 정보특성을 효율적으로 처리하기 위한 정보모델로 의미구조그래프를 사용하여 기존 시스템의 문제점을 해결하기 위한 방법을 제안한다. 의미구조그래프를 사용한 정보구조화는 정보의미를 분석할 수 있으며, 정보의 표현의 융통성을 제공한다. 의미구조그래프는 노드와 링크를 갖는 확장된 하이퍼그래프를 사용하였으며, 정보구조화를 위한 대상데이타로 문화예술 분야의 관련 정보를 실험하였다.

1. 서 론

실세계(real world) 정보는 매우 방대하고 다양한 형태로 표현되며 동적으로 변환되고 전이되는 중의성(polyseme)을 갖는다. 과거 실세계 정보를 처리하기 위한 방법으로는 관계형 데이타베이스관리시스템(Relational DBMS)[5], 객체지향데이타베이스관리시스템(Object Oriented DBMS)[4], 지식베이스 시스템(Knowledge Base System), 전문가시스템(Expert System)[6] 등이 있다.

이들 시스템은 제한된 영역(closed world)의 데이타만을 가공하여 처리함으로써 실세계의 정보에 대해 의미적인 처리를 할 수 없다는 단점을 갖는다. 효과적인 실세계 정보 처리방법인 정보의 구조화는 정보의 의미를 분석하여 정보 특성에 적합하고 융통성을 제공하는 정보모델을 필요로 한다. 본 논문에서는 문화예술분야의 관련 정보를 의미구조화 하기 위해서 확장된 하이퍼그래프를 제시하고, 정보의 중간표현을 위해 확장된 하이퍼그래프를 사용하여 정보를 구조화하였다.

2. 정보의 특성

정보는 관점에 따라 다른 의미를 가지며, 정보의 조직화는 기존의 정보로부터 학습 및 추론을 통해 얻어지는 지식 획득(knowledge acquisition) 과정이다.

2.1 의미구조

정보는 단독적으로 사용되지 않으며, 용어들간의 관련성에 의해 구성되는 개념의 형태로서 존재한다. 개념은 다음과 같은 특징을 갖는다.

- 속성들로 이루어진 내부구조를 갖는다.
- 내부구조를 갖는 개념은 계층구조를 표현할 수 있다.
- 상호 중첩되고, 중첩되는 부분은 개념의 유사성을 표현한다.
- 방향성을 갖는다.

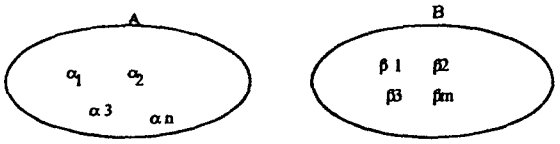

그림 1: 개념 관계(계층관계)

그림 1은 개념 A, B 사이의 계층관계를 보인것이다. 개념 A는 속성집합 α로 이루어지고 α는 $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$의 속성들을 갖는다. $\{\alpha_1\ \alpha_2 \ldots \alpha_n\}$은 개념 A를 구성하는 내부구조이며 개념 A의 하위개념이다. 개념 B는 속성집합 β로 이루어지고 속성집합 β는 $\{\beta_1\ \beta_2 \ldots \beta_m\}$의

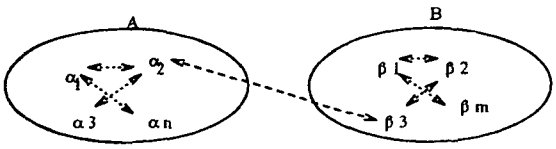

그림 2: 개념 관계(연관관계)

속성들을 갖는다. $\{\beta_1\ \beta_2\ \ldots\ \beta_m\}$는 개념 B를 구성하는 내부구조이며 개념 B의 하위개념이다.

그림 2는 개념 A, B 사이의 연관관계를 보인것이다. 연관성은 단일개념내의 속성사이에 존재할 수 있으며 다른 개념간에 존재할 수 있다.

개념관계에 따른 정보의 표현을 위해서는 계층적 관계에 의한 네트워크 구조, 정보의 운영을 위한 정보 모델을 필요로 한다. 의미구조는 개념관계의 구조화된 형태로서 이를 정형화하기 위해서는 의미구조 모델을 사용한다.

3. 의미구조모델

개념구조 표현을 위해 사용되는 의미구조 그래프는 DRLH, Higraph, AHOM, GROOVY 등이 있으며[2]. 본 연구에서는 하이퍼링크로 이루어지는 확장된 하이퍼그래프를 의미구조 모델을 위해 사용하였다. 본 장에서는 의미구조 표현을 위한 확장된 하이퍼그래프와 의미구조의 소스입력형태, 확장된 하이퍼그래프의 구성을 기술한다.

3.1 확장된 하이퍼그래프

개념구조를 표현하기 위한 중간표현으로 의미구조 그래프를 사용한다. 정보표현을 위한 의미구조 그래프는 개념간을 연결하는 링크와 정보의 의미를 기술하기 위한 정보를 표현하는 노드로 구성되는 확장된 하이퍼그래프(extended hypergraph, EHG)를 사용한다.

의미구조 그래프상의 개념간의 연결은 의미구조 그래프내에 부그래프로 기술되며, 그래프의 내부는 외부에 존재하는 관련된 외부 하이퍼 링크내에 연결한다. 유한집합 X로 이루어지는 하이퍼그래프 HG는 X의 서브집합으로 정의 1과 같다[1].

정의 1 (하이퍼그래프)

$$X = \{x_1, x_2, \ldots, x_n\}$$
$$HG = \{E_1, E_2, \ldots, E_m\}$$
$$E_i \neq \phi (i = 1, 2, \ldots, m)$$
$$\bigcup_{i=1}^{m} E_i = X \square$$

그림 3은 $x_1, x_2, \ldots, x_6$ 의 노드와 $E_1, E_2, \ldots, E_4$ 의 링크로 구성된 하이퍼그래프의 예이다.

정의 1의 HG는 개념의 집합관계와 중첩을 표현하지만, 개념의 내부구조와 관련성 표현이 어렵다. 이를 해결

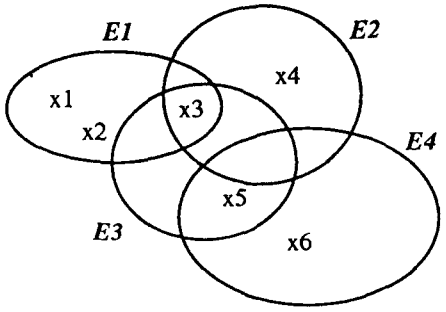

그림 3: 하이퍼그래프의 예

하기 위해 개념의 중첩, 개념관계를 표현하는 레이블, 개념의 방향성을 표현할 수 있는 확장된 하이퍼그래프를 사용한다.

정의 2 (확장된 하이퍼그래프)

$$EHG = \{HL_1, HL_2, \ldots, HL_n\} \square$$

정의 2의 EHG는 하이퍼링크 HL의 집합으로 이루어지며, HL은 하나의 구성요소를 표현하는 단순하이퍼 링크와 두개이상의 구성요소를 갖는 복합하이퍼링크로 구성된다[7]. 하이퍼링크는 개념을 구성하는 구성요소들로 이루어진다.

정보공간은 노드와 링크로 구성되며, 정보공간을 구성하는 구성요소는 고정되지 않고 변화한다. 변화는 추가된 HL 집합에 의해 생성되어진다. 하이퍼링크의 중첩은 개념의 내부구조를 표현한다[2].

3.2 입력 의미구조 표현

실세계 정보의 기술을 위한 방법으로 정보표현의 융통성을 제공하고 운영의 효율성을 위해 의미구조 기술문법을 표 1와 같이 EBNF로 표현하였다.

개념간의 관계는 관계명으로 설정되며 두개의 개념은 서로 관계명으로 링크된다. 관계명은 방향성을 갖으며, 관계정의의 틀이 된다. 의미구조 표현문법은 정보 분석, 정보 관계의 명확한 표현, 정보 구성이 가능하며, 하이퍼링크를 갖는 확장된 하이퍼그래프 구성을 위한 입력형태이다.

그림 4는 표 1의 문법에 따라 작성된 의미구조의 예이다.

3.3 확장된 하이퍼그래프의 구성

의미구조기술 문법에 따라 쓰여진 정보표현은 정보모델을 사용하여 정형화할 수 있다. 정형화된 의미구조는 표 1의 표기로 기술된 의미구조 소스로 부터 하이퍼노드와 하이퍼링크를 갖는 확장된 하이퍼그래프를 구성한다. 정의 3은 표 1를 기본으로 확장된 하이퍼그래프 EHG'을 정의한것이다.

표 1: 의미구조기술 문법

```
<Term> ::= '('<Entry>')'['{'{<Desciptor>}'}']
<Entry> ::= <String>
<Descriptor> ::= '['<Relation>']'
          {'{'<Label>'}'} {','<Descriptor>}
<Label> ::= (<Entry>{','<Label>}
            |<Term>{','<Label>})
<Relation> ::=  (<Term> | <Entry>)
```

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

그림 4: 의미구조 예

정의 3 (확장된 하이퍼그래프2)

$$TD = \{entry, relation, label\}$$
$$EHG' = \{TD_1, TD_2, TD_3, \ldots, TD_n\}\square$$

*EHG'*는 하이퍼링크인 *TD(Term Description)* 집합으로 구성되며, 개념을 표현한다. 하이퍼링크 *TD*의 구성요소는 다음의 특징을 갖는다.

- *entry*는 정보의 단위인 그래프식별자를 나타낸다.
- *label*은 *entry*로 부터 *relation*에 의해 연결되며, *term*에 대한 기술이다.
- *label*은 *entry* 또는 다른 *TD*를 포함할 수 있고 *entry*와 *label*은 개념을 표현한다.
- *relation*은 두개의 개념을 연결하는 관련성을 표현한다.

개념 연결은 2개의 개념이 직접연결되는 경우인 직접연결과 2개의 개념이 2번이상의 연결에 의해 이루어지는 간접연결로 구분된다. *TD*는 *relation*에 의해 *entry*에서 *label*로의 TD_o와 *label*에서 *entry*로의 방향성을 갖는 TD_r로 이루어진다. 즉, *relation*에 의해 연결된 2개의 개념구조를 표현하는 방향그래프이다. 그림 4는 $TD_1 = \{$유예지, 대별하여, $TD_6\}$, $TD_2 = \{$형금자보, 표기, $TD_3\}$ 등의 *TD* 집합을 갖는 확장된 하이퍼그래프 *EHG'*로 구성된다.

4. 시스템 설계 및 구현

의미구조 소스로 부터 구성된 *TD* 집합인 *EHG'*를 운영하기위한 의미구조 관리시스템(semantic structure management system)의 구성은 입력 의미구조 소스로 부터 관련 정보를 추출하기 위한 전처리단계가 필요하다.

확장된 하이퍼그래프의 구성을 위한 과정은 다음의 과정을 통해 이루어지며, 그림 5은 시스템 구성을 보인 것이다.

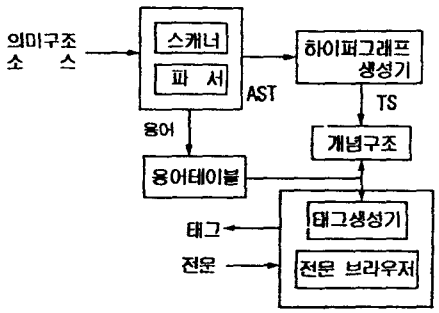

그림 5: 시스템 구성

의미구조 관리 시스템의 하부 물리구조는 서지정보와 인덱스로구성되며, 저장의 인덱스와 키워드에 의한 검색을 가능하게 한다.

4.1 스캐너 및 파서 구성

하이퍼그래프 구성을 위한 전단계로 의미구조 소스를 처리하기위한 스캐너와 파서를 구성한다. 스캐너는 식별자 토큰 추출 및 기타 토큰의 식별을 위해 사용하며, EBNF 표기로 기술된 의미구조를 TD로 구성하기 위해 컴파일러 자동화 도구인 YACC을 사용하여 파서를 구성한다[3].

4.2 용어테이블의 생성

파서의 출력형태인 AST의 운행과정에서 노드식별자에 따라 관계명, 용어명, 기술자를 종류로 갖는 용어테이블을 생성한다. 표 2은 그림 4로 부터 생성된 용어테이블를 보인 것이다. 생성된 용어테이블은 전문(fulltext)의 tagging 및 전문의 검색과정에서 검색 키워드로 사용된다. 표 2의 종류값 D는 Descriptor, T는 Term, R은 Relation을 나타내며, 하이퍼그래프에서 "당금자보"는 Term과 Descriptor로 서로 다른 TD내에서 존재함을 나타낸다.

4.3 TD의 구성

TD 집합은 의미구조의 하위의 개념에서 상위개념으로 연결되며 내포되어지는 그래프(nested subgraph)를 검출한다. 검출된 그래프는 그림 6와 같다. 그림 6의 TERM 그래프는 방향성에 의해 새로운 *TD*인 TD_r를 생성하게 되며 그림 6의 예로 부터 생성된 전체 *EHG'*는 그림 7과 같다. 그림 7의 *EHG'*는 11개의 *TD*로 이루어진다.

표 2: 생성된 용어테이블의 예

번호	식별자	종류
7	옥보	T
8	형금자보	TD
9	당금자보	TD
16	한자명	R
17	표기	R

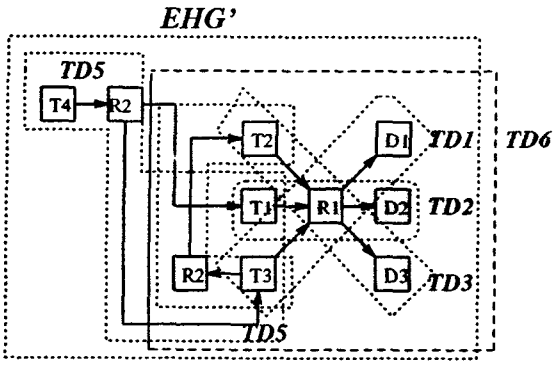

그림 6: TD

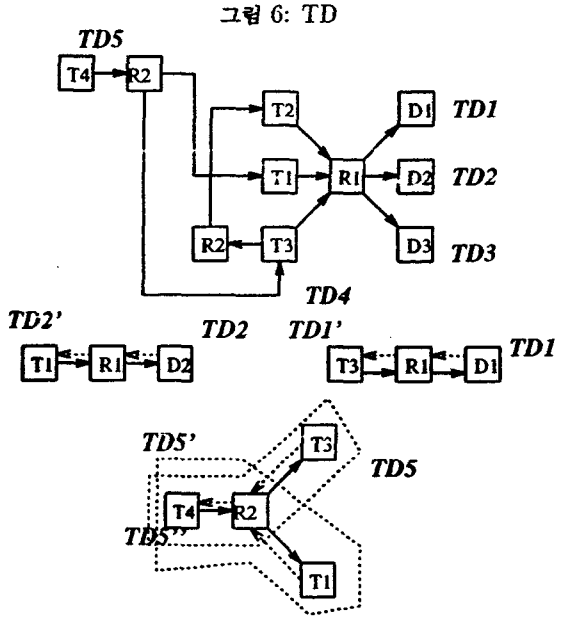

그림 7: 재구성된 TD

5. 결론

본 논문에서는 정보의 효율적 표현 및 처리를 위해 의미구조 그래프를 사용한 정보 모델에 의한 정보정형화 방법을 제안하였다. 의미구조그래프의 형태로는 확장된 하이퍼그래프를 채용하였으며, 확장된 하이퍼그래프를 사용한 정보검색 및 운영은 저장된 전문의 의미구조를 관리함으로서 정보의 변경과 추가에 따른 효율적인 정보운영이 가능하며, 그래프의 운행과정에서 TD_r의 새로운 정보의 생성이 가능하였다.

의미구조 관리 시스템 구성은 개념구조(conceptual structure)를 구성하기 위한 시소러스와 논리구조구성을 위한 분류표(taxonomy)로 부터 가능하다.

현재의 의미구조는 용어간의 관계가 명사의 형태를 가지므로 정확한 관계 표현이 어렵고 효율적인 정보의 검색 및 정보의 재생산을 위해서는 사전의 구축, 관계명의 구조표현, 동적분류표의 작성이 필요하다.

참고 문헌

[1] C. Berge, Hypergraphs, North-Holland, 1989

[2] Norihiko Uda, Information Analysis for Modeling and Representation of Meaning, Ph.D thesis, pp. 6-30, 1994

[3] A.V.Aho, R.Sethi and J.D Ullman, Compiler principles, Techniques and Tools, Addison Wesley, 1986

[4] 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

[5] E.F. Codd, "Extending the Database Relational Model to Capture More Meaning", ACM Transations on Database Systems, Vol. 4, No. 4, pp. 397-434, 1979

[6] R.S. Michalski, "A Theory and Methodlogy of Inductive Learning", Artifical Intelligence, Vol. 20, pp. 111-116, 1983

[7] N.Uda, W.G. Lee and Y. Fujiwara, "Construction of Semantic Structures in the Self-Organizing Information-Base Systems", Journal of Janpan Society of Information and Knowledge, Vol. 3, No. 1, 1994