온톨러지 재사용을 위한 범주 재분류

양 재 군[†] · 배 재 학[††] · 이 종 혁[†††]

요 약

본 논문에서는 기존 온톨러지를 가공하여 용도에 맞게 변환하는 방안을 모색하였다. 변환방법으로서는 범주정보 단순화와 구체화를 고안하였다. 이 각각은 다시 범주표제와 기저범주를 이용한 방법으로 나누어 생각하였다. 또한 상이한 범주집합들 사이의 관계를 밝히는 방법도 도출하였다. 정립한 변환 방법론을 활용하여, (1) Roget 시소러스로부터 7개의 범주로 구성된 '이야기 분석용 온톨러지'[32]의 원형을 구축하였고, (2) 이미 알려진 22가지 멀티미디어 게임 흡인요소를 바탕으로 세분화된 흡인요소 207가지를 발견할 수 있었으며[35], 그리고 (3) 10개의 심소와 22가지 멀티미디어 게임 흡인요소 사이의 관계를 밝혀낼 수 있었다[36].

Category Reorganization for Ontology Reuse

Jae-Gun Yang[†] · Jae-Hak J. Bae[††] · Jong-Hyeok Lee[†††]

ABSTRACT

This paper introduces a methodology of transforming an existing ontology into the one that satisfies its application. The transformation consists of simplification and realization of word category information. They are based on category headings and base categories. Furthermore, this paper describes a method by which we can identify relationships between category sets. Through the transformation, (1) Roget's thesaurus is reorganized into 7 categories and the base of "Ontology for Narrative"[32], (2) 22 immersion factors of multimedia games can be subdivided into 207 factors in [35], and (3) the relationships between 10 mental factors and 22 immersion factors of multimedia games are identified in [36].

키워드 : 온톨러지(Ontology), 시소러스(Thesaurus), 범주 재분류(Category Reorganization)

1. 서 론

온톨러지(Ontology)는 철학의 한 갈래로 존재론 또는 존재학이라고도 한다. 학문의 이름이 의미하는 바와 같이, 존재의 본질을 연구하는 형이상학이다. 한편, 인공지능 영역에서의 온톨러지는 인지체에 의하여 개별화된 실체(Entity)들에 대하여 논의나 분류를 가능하게 하는 범주(Category) 시스템으로 본다[1]. 이와 같은 온톨러지는 인지체가 세계를 인식 및 분할하여 얻은 개체(Individual)와 이들의 속성 및 관련성을 파악한 것의 총화이다. 모든 인지체들은 정도의 차이는 있지만 어떤 식으로든지 온톨러지를 내장하고 있다고 본다. 그 이유는 온톨러지가 이해를 포함한 인지능력의 발휘에 필수 불가결한 기본지식이기 때문이다. 언어이해뿐만 아니라 우리가 접할 수 있는 여러 가지 문제나 현상의 파악에도 적절한 온톨러지가 필요함은 익히 짐작할 수 있다.

1.1 연구동기

컴퓨터와 인터넷 기술이 발전함에 따라서, 개인이 접하는 문서의 양은 스스로 처리할 수 있는 수준을 훨씬 상회하고 있다. 따라서 문서 처리를 도와줄 자연어 처리 도구 또는 지식기반의 시스템이 필요하다. 이러한 지능형 시스템의 필수적인 구성요소가 온톨러지[1]이다. 한편 정보의 형태에 기반한 현재의 웹은 점차 정보의 내용을 중요시하는 시맨틱 웹[2]으로 발전하는 추세이다. 온톨러지는 이러한 정보의 내용을 파악하기 위해 필요한 요소이다. 하지만, 각 에이전트[3]들의 온톨러지는 상이한 구조나 형태로 존재할 가능성이 높다. 따라서 시맨틱 웹에서 에이전트들이 검색 혹은 교환한 정보를 이해하기 위해서는, 검색된 결과를 자신의 온톨러지와 정련(Alignment)시키는 과정이 필요하다. 정련과정에서는 상이한 구조나 형태를 단순화시키거나 구체화시키는 재편성 과정이 수반된다. 본 논문에서는 이러한 정련의 전처리 과정에 응용할 수 있는 온톨러지 재편성 방법론을 모색하였다.

※ 본 연구는 한국과학재단 목적기초연구 R05-2004-000-12362-0 지원으로 수행되었음. 또한 부분적으로 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었음.
† 준 회 원 : 울산대학교 대학원 컴퓨터·정보통신공학부
†† 종신회원 : 울산대학교 컴퓨터·정보통신공학부 교수 (교신저자)
††† 정 회 원 : 포항공과대학교 컴퓨터공학과 교수
논문접수 : 2004년 6월 26일, 심사완료 : 2005년 1월 31일

1.2 관련연구

온톨러지 구축 방법론[4]은 크게 두 가지로 나누어 생각할 수 있다. 첫 번째는 새로운 온톨러지를 제작하는 방법이고, 두 번째는 기존의 온톨러지를 재사용하는 방법이다. 전자의 경우에는 주로 필요한 어휘사전(Lexicon)과 분류법 등을 수동으로 취합하여 구축한다. 후자의 경우에는 기존 온톨러지에서 필요한 정보를 추출, 통합하는 방식을 취한다. 비록 온톨러지 재사용에 대한 선행 연구[5, 6]가 있기는 하지만, 목적하는 용도를 충족하면서도 효과적인 재사용 방법론에 대한 지속적인 연구가 필요하다.

1.2.1 온톨러지 구축

ACQUILEX LKB(Lexical Knowledge Base)[7, 8]는 다국어 어휘 정보를 표현하기 위해서 고안되었다. 이 지식 베이스는 DGILE[9]과 LDOCE[10]에서 추출한 스페인어와 영어 어휘를 대응시켜서 구축하였다. 여기에서 각 어휘간 대응관계는 Tlink(Translation Link)로 표현하였다. TGE(Tlinks Generation Environment)[11]는 LKB의 구축을 위해서 Tlink를 생성하는 시스템이다. 이와 유사한 연구로, DGILE에서 추출한 스페인어 개념 분류법을 WordNet[12]의 유의어 집합과 자동으로 대응시키는 접근법도 있다[13]. 전술한 두 연구에 적용된 방법론은 사전에 명시된 정보를 이용해서 어휘를 대응시켰다. 이와 달리, 본 논문에서는 분류법의 계층정보와 참조정보를 이용해서 개념간 유사도를 계산해 내는 방법론을 시도하였다.

Sensus는 스페인어-영어 기계번역 시스템인 Pangloss[14]에서 사용하기 위해서 고안되었다. 이 온톨러지는 다양한 온라인 사전들과 의미망, 양국어(Bilingual) 자원들을 반자동으로 병합해서 구축되었다[15]. 기계번역용 온톨러지 구축에 대한 다른 연구[16]에서는, 일본어 어휘사전을 영어 온톨러지에 대응시키는 방법을 적용하기도 하였다. 또 다른 다국어(Multilingual) 온톨러지의 구축 사례로는, EDR(Japan Electronic Dictionary Research Institute) 전자사전[17]과 WordNet의 정련에 관한 몇 가지 연구들[18]이 있다. 또한 양국어 사전내의 스페인 어휘를 WordNet에 대응시킨 사례가 있다[19]. 이 대응 프로세스를 지원하는 방안으로 단국어(Monolingual) 기계가독형 사전(Machine Readable Dictionary : MRD)에서 유도된 분류법 구조의 사용법이 제안되었다[20].

한편 제약만족 알고리즘(Constraint Satisfaction Algorithm)을 이용한 다국어 계층구조의 대응에 관한 연구[21]에서는 분류법과 WordNet을 대응시키기 위해서 양국어 사전이 사용되었다. 이러한 대응 방법을 이용해서 WordNet 1.5의 명사부분을 WordNet 1.6에 투영시킨 사례[22]도 발표되었다. Factotum SemNet의 MRD 변환에 관한 연구[23, 24]에서는, 기계가 개연규칙을 반자동으로 발견하는데 활용할 목적으로 Semantic Network의 일종인 Factotum SemNet을 온라인화 하였다.

1.2.2 온톨러지 재사용

온톨러지 재사용 방법론은 크게 두 가지로 구분할 수 있다[5]. 온톨러지 병합(Ontologies Merging)과 온톨러지 집적(Ontologies Integration)이 그것이다 : (1) 온톨러지 병합이란 주제가 같거나 비슷한 온톨러지들을 단일화하는 방법이고, (2) 온톨러지 집적이란 다른 온톨러지들을 수집하거나 수정해서 하나로 만드는 방법을 말한다. 이 두 방법의 차이점은 원 온톨러지의 각 부분이 재사용 된 후에 여타 부분과 구분할 수 있느냐 여부이다. 전자는 각 부분을 서로 구분할 수 없는 반면에 후자는 재사용 후에도 원 온톨러지의 각 부분을 구분할 수 있다.

온톨러지 병합의 한 사례로는 ACP(Air Campaign Planning)[25] 온톨러지를 들 수 있다[5]. 이것은 공중 군사작전을 입안하는데 필요한 계획, 행위, 수행과정 등을 표현하기 위해서 ARPI(DARPA/Rome Planning Initiative)와 JFACC(Joint Force Air Component Commander)의 공중전 작전계획 수립용 어플리케이션에서 사용된다. 이 온톨러지는 공중전에 대한 전문용어를 SENSUS에 정련시킨 후 해당 용어에서 근 노드 사이에 있는 노드들을 취합하는 방식으로 구축되었다.

온톨러지 집적의 사례로는 레퍼런스 온톨러지(Reference Ontology)[26]의 경우를 들 수 있다. 이것은 이름에서 짐작할 수 있듯이 다른 온톨러지들을 표현하기 위해서 고안되었다. 이 온톨러지의 개발은 두 과정으로 구분할 수 있다.

첫째, 개념적인 구조와 주요개념, 분류법, 관계, 역할, 공리들을 정의한다. 이 과정에서는 METHONTOLOGY[27]와 ODE(Ontology Design Environment) [28]를 이용하였다.

둘째, 레퍼런스 온톨러지를 구성하게 될 특정 온톨러지들을 추가한다. 이 과정에서 $(KA)^2$ 온톨러지(Knowledge Acquisition Ontology)[29]를 재사용 하였다.

1.3 연구 내용 및 기여도

현재, 지능형 정보 시스템이 요구하는 분야 지식이 다양해졌다. 이에 비해서 개발된 온톨러지 자원이 부족한 실정이다[30]. 온톨러지를 새로 개발하는 데는 상당한 시간과 노력이 투입되어야 한다. 그러나 가용 시간과 노력은 제한적이기 마련이다. 이러한 이유로 온톨러지를 보다 용이하게 구축할 방법을 강구하게 되었다. 그 방법으로서 본 논문에서는 기존의 어휘자원과 온톨러지를 목적하는 바에 맞게 변환하여 재사용하는 범주 재분류 방법론을 개발하였다.

본 논문에서 개발한 범주 재분류 방법론을 통하여 (1) 온톨러지 범주정보를 단순화시킬 수 있고 또는 구체화시킬 수도 있다. 이 과정에서 온톨러지의 범주표제와 참조정보를 활용한다. 또한 (2) 상이한 범주들 사이의 관계를 밝히는 방법도 도출하였다. 마지막으로 (3) 이 방법론을 로젯 시소러스[31]에 적용한 세 가지 예[32, 35, 36]를 통하여 그 유용성을 살펴보았다.

본 논문에서 개발한 범주 재분류 방법론을 사용할 경우 다음 사항을 기대할 수 있다 : (1) 기존 온톨러지의 재사용성을 높일 수 있다. (2) 온톨러지에서 범주들 사이의 새로운 관계를 밝힐 수 있다. (3) 대응되는 범주를 기반으로 이종 온톨러지를 가상으로 통합하는 것을 도와준다. (4) 범용 지식을 매개로 분야 지식의 다양성을 증대시킬 수 있다.

2. 온톨러지 범주 재분류 방법론

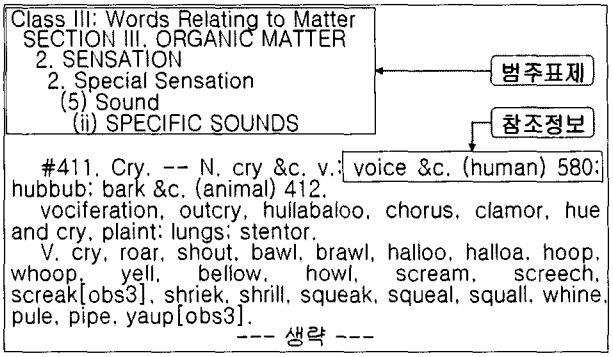
사람이 어떤 의미를 전달한다는 것은 언어를 이용하여 추상화된 개념을 전달한다는 것을 의미한다. 달리 말하여 전달하고자 하는 개념은 전달 과정에서 어휘로 표상 되며, 표상된 어휘를 상호 전달함으로써 정보전달의 목적을 성취한다. 결국은 정보전달 과정에 간여하는 중요한 요소로서 어휘를 들 수 있다. 이런 점에서 보면 견실한 의미 분류체계를 갖춘 유의어 사전인 로젯 시소러스[31]는 정보처리에 중요한 자원이라 할 수 있다. 본 논문에서는 주어진 정보를 범주화하고 그 범주들을 단순화 또는 구체화하는데 로젯 시소러스를 활용할 것이다.

본 논문에서는 기존 온톨러지를 재사용하는 방법으로서 온톨러지 범주정보의 단순화와 구체화를 생각하였다. 기존 온톨러지를 변환할 때 온톨러지의 범주표제와 참조정보를 활용할 수 있다. 온톨러지 변환 작업에서 범주 단순화와 범주 구체화는 서로 밀접한 관계를 가지고 있다 : (1) 범주 구체화는 주어진 범주 체계의 내용을 결정하는 것으로서, 그 과정에서 범주 단순화 방법을 활용할 수 있다. (2) 범주 단순화는 세분화된 기존 범주를 그 내용은 유지하되, 분류 체계를 주어진 목적에 맞추어 간략하게 만드는 것이다. 그리하여 주어진 분류 체계의 내용을 구체화할 수 있다. 한편 범주표제나 참조정보 역시 이와 같은 범주 재편성 과정에서 상호보완적으로 작용한다. 본 절에서는 세부적인 방법을 기술하기 앞서 OfN(Ontology for Narrative)[32] 구축과정을 통해 온톨러지 변환 방법론을 개관 하고자 한다.

2.1 로젯 시소러스

로젯 시소러스[31]는 의미 분류를 기초로 총6개의 강(Class)으로 구성되었다. 각 강은 하부에 부(Division), 과(Section) 등의 계층구조로 세분화되었다. 각 계층은 고유한 범주표제를 가지고 있으며 계층구조의 말단에는 총1044개의 범주(Category)가 존재한다. 각 범주에는 품사별로 유의어 목록이 나열되어 있다. 한편, 유의어 목록에서 특정 어휘가 다른 범주를 참조하는 경우에는 "어휘 &c.(표제어) 표제번호"의 형식으로 표현한다. 이를 참조정보라 하였다(그림 1). 본 논문에서는 이러한 범주표제와 참조정보를 범주 재분류에 활용하고자 한다.

범주 재분류 과정에서 로젯 시소러스의 원형을 그대로 이용하는 것은 무리이다. 따라서 로젯 시소러스의 원형을 기계가 사용할 수 있는 형태로 미리 전처리한 어휘사전 ROTIP (ROget's Thesaurus In Prolog)[33]을 이용할 것이다.



(그림 1) 로젯 시소러스

2.2 범주 재분류 개관

어휘사전 ROTIP의 범주체계를 온톨러지 OfN(Ontology for Narrative)의 체계에 대응시키기 위해서 ROTIP의 범주표제와 참조정보를 활용하였다. ROTIP의 범주표제는 계층적으로 붙여져 있고 관련범주 참조정보는 '->'로 표현되었다. 참조정보를 활용하기 위해서는, 우선 OfN의 각 범주에 대한 ROTIP 기저범주(Base Category)를 설정할 필요가 있었다. 이와 함께 ROTIP 범주 *Cr*의 OfN 범주 *Co*에 대한 대응선호도를 규정할 필요도 있었다. 그 이유는 다음과 같다 : (1) 일반적으로 ROTIP의 한 범주는 OfN의 각 범주와 정도는 차이가 있지만 서로 관련되어 있다. (2) 범주 *Cr*에 속하는 어휘들이 이야기에서 자주 쓰이는 뜻을 *Co*로 반영한다.

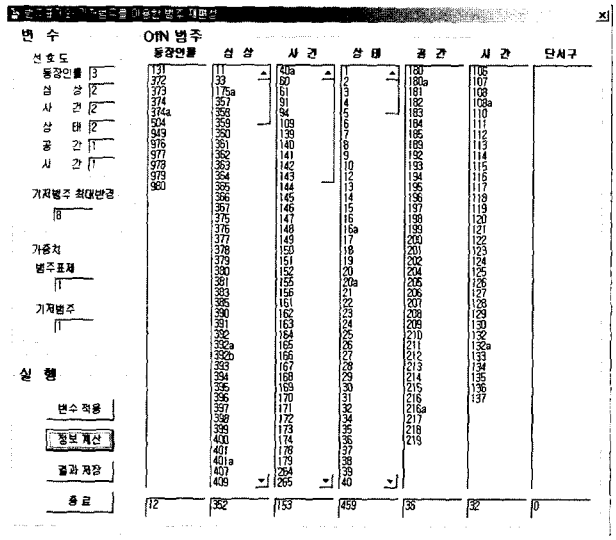
〈표 1〉 ROTIP - OfN 범주대응시 고려사항

분 류	세 부 사 항					
OfN 대응범주	등장인물 (Characters)	심상 (Affect State)	사건 (Event)	상태 (State)	공간* (Space)	시간* (Time)
ROTIP 기저범주	373, 374	821	151	7	181	106
OfN 대응 선호도	3	2	2	2	1	1
범주표제의 대응성	범주표제의 의미와 계층수준(세부범주가 나타나는 하위계층의 표제정보를 우선적으로 감안한다)					
기저범주에서의 거리	기저범주의 최대반경에 대한 보수 값(가까운 기저범주와 대응 가능성이 높다)					

<표 1>의 내용은 ROTIP-OfN 범주체계 대응시 고려해야 할 이와 같은 사항을 정리한 것이다. 참고로, OfN 범주에는 시간 및 공간이라는 범주는 없다(표중 * 참고). 그 대신, 시공의 변화(Delta-Time 및 Delta-Space)가 있다. 시공의 변화라는 범주는 문장추상화 단계에서 시간 및 공간이라는 범주에서 파생되는 것이다. 문장추상화가 실시되

기 전의 문장 구성성분은 시간이나 공간의 범주를 가질 수 있다. 따라서 시간 및 공간에 대한 범주를 시공의 변화라는 범주에 대한 원형범주 자격으로 편의상 다른 OfN 범주와 함께 고려한다.

ROTIP의 범주 *Cr*을 OfN 범주 *Co*에 대응시키는 절차는 다음과 같다 : (1) *Cr*의 계층적 범주표제 *Hc*의 OfN 대응성 *Mh*를 계산한다. (2) ROTIP에 설정해둔 OfN 기저범주에서 *Cr*까지의 참조경로 길이로써 *Cr*의 OfN 대응성 *Mp*를 계산한다. (3) *Mh*와 *Mp*의 가중평균치를 근거로 *Cr*에 대응하는 OfN 범주 *Co*를 결정한다. 이 절차를 구현한 프로그램이 (그림 2)와 같은 결과를 보였다. 그림에서 OfN 범주에 대응하는 ROTIP 범주번호들과 함께 범주개수를 확인할 수 있다. 구현의 세부내용은 다음 두 개의 절에서 논의하기로 한다.



(그림 2) ROTIP-OfN 범주체계 대응 결과

'411'=>cry:[ii:[specific, sounds], 5:sound, 2:[special, sensation], 2:sensation, [section, iii]:[organic, matter], [class, iii]:[words, relating, to, matter]].
↓ Prolog 술어표현으로 변환
'411'=>[cry, sound(specific), sound, sensation(special), sensation, matter(organic), word(matter)].
↓ OfN 범주로 변환
'411'=>[cry, state, state, affect_state, affect_state, state, state].
↓ 계층성과 선호도를 감안하여 대응하는 OfN 범주를 결정
'411'=>state.

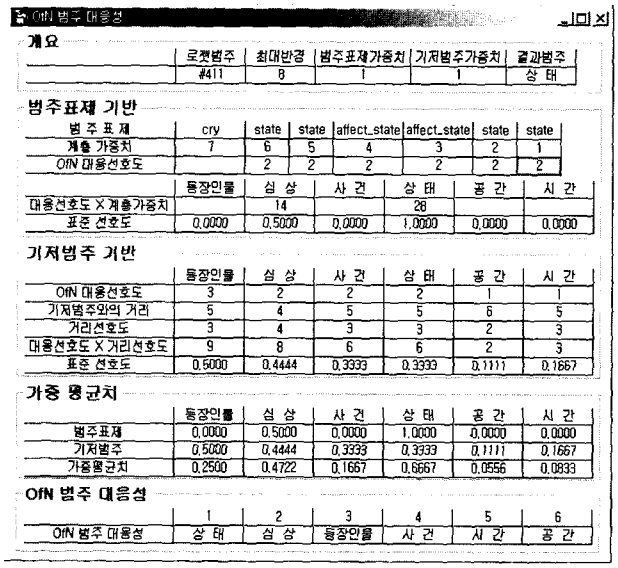
(그림 3) '#411. Cry'의 OfN 범주 변환과정(범주표제 기반)

2.2.1 ROTIP-OfN 범주대응 : 범주표제 기반

ROTIP 범주 *Cr*에는 계층적 범주표제 *Hc*가 붙어 있다. *Hc*의 의미와 계층성을 기반으로 *Cr*과 OfN 범주 *Co*의 대응가능성 *Mh*을 계산한다. 그 절차는 (그림 3)에 나타나 있다 : (1) ROTIP의 범주표제를 Prolog 술어로 표현한다. 이때 강(Class), 부(Division), 과(Section) 등과 같은 계층 식별자는 생략한다. (2) 표제를 구성하고 있는 Prolog 술어를 OfN 범주명으로 변환한다. 이때 사용하는 변환표의 일부를 <표 2>에 보였다. (3) *Hc*의 계층성과 OfN 대응선호도를 감안하여 OfN 범주를 결정한다.

〈표 2〉 표제 구성 술어 - OfN 범주명 변환표 (일부)

표제 구성 술어	OfN 범주명
act	event
being	state
dimension	space*
future	time*
sensation	affect_state
sound	state
affection(general)	affect_state
matter(organic)	state
sensation(special)	affect_state
sound(specific)	state
word(matter)	state
word(sentient, power(moral))	affect_state



(그림 4) ROTIP-OfN 범주체계 대응 상세출력

범주표제는 하위계층으로 갈수록 범주특유의 속성을 보다 명확하게 반영한다. 따라서 OfN 범주명으로 변환된 표제에서, 범주표제에 대한 계층 가중치(Weight)는

최상위 것을 1이라고 하고 순차적으로 하위계층에 대한 가중치를 부여하였다. 이에 대한 예를 <표 3>에서 볼 수 있다. 로젯범주 '#411. Cry'의 경우, 변환된 범주표제에는 state와 affect_state 만이 나타나 있다. 이들에 대한 범주표제 기반의 표준 선호도 *Mh*를 각각 계산한 결과가 <표 3>에 보인다. 표의 내용을 (그림 4)의 '범주표제 기반' 부분에서 다시 확인할 수 있다. 그림의 내용은 ROTIP-OfN 범주체계 대응을 구현한 프로그램의 출력이다.

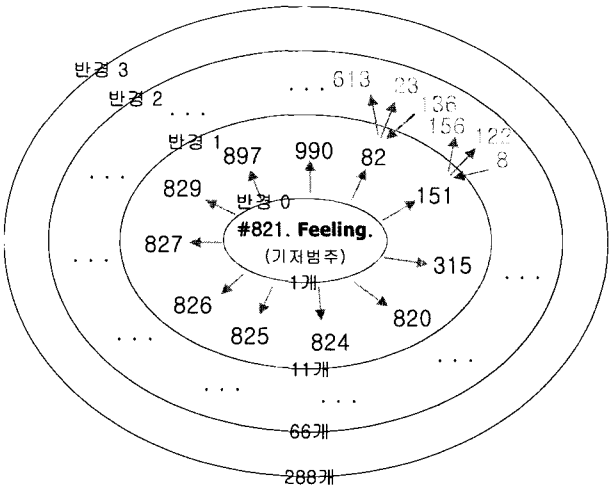
〈표 3〉 '#411. Cry'의 OfN 범주 변환과정 (범주표제 기반)

범주표제		cry	state	state	affect state	affect state	state	state
범주표제 계층 가중치		7	6	5	4	3	2	1
OfN대응선호도		0(없음)	2	2	2	2	2	2
범주표제 기반의 표준선호도	state	(6*2 + 5*2 + 2*2 + 1*2) / (1 + 2 + 3 + 4 + 5 + 6 + 7) = 28 / 28 = 1.00						
	affect state	(4*2 + 3*2) / (1 + 2 + 3 + 4 + 5 + 6 + 7) = 14 / 28 = 0.50						

2.2.2 ROTIP-OfN 범주대응 : 기저범주 기반

<표 3>과 같이 OfN 기저범주를 ROTIP 안에 미리 설정해 두었다. (그림 5)는 심상(Affect State)에 대한 기저범주 821에서 반경값을 1씩 증가시켜 다른 범주에 도달하는 모습을 보인다. 기저범주 안에 있는 참조정보를 이용하여 기저범주로부터 1만큼 떨어진 곳에 있는 범주를 밝혀낸다. 이와 아울러 기저범주를 가리키는 참조정보가 있는 범주도 기저범주와 거리로 1만큼 떨어져 있다고 정한다. 이렇게 해서 밝혀낸 범주들이 기저범주를 중심으로 반경 1안에 존재하는 범주들이 된다. 일반적으로 기저범주를 중심으로 반경(i + 1) 안에 있는 범주는, (1) 반경 i 밖에 있는 범주로서, (2) 반경 i 안에 존재하는 참조정보로 도달할 수 있는 범주이거나 또는 (3) 반경 i 안의 범주를 가리키는 참조정보를 가진 범주들로서 구성된다.

각 기저범주에서 ROTIP 범주 *Cr*까지의 참조경로 길이로써 *Cr*의 OfN 범주 대응성 *Mp*를 계산한다. 구체적인 계산과정은 <표 4>에 정리하였다 : (1) ROTIP 범주 *Cr*이 각 기저범주에서 얼마나 떨어져 있는가를 계산한다. 이 값을 *Lp*라고 하자. (2) OfN 범주에 각각에 대하여 *Cr*이 가지는 거리 선호도를 (8-*Lp*)로 정한다. 이렇게 하면 거리 선호도가 기저범주와 가까운 범주일수록 그 값이 커진다. 8은 ROTIP의 참조경로 최대길이이다. 각 기저범주에 대해서 확인한 바, 9이상의 길이를 가진 참조경로로써 새로운 범주를 탐색할 수 없었다. (3) OfN 대응 선호도와 거리 선호도 (8-*Lp*)를 곱한 다음, (4) 거리 선호도의 합으로 나누어 정규화 시킨다. 이 값이 *Cr*의 *Mp*가 된다.



(그림 5) 참조정보를 이용한 범주탐색

〈표 4〉 '#411. Cry'의 OfN 범주 변환과정 (기저범주 기반)

분 류	세 부 사 항					
OfN 대응범주	등장인물 (Character)	심상 (Affect State)	사건 (Event)	상태 (State)	공간* (Space)	시간* (Time)
OfN 대응선호도	3	2	2	2	1	1
ROTIP 기저범주	373, 374	821	151	7	181	106
기저범주에서 떨어진 거리	5	4	5	5	6	5
거리 선호도	3 (=8-5)	4 (=8-4)	3 (=8-5)	3 (=8-5)	2 (=8-6)	3 (=8-5)
기저범주기반 표준선호도	0.50 (=3*3/18)	0.44 (=2*4/18)	0.33 (=2*3/18)	0.33 (=2*3/18)	0.11 (=1*2/18)	0.17 (=1*3/18)

2.2.3 ROTIP-OfN 범주대응 : 가중평균

ROTIP의 범주 *Cr*을 OfN 범주 *Co*에 대응시키기 위하여 (1) 범주표제 기반의 대응성 *Mh*와 (2) 기저범주 기반의 대응성 *Mp*를 각각 계산하였다. 이제 *Mh*와 *Mp*의 가중평균치를 근거로 *Cr*에 대응하는 OfN 범주 *Co*를 결정할 수 있다. 전술한 바 있는 구현 시스템으로 실험한 결과, 가중비율은 (*Mh* : *Mp* = 1 : 1)로 정하는 것이 적절함을 알았다. 이것은 ROTIP의 범주 *Cr*을 OfN 범주 *Co*에 대응시킬 때, 범주표제의 계층성과 참조정보가 가지는 비계층성을 동등하게 고려함을 의미한다. 이로써 로젯 시소러스가 가지는 계층적 및 비계층적 어휘 분류체계를 균형 있게 감안하여 새로운 온톨러지 OfN의 분류체계를 얻게 되었다.

<표 5>는 로젯범주 '#411. Cry'의 OfN 범주변환 마지막 단계를 나타내고 있다. 이 경우, 상태(State)에 대한 가중평균값이 최대이다. 따라서 로젯범주 '#411. Cry'는 OfN 범주의 상태(State)에 대응된다고 할 수 있다.

〈표 5〉 로젯범주 '#411. Cry'의 OfN 범주 변환과정 (가중평균)

분 류	세 부 사 항					
OfN 대응범주	등장인물 (Character)	심상 (Affect State)	사건 (Event)	상태 (State)	공간* (Space)	시간* (Time)
범주표제 기반의 표준선호도	0.00	0.50	0.00	1.0	0.00	0.00
기저범주 기반의 표준선호도	0.50 (=3*3/18)	0.44 (=2*4/18)	0.33 (=2*3/18)	0.33 (=2*3/18)	0.11 (=1*2/18)	0.17 (=1*3/18)
가중평균치	0.25	0.47	0.17	0.67	0.06	0.09

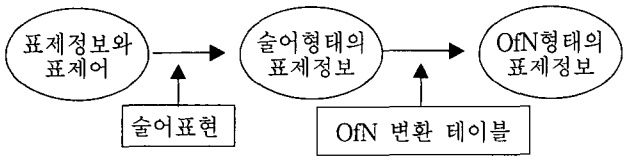
2.3 범주 단순화

범주 단순화란 여러 개의 범주들을 의미의 유사성이나 시소러스의 분류규칙을 감안하여 하나의 범주로 통합하는 것을 말한다. 범주 단순화를 통해서 복수개의 범주들에 대한 범주대표를 발견할 수 있다. 또한 개별적인 어휘정보 혹은 범주집합을 하나의 범주로 결합시키거나 각 범주들에 대하여 새로운 관계를 맺어줄 수 있다.

2.3.1 범주표제를 이용한 단순화

로젯 시소러스는 그 구조가 의미체계에 기초한 계층적 분류체제를 따르고 있다. 또한 각 계층에는 고유한 범주표제가 존재한다. 범주표제는 해당 계층 및 하부 계층에 대한 함축적인 정보를 가지고 있다. 이 범주표제를 범주 단순화에 이용하는 방법을 생각해 보았다.

범주표제를 이용한 범주 단순화는 단순화 범주 *Co*와 범주표제 *Hc* 사이의 어휘적 밀접도를 바탕으로 단순화하는 방법이다. 즉, *Co*의 각 범주들에 대한 *Hc*와의 어휘 밀접도를 조사한 후 ROTIP의 범주 *Cr*을 밀접도가 가장 높은 *Co*에 단순화시킨다. 이를 위해 (그림 6)과 같은 과정으로 *Co*를 묘사하는 어휘를 참고로 *Hc*의 형태를 바꾼다. 그 후 *Co*를 묘사하는 어휘를 범주표제와 ROTIP 본문에서 검색한 후 *Cr*과 *Co*의 대응성 *Mh*를 구한다. 이 *Mh*를 비교해서 단순화 범주를 결정한다.



(그림 6) 범주표제 변환 과정

범주 단순화의 개략적인 과정은 앞 절의 ROTIP-OfN 대응 예에서 이미 살펴보았다. (그림 3)은 로젯범주 '#411. Cry'에 대한 (그림 6)의 처리 과정을 나타내고 있다. 이 과정에서는 미리 준비한 범주 변환표(<표 2>)가 필요하다. 이러한 범주 단순화 과정을 알고리즘으로 표현하면 (그림 7)과 같다.

```
// 주요 용어
Co      // 단순화 범주
Wi      // 단순화 범주를 묘사하는 어휘
Tc      // 범주 묘사 어휘와 Co를 대응시킨 테이블
Cr      // 로젯 시소러스의 범주
Hc      // 로젯 시소러스의 범주표제
Hp      // Prolog 술어 형태로 변형시킨 범주표제
Ho      // 단순화 범주 형태로 변형된 범주표제
Mh      // 로젯 범주에 대한 단순화 범주의 범주표제 대응성
Hw      // 범주표제 계층 가중치(범주표제의 특수성과 일반성)
Mc      // 대응선호도(단순화 범주 사이의 우선순위)


// 범주표제를 변형시키는 과정
repeat (모든 Hp에 대해서)
  if Hp에 Wi가 포함되어 있다 then
    Wi에 대응하는 Co를 Tc에서 찾는다.
    Wi를 Co로 대치한다.
  end if // 술어 표제정보가 단순화 범주 형태로 변형된다.
  Ho = Wi가 Co로 치환된 Hp
end repeat

// 로젯 범주와 단순화 범주의 대응
repeat (모든 Ho에 대해서)
  // 범주 대응성 결정
  repeat (Ho의 모든 표제에 대해서)
    Hco = Ho내의 Co 개수
  end repeat
  Mho = Hco × Hwo
    // Mho, Hco, Hwo의 첨자 o는 Co의 각 범주에 해당

  // 로젯 범주를 단순화 범주에 종속시키는 과정
  if (가장 큰 Mho가 하나 존재한다.) then
    repeat (Mho)
      가장 큰 Mho를 찾는다.
    end repeat
    가장 큰 Mho를 갖는 Cr을 Co에 대응시킨다.
  else
    repeat (동일 값의 Mho)
      Mho간의 Mc를 비교한다.
      Mc가 큰 Co를 찾는다.
    end repeat
      Cr을 Mc가 큰 Co에 대응시킨다.
  end if
end repeat
```

(그림 7) 범주표제를 이용한 단순화 알고리즘

〈표 6〉 범주표제의 계층 가중치와 대응 선호도

범주표제 (H_c)	cry	specific, sounds	sound	special, sensation	sensation	organic, matter	words, relating, to, matter
범주표제 (H_o)		범주$_4$	범주$_1$	범주$_2$	범주$_2$	범주$_4$	범주$_1$
대응선호도 (M_c)	0	3	3	6	6	3	3
계층가중치 (H_w)	7	6	5	4	3	2	1

(그림 7)의 범주 단순화 알고리즘을 범주 단순화 예와 함께 자세히 살펴보고자 한다. 범주표제에 대한 수치화 과정을 살펴보면, 범주표제는 하위계층으로 갈수록 범주특유의 속성을 보다 명확하게 반영한다. 따라서 단순화 범주로 변환된 표제에서, 범주표제에 대한 계층 가중치(Weight)는 최상위 것을 1이라고 하고 순차적으로 하위계층에 대한 가중치를 부여하였다. 이 계층 가중치를 *Hw*로 표현한다.

한편, 단순화 범주 사이에도 중요도의 차이가 있을 것이

다. 이 경우 중요한 정도를 대응 선호도 *Mc*로 표현하였다. 대응 선호도는 우선 순위가 높을수록 큰 값을, 우선 순위가 낮을수록 작은 값을 지정하였다. 이러한 계층 가중치과 대응 선호도를 곱한 수치가 범주표제의 대응성 *Mh*이다.

상기 내용을 앞 절의 ROTIP-OfN 대응의 경우와 비슷한 구체적인 예를 들어 살펴보고자 한다. 이 예에서는 단순화 범주 *Co*를 "범주1, 범주2, 범주3, 범주4, 범주5, 범주6" 총 여섯 개의 범주로 설정하였다. 각 범주의 대응 선호도 *Mc*는 "범주1 : 7, 범주2 : 6, 범주3 : 5, 범주4 : 3, 범주5 : 2, 범주6 : 1"로 설정했다. 또한, 처리대상인 로젯 범주는 "#411. Cry."를 예로 들었다. 해당 로젯 범주의 범주표제 *Hc*는 말단 범주까지 총7단계의 계층으로 구성되었다. 따라서 7개의 범주표제를 가지고 있다. 이 범주표제를 범주 묘사 어휘를 통해 가능한 단순화 범주와 대응시킨다. 그 결과 한개의 범주표제를 제외하고 나머지 여섯 개의 범주표제를 단순화 범주로 대응시킬 수 있었다-"cry, 범주4, 범주4, 범주2, 범주2, 범주4, 범주4"-. 나열 순서는 전방이 하위계층이며 후방이 상위계층이다<표 6>.

단순화 범주를 결정하기 위해서 로젯 범주에 대한 단순화 범주의 대응성을 정규화 하는 것이 바람직 할 것이다. <표 6>을 범주표제의 계층 가중치 *Hw*에 대한 대응 선호도 *Mc*로 <표 7>과 같이 달리 표현할 수 있다.

〈표 7〉 계층 가중치에 대한 대응 선호도

H_o / C_o	계층 가중치(H_w)						
	7	6	5	4	3	2	1
범주$_1$							
범주$_2$				6	6		
범주$_3$							
범주$_4$	3	3	3			3	3
범주$_5$							
범주$_6$							

수치화 과정의 첫 번째 단계로 <표 7>을 행렬로 변환시킨다.

$$H=\begin{pmatrix}0&0&0&0&0&0&0\\0&0&0&6&6&0&0\\0&0&0&0&0&0&0\\0&3&3&0&0&3&3\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{pmatrix}$$

[식 1] 계층 가중치에 대한 대응 선호도

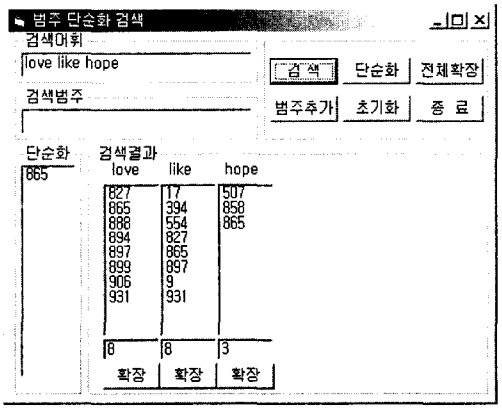
계층 가중치 *Hw*에 대한 대응 선호도 행렬을 *H*라 할 때 행렬 *H*의 원소 *aij*의 첨자 *i*는 단순화 범주 *Co*의 각 범주이고 *j*는 계층 가중치 *Hw*이다. 즉, 범주표제의 계층수준이다. 따라서 원소 *aij*는 단순화 범주의 대응 선호도 *Mc*이다. 이 때, 각 단순화 범주의 범주표제 대응성 *Mh*는 계층 가중치 *Hw*와 대응 선호도 *Mc*의 곱에 대한 합이라고 볼 수 있다. 최종적으로 범주표제 *Hc*의 계층수준 개수를 l이라 하면 단순화 범주 대응성 *Mh*은 다음과 같이 표현할 수 있다.

$$Mh_i=\sum_{j=1}^{l} a_{ij}$$

[식 2] 단순화 범주 값

하나의 로젯 범주에 대해서 범주표제의 대응성을 [식 2]를 통해 구한다. 그 중 가장 큰 값을 찾아서 그에 해당하는 단순화 범주로 결정한다. 이 과정을 모든 로젯 범주에 적용하면 로젯 범주를 단순화시킬 수 있다. 경우에 따라서 단순화 범주 값이 동일할 수 있다. 이 때에는 단순화 범주의 대응 선호도 *Mc*를 따른다.

단순화 방법중 범주 묘사 어휘 *Wi*를 범주표제와 시소러스 본문에서 검색할 수 있는 검색 인터페이스를 구현하였다. (그림 8)에서 보듯이 복수의 어휘를 동시에 검색하면 각 어휘에 해당하는 로젯 범주를 검색한다. 또한 검색된 범주들을 바탕으로 단순화 범주를 계산한다. 만일 검색된 범주들에서 단순화 범주를 추출할 수 없다면 범주 묘사 어휘에 대해서 개별적 혹은 일괄적으로 참조 반경을 확장할 수도 있다. 이 검색 인터페이스는 다음절에서 살펴볼 참조정보를 이용한 범주 단순화에도 활용할 수 있도록 설계되었다.



(그림 8) 범주 단순화 검색 인터페이스

이렇게 단순화한 범주는 모든 범주가 서로 중복되지 않는 장점이 있다. 또한 이 범주는 로젯 시소러스의 범주 분류 취지를 그대로 계승한다. 한편 범주표제가 해당 범주의 유의어 집합을 모두 대변할 수는 없다. 또한 범주에 나타나는 어휘의 다의성도 고려되어야 할 것이다. 따라서 다른 각도의 범주 단순화 방법을 모색할 필요가 있다.

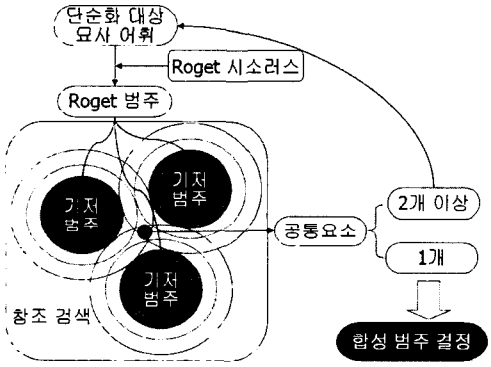
2.3.2 기저범주를 이용한 단순화

다른 사전들처럼 로젯 시소러스에서도 어휘에 대한 부가적인 설명이나 참조가 필요한 경우, 어휘를 다른 표제어로 참조시킨다. 이런 점을 범주단위로 확장해서 생각해 보면, 참조하는 범주사이가 그렇지 않은 경우에 비해 어휘적 밀접도가 높다고 볼 수 있다. 이 점을 활용해서 범주 재분류에

응용하고자 한다. 바꿔 말해서 참조관계들의 연결인 참조 네트워크를 특정 기준으로 걸러 내거나, 참조들 사이에 관계를 맺은 후 재구성하면 범주 단순화 또는 구체화가 가능하다.

구체적인 방법은 우선 단순화 범주와 가장 밀접한 로젯 범주를 어휘적 밀접도를 고려해서 선정한다. 이 범주를 기저범주라 하고, 참조 탐색의 출발점으로 이용될 것이다. 이 기저범주의 반경 값을 0으로 정한다. 그 후 참조정보를 이용하여 이 기저범주가 참조하는 범주들과 이 기저범주를 참조하는 범주들을 검색한다. 이 과정에서 검색된 범주들의 반경 값은 1로 정한다. 같은 방식으로 계속 반경을 확장한다(그림 5).

만일 반경을 7까지 확장하면 모든 로젯 범주가 검색될 것이다. 한 범주에 대해서 모든 로젯 범주가 검색된다면 의미가 없기 때문에 적절한 반경을 결정해야 할 것이다. 그 한가지 예로 검색된 로젯 범주가 전체의 절반을 넘지 않는 경우를 들 수 있을 것이다. 실험을 통해서 반경을 3으로 조절하면 전체의 절반을 넘지 않는 범위에서 로젯 범주가 검색됨을 알 수 있었다. (그림 5)는 로젯 범주 "#821. Feeling."을 기저범주로 정하여 참조정보를 검색한 경우이다. 반경을 3까지 확장했을 때 366개의 로젯 범주가 검색되었다.



(그림 9) 기저범주를 이용한 단순화 과정

기저범주를 이용한 단순화의 경우에는 우선 단순화 대상인 복수의 정보 혹은 범주들에 대해서 기저범주를 결정한다. 결정된 각 기저범주의 참조 반경을 확장한다. 각 참조 반경들을 집합으로 가정하고 집합들 사이의 공통요소를 취한다. 이 공통요소가 복수의 정보 혹은 범주에 대한 단순화 결과이다. 만일 단순화 결과가 복수인 경우에는 순환적 관계를 보이지 않는 범위에서 그 결과를 대상으로 다시 단순화시킨다. 이 과정을 (그림 9)에 나타내었다.

상기 내용을 구체적으로 살펴보면, 우선 범주표제를 이용한 범주 단순화의 경우와 마찬가지로 기저범주에 대한 거리 선호도를 설정할 수 있다. 거리 선호도는 기저범주와 단순화 범주 Cr과의 거리 Lp에 영향을 받는다. 기저범주의 반경은 0이고 이 기저범주와 가까운 Cr일수록 작은 반경 값을 갖는다. 그러므로 거리 선호도는 반경이 작을수록 크게 적용되어야 할 것이다. 반경을 7까지 확장하면 모든 로젯 범주를 포함할 수 있으므로 반경 값에 대한 8의 보수가 기저범주의 거리 선호도($Lc = 8-Lp$)이다. 기저범주에 대한 단순화 범주의 대응성 Mp는 단순화 범주에 지정한 대응 선호도 Mc와 거리 선호도 Lc의 곱이다. 이상을 <표 8>에 나타내었다.

〈표 8〉 로젯 범주 '#411. cry.'에 대한 기저범주

분 류	단순화범주 (C_o)	대응선호도 (M_c)	기저범주에서 떨어진 거리(L_p)	거리선호도 (L_c)	비 고
Major	범주1	7	5	3	참조정보의 선호도는 각 참조정보 반경의 최대반경에 대한 보수
	범주2	6	4	4	
	범주3	5	5	3	
Minor	범주4	3	5	3	
	범주5	2	6	2	
	범주6	1	5	3	

참조정보의 거리 선호도를 Lc이라 하고 대응 선호도를 Mc라 하면 단순화 범주 Co의 각 범주 i에 대한 기저범주의 대응성 Mp는 다음 식으로 표현할 수 있다.

$$Mp_i = Lc_i \times Mc_i \qquad [식\ 3]$$

하나의 로젯 범주 Cr에 대해서 각 단순화 범주 Co와의 대응성 Mp를 [식 3]을 통해 구한다. 그 중 가장 큰 값을 찾아서 그에 해당하는 단순화 범주로 결정한다. 이 과정을 모든 로젯 범주에 적용하면 로젯 범주를 단순화시킬 수 있다.

이러한 범주 단순화 방법의 유용성을 예를 통해 살펴보고자 한다. "유식학 선심소 분석"을 위한 범주 단순화 결과[36]를 보면, 기저범주를 이용한 단순화 결과와 사람의 인지적 판단에 의한 단순화 결과가 근소한 차이를 보이고 있다. 총11가지 단순화 범주에 대해서 범주 단위로 일치하는 경우가 4개 범주였다. 이외에 의미상 유사한 범주를 결정한 경우도 있었고 나머지 범주들에 대해서도 로젯 범주 내에서 근소한 차이를 보였다. 이 점은 로젯 시소러스의 구조를 감안하면 기저범주를 이용한 단순화 방법이 유용함을 시사하고 있다.

2.4 범주 구체화

범주 구체화란 하나 또는 그 이상의 범주를 시소러스의 구조와 의미를 유지하면서 범주 재분류 목적에 맞게 분해하는 것을 말한다. 이를 통해 시소러스 내에 함축된 새로운 범주구성요소들을 밝혀낼 수 있다. 이 절에서는 ROTIP를 활용하여 범주 단순화 방법에서처럼 두 가지 구체화 방법을 모색하였다. 첫 번째는 로젯 시소러스의 표제정보를 이용하는 방법이고 두 번째는 참조정보를 탐색하여 구체화하는 방법이다.

2.4.1 범주표제를 이용한 구체화

비성형적(Informal) 온톨러지의 범주를 구체화시킬 때 기존의 온톨러지 범주 분류체계를 활용할 수 있다. 우선 구체화시킬 범주에 부합하는 기존 온톨러지의 범주들을 탐색한다. 이 결과를 비성형적 온톨러지의 해당 범주에 귀속시킨다. 이러한 과정을 반복하면 비성형적 온톨러지가 기존의 온톨러지를 통하여 구체화된다. 이점에 착안하여 로젯 시소러스의 범주표제를 이용해서 범주 구체화를 시도하였다. 구체적인 방법은 범주표제를 이용한 단순화의 경우를 따른다.

2.4.2 기저범주를 이용한 구체화

로젯 시소러스의 범주 간 참조정보를 탐색하기 위해서 구체화 대상범주의 각 어휘를 선정한다. 대상 범주를 로젯 범주에 투영시키기 위해 선정한 어휘에 대응하는 로젯 범주를 기저범주로 정한다. 이 기저범주를 결정하기 위하여 로젯 시소러스의 범주표제를 이용하여 탐색한다. 이 경우에 범주표제에서 탐색할 수 없다면 시소러스 본문 중에 등장하는 범주들의 범주표제를 참고하여 택일한다. 결국 각 기저범주는 대상 범주를 가장 잘 나타내는 범주이다. 이 기저범주들과 나머지 로젯 범주 사이의 참조관계를 근거로 세부 범주들을 유도해낼 것이다. 이 경우 역시 구체적인 방법은 참조정보를 이용한 단순화의 경우를 따른다.

구체화 과정 중 참조범주를 검색할 수 있는 검색 인터페이스를 구현하였다. 이는 기저범주를 결정하고 참조 검색하고자 하는 반경을 설정하면 지정한 반경까지 기저범주를 검색할 수 있다.

이러한 범주 재분류 방법은 로젯 시소러스의 내재적 구조를 밝힌 점에서 의미가 있다. 또한 범주의 반경을 정함으로서 해당 범주가 기저범주와 얼마나 밀접한지를 알 수 있다. 반경 정보는 다른 범주간의 우선순위 결정에도 이용할 수 있다. 한편 참조정보가 잘못된 경우 잘못된 정보가 파생시키는 오류의 범위가 크다. 또한 반경이 커질수록 범주들 사이의 교차 참조가 빈번하다. 그러므로 단순화시키고자 하는 목적에 맞게 탐색반경을 한정시킬 필요가 있다.

2.5 범주표제·기저범주를 이용한 범주 재분류

범주 재분류 방법으로 제시한 범주표제 재분류 방법과 기저범주 재분류 방법은 각각 장단점을 가지고 있다. 결국 한 가지 범주 재분류 방법보다는 각각의 장점을 취할 수 있는 방법이 필요할 것이다. 그 방법으로서 범주표제와 기저범주의 대응성을 수치적으로 결합하는 방식을 택하였다.

범주표제와 기저범주는 서로 다른 부류의 값이다. 따라서 이 두 수치를 결합하기 위해서 표준 선호도를 산출한다. 우선 범주표제의 계층 가중치는 범주표제의 계층수준이다. 이 계층 가중치를 Hw라 하면 범주표제 대응성 Mh의 표준 선호도는 Mhu는 다음 식으로 구할 수 있다.

$$Mh_{ui} = \frac{Mh_i}{\text{계층가중치}(Hw)\text{의 합}} = \frac{Mh_i}{\text{등차수열 } Hw\text{의합}} \quad [식\ 4]$$

마찬가지로 기저범주 대응성 Mp의 표준 선호도는

$$Mp_{ui} = \frac{Mp_i}{\text{거리선호도}(Lc)\text{의 합}} = \frac{Mp_i}{\text{등차수열 } Lc\text{의합}} \quad [식\ 5]$$

이다.

최종적으로 범주 재분류에 사용될 값은 범주표제의 표준 선호도와 기저범주의 표준 선호도의 평균이다.

$$S_i = \frac{Mh_{ui} + Mp_{ui}}{2} \quad [식\ 6]$$

경우에 따라서 범주표제의 대응성과 기저범주 대응성중 어느 한쪽에 더 큰 비중을 줄 수 있을 것이다. 이 경우에는 각 표준 선호도에 범주표제와 기저범주의 중요도를 고려해서 가중 산술 평균을 구한다.

범주표제 대응성에 가중치 2를 설정하고 기저범주 대응성에 가중치 1을 설정한다면 범주표제와 기저범주의 가중 산술 평균은 다음과 같다.

$$S_i = \frac{Mh_{ui} \times 2 + Mp_{ui} \times 1}{3} \quad [식\ 7]$$

각 i에 대한 Si의 값을 구한 후 가장 큰 Si를 보이는 범주를 해당 로젯 범주의 재분류 범주로 결정한다.

2.6 범주 재분류 결정 인터페이스

범주 단순화 및 구체화의 경우에 계산양이 방대하기 때문에 범주 결정의 효율을 높이고 자동화하기 위하여 범주 재분류 결정 시스템을 구현하였다(그림 2), (그림 4), (그림 8). 이 시스템은 재분류 대상 범주의 범주 묘사 어휘 결정을 도우며 선정한 범주 묘사어휘로부터 범주표제와 기저참조의 대응성을 산출한다. 또한 기저범주와 재분류 범주사이의 참조 경로 파악이 가능하며, 복수의 범주에 대한 공통 범주 산출이 가능하다. 산출한 범주표제와 기저참조 대응성에 대한 가중평균치를 산출하고, 그 값에 따라 재분류 범주를 결정하는 전 과정을 처리한다. 또한, 범주 결정에 변인으로 작용하는 재분류 범주의 대응 선호도와 참조정보의 최대 반경 값, 범주표제와 기저범주의 중요도를 조절하는 가중치를 변경할 수 있게 설계하였다. 마지막으로 계산과정을 검토할 수 있도록 재분류 범주 결정 과정의 상세 정보를 출력하도록 하였다.

3. 범주 재분류 방법론 적용예

지금까지 살펴본 로젯 범주 정보를 이용한 범주 단순화 및 구체화 방법을 문장추상화를 위한 온톨러지와 게임의 흡인요소분석 등에 활용해 보았다.

3.1 문장추상화를 위한 온톨러지 구축

설화문장을 추상화시키는데 사용할 목적으로 일곱 가지 범주[34]로 구성된 온톨러지를 재구성하였다. 재구성에는 로젯 시소러스를 범주 단순화하는 방법을 이용하였다. 이 온톨러지의 범주는 다음과 같은 7가지 유형이 포함되어 있다 : (1) 등장인물, (2) 심상, (3) 사건, (4) 상태, (5) 공간, (6) 시간, (7) 담화 표지.

온톨러지를 재분류하기 위해서 로젯 시소러스 범주표제와 기저범주의 값을 산출하였다. 산출된 각 범주 값의 가중치 평균을 근거로 재분류 범주를 결정하였다. 각각의 범주 값을 병합하는 과정에서는 범주표제와 기저범주의 중요도를 조절하기 위해 가중치를 적용하였다. 이러한 과정을 통해서 설화문장의 추상화를 위한 로젯 시소러스 단순화 재분류 결과를 얻었다[32].

3.2 멀티미디어 게임 흡인력 분석

인지 및 감성을 고려한 22가지 흡인요소를 로젯 시소러스의 범주정보를 바탕으로 22가지 범주로 분류하였다. 이 범주를 바탕으로 기저범주를 탐색하여 가까운 반경의 범주를 병합하였다. 이 22가지 범주를 구체화시켜서 207가지 범주들을 찾을 수 있었다. 이 범주들은 22가지 흡인요소와는 다른 세분화된 흡인요소들이다. 이 요소는 개별적이고 독립적인 흡인요소로도 존재할 수도 있으며, 여러 가지 요소가 함께 할 때 더 강한 흡인력으로 작용하기도 한다[35].

3.3 선심소와 게임 흡인요소 대응

유식학의 선심소와 게임 흡인요소간의 대응관계를 범주 구체화를 응용하여 밝혀 보았다. 우선 게임의 22가지 흡인요소와 11종의 선심소에 대한 범주 재분류를 시행하였다. 이 범주 집합들의 반경을 각기 달리하여 3가지 경우로 나누어 대응시켜 보았다. 그 결과 흡인요소 범주의 반경이 0, 1, 2이고 선심소 범주의 반경이 0, 1, 2일때 두 범주간의 반경이 적절하게 유지되어 가장 만족할 만한 결과를 얻을 수 있었다[36].

4. 결론 및 향후연구

본 논문에서는 자연어 처리 분야와 지식기반 시스템에서 요구하는 온톨러지 자원의 개발 방안으로 기존 자원의 재사용 방법론을 제안하였다. 그 구체적인 방안은 범주 재분류이다. 재사용할 온톨러지로서 전처리된 로젯 시소러스를 이용하였다. 재분류 방법으로는 범주정보 단순화와 구체화를 고안하였다. 각각은 다시 범주표제와 기저범주를 이용한 방법으로 나누어 생각하였다. 또한 이 과정의 자동화에 사용할 범주 결정 시스템을 구현하였다.

본 논문에서 고안한 범주 재분류 방법을 활용한 실험에서 확인할 수 있었던 장점은 다음과 같다 : (1) 범주 단순화 방법은 복수의 정보 또는 범주에 대한 대표성을 발견할 수 있게 한다. 또한 개별적인 정보 혹은 범주집합을 하나의 범주로 통합할 수 있게 한다. (2) 범주 구체화 방법은 하나 혹은 그 이상의 범주를 의미와 구조는 유지시키면서 원하는 목적에 맞게 분할 할 수 있게 한다. 이를 통해 범주를 구성하는 새로운 요소들을 발견할 수 있었다. (3) 또한 범주 단순화와 구체화를 응용해서 서로 다른 범주집합들 사이의 관계를 밝혀낼 수 있었다. 이 점은 상이한 온톨러지 범주집합 사이의 관계를 기준 온톨러지 범주정보를 활용해서 명시화할 수 있음을 시사한다. 이 방법을 이용한다면 서로 다른 온톨러지를 통합하는데 유용할 것이다.

본 논문에서 제안한 방안을 적용한 예는 다음과 같다 : (1) 일곱개의 범주로 구성된 '설화문장을 위한 온톨러지'를 로젯 시소러스를 활용해서 적은 비용으로도 구축 할 수 있었다. (2) '멀티미디어 게임 흡인력 분석'에서는 이미 알려진 22가지 게임 흡인요소를 바탕으로 범주 구체화를 통해 새로운 흡인요소 207가지를 발견했다. (3) '선심소와 게임 흡인요소 대응'에서는 선심소를 구성하는 10개의 심소와 게임 흡인요소를 구성하는 22가지 요소 사이의 관계를 밝힐 수 있었다.

정보시스템의 지능화는 피할 수 없는 추세이다. 이러한 지능형 시스템의 필수적인 구성요소가 온톨러지이다. 그러나 온톨러지를 새로이 개발하려면 많은 시간과 노력이 소요된다. 가용한 자원이 한정적이라는 측면에서 온톨러지 구축에 관해 다른 접근방식이 필요하다고 보았다. 이에 본 논문에서는 기존의 온톨러지를 재사용하기 위한 범주 재분류 방법론을 고안하였다. 이 방법론은 온톨러지에 대한 연산의 특별한 경우이다. 온톨러지에 대한 적절한 수학적인 모형[37, 38]이 개발된다면 병합과 집적, 공통부분, 차이 등과 같은 온톨러지 연산을 다양하게 정의할 수 있을 것이다. 이론적으로 뒷받침되는 이러한 연산을 적용하여 온톨러지 개발에 투입되는 시간과 노력을 대폭 줄일 수 있을 것이라 기대한다.

참 고 문 헌

[1] Sowa, J., Ontology, http://www.jfsowa.com/ontology/index.htm.

[2] "W3C Semantic Web", http://www.w3c.org/2001/sw/.

[3] Wooldridge, M. J. and Jennings, N. R., "Agent Theories, Architectures and Languages : A Survey," Springer-Verlag Lecture Notes in AI Volume 890, February, 1995.

[4] Uschold, M. and Gruninger, M., "Ontologies : Principles, Methods and Applications," Knowledge Engineering

Review, Vol.11, No.2, 1996.

[5] Pinto, H. S. and Martins, J. P., 'Reusing ontologies', AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes, pp.77-84, (2000), 2000.

[6] Michael Uschold, Michael Healy, Keith Williamson, Peter Clark, and Steve Woods, Ontology reuse and application. In Nichola Guarino, editor, Proc. of the Int. Conf. on Formal Ontology in Information Systems - FOIS 1998(Frontiers in AI and Applications v46), Amsterdam, 1998. IOS Press, pp.179-192, 1998.

[7] "ACQUILEX", http://www.cl.cam.ac.uk/Research/NL/acquilex/acqhome.html.

[8] Copestake, A., The Acquilex LKB : represention issues in semi-atomatic acquisition of large lexicons. In Proceedings of 3rd Conference on Applied Natural Language Processing. Trento. Italy. 1992.

[9] Alvar, M., editor. Diccionario General Ilustrado de la Lengua Española VOX. Biblograf S.A, Barcelona, Spain, 1987.

[10] Procter, P., editor. Longman Dictionary of Common English. Longman Group, Harlow, Essex, England, 1987.

[11] Ageno, A., Castellón, I., Ribas, F., Rigau, G., Rodríguez, H. and Samiotou, A., TGE : Tlink Generation Environment, In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), Kyoto, Japan, 1994.

[12] WordNet.http://www.cogsci.princeton.edu/~wn/.

[13] Rigau, G., Rodríguez, H. and Turmo, J., Automatically extracting Translation Links using a wide coverage semantic taxonomy. In Proceedings of 15th International Conference AI 1995, Montpellier, France, 1995.

[14] "Pangloss at Carnegie Mellon University CMT," http://www.lti.cs.cmu.edu/Research/Pangloss/.

[15] Knight, K. and Luk, S., Building a Large-Scale Knowledge Base for Machine Translation. In Proceedings of the American Association for Artificial Inteligence(AAAI 1994), 1994.

[16] Okumura, A. and Hovy, E., Building japanese-english dictionary based on ontology for machine translation. In Proceedings of ARPA Workshop on Human Language Technology, pp.236-241, 1994.

[17] "EDR Home Page," http://www2.crl.go.jp/kk/e416/EDR/index.html.

[18] Utiyama, M. and Hasida, K., Bottom-Up Alignment of Ontologies. In Proceedings of IJCAI workshop on Ontologies and Multilingual NLP, Nagoya, Japan, 1997.

[19] Atserias, J., Climent, S., Farreres, X., Rigau, G., Rodríguez, H., Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In Proceedings of International Conference on Recent Advances in Natural Language Processing(RANLP 1997), Tzigov Chark, Bulgaria, 1997.

[20] Farreres, X., Rigau, G., Rodríguez, H., Using WordNet for Building WordNets. In Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montr'eal, Canada, 1998.

[21] Daudé, J., Padró, L., Rigau, G., Mapping Multilingual Hierarchies Using Relaxation Labeling. In Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora(EMNLP/VLC 1999), Maryland, US, 1999.

[22] Daude, J., Padro, L., Rigau, G., Mapping WordNets using Structural Information, In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics ACL 2000. Hong Kong, China, 2000.

[23] 양재군, 배재학, 유해영, 이종혁, "Factotum SemNet을 활용한 개연규칙 검증", 한국정보처리학회 제19회 춘계학술발표대회 논문집, 제10권, 제1호, pp.349-352, 2003.

[24] 양재군, 강인수, 배재학, 이종혁, "Factotum SemNet의 MRD 변환에 관한 연구", 한국인지과학회 춘계학술대회 논문집, pp.279-284, 2003.

[25] Swartout, B., Patil, R., Knight, K. and Russ, T., Toward distributed use of largescale ontologies. In Proceedings of the Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop-KAW-1996, Banff, Canada, 1996.

[26] Arpirez, J. C., Gomez-Perez, A., Lozano-Tello, A., and Pinto, H. S., Reference ontology and(ONTO)2 agent : the ontology yellow pages. Knowledge and Information Systems, 2(4):387-412, 2000.

[27] Fernandez M., Gomez-Perez A, Juristo N., METHONTOLOGY : from ontological art toward ontological engineering. In Spring symposium series on ontological engineering, AAAI 1997, Stanford, CA, March, 1997.

[28] Blazquez M., Fern andez M., Garca-Pinar J. M., Gomez-Perez A., Building ontologies at the knowledge level using the ontology design environment. In Knowledge acquisition workshop, KAW 1998, Ban, Alberta, Canada, 1998.

[29] Benjamins, R. and Fensel, D., The ontological engineering initiative(KA)^2. In Proceedings of the First International Conference on Formal Ontology in Information Systems-FOIS 1998, pp.287-301, Trento,

Italy. IOS Press, 1998.
[30] "Stanford Knowledge Systems Laboratory", http://www-ksl.stanford.edu/.
[31] Roget's Thesaurus, http://promo.net/cgi-promo/pg/t9.cgi?entry=22&full=yes&ftpsite=ftp://ibiblio.org/pub/docs/books/gutenberg/.
[32] 양재군, 배재학, "온톨로지 정보를 이용한 범주 재편성 : 로젯 시소러스의 경우", 한국정보처리학회 제18회 춘계학술발표대회 논문집, 제9권, 제1호, pp.515-518, 2002.
[33] 양재군, "시소러스의 기계 가용화에 대한 연구", 울산대학교 석사학위논문, 2000.
[34] Bae J.-H. J. and Lee J.-H., Mid-Depth Text Understanding by Abductive Chains for Topic Sentence Selection, International Journal of Computer Processing of Oriental Languages, Vol.15, No.3, pp.341-357, 2002.
[35] 정혜영, 조윤경, 배재학, "온톨러지 정보를 이용한 멀티미디어 게임 흡인력 분석", 한국인지과학회 춘계학술대회 논문집, pp.15-20, 2002.
[36] 조윤경, 손인숙, 배재학, "멀티미디어 게임 흡인요소의 순화 : 유식학 응용", 한국정보처리학회 제18회 추계학술발표대회 논문집, pp.2451-2454, 2002.
[37] Boris Motik, Alexander Maedche, Raphael Volz, A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications, R. Meersman, Z. Tari, et al. (Eds.) : CoopIS/DOA/ODBASE 2002, Lecture Notes in Computer Science 2519, Springer-Verlag Heidelberg, pp.1082-1099, 2002.
[38] Sowa, John F. Conceptual Structures : Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984.

양 재 군

e-mail : jgyang@mail.ulsan.ac.kr
1997년 울산대학교 공과대학 건축학과 (공학사)
2001년 울산대학교 정보통신대학원 정보통신공학전공(석사)
2001년~현재 울산대학교 대학원 컴퓨터·정보통신공학부 박사과정 중
관심분야 : 자동프로그래밍, 인공지능, 온톨러지, 문서요약

배 재 학

e-mail : jhjbae@ulsan.ac.kr
1981년 중앙대학교 전자계산학과(이학사)
1983년 한국과학기술원 전산학과(공학석사)
2003년 포항공과대학교 컴퓨터공학과 (공학박사)
1985년~현재 울산대학교 컴퓨터·정보통신공학부 교수
관심분야 : 자동문서요약, (자동, 논리)프로그래밍, 지식경영 및 기술, 전략경영정보시스템, 교육인적자원정보시스템

이 종 혁

e-mail : jhlee@postech.ac.kr
1980년 서울대학교 수학교육과(이학사)
1982년 한국과학기술원 전산학과(공학석사)
1988년 한국과학기술원 컴퓨터공학과 (공학박사)
1991년~현재 포항공과대학교 컴퓨터공학과 교수
관심분야 : 자연어처리, 한국어정보처리, 한중일영 기계번역, 교차언어 정보검색, 문서요약, 문서분류