제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

온톨로지 범주의 재분류: Roget 시소러스의 경우

양재군, 배재학
울산대학교 컴퓨터 · 정보통신공학부
e-mail: {jgyang, jhjbae}@ulsan.ac.kr

Category Reorganization of Ontology: Roget Thesaurus Case

Jae-Gun Yang, Jae-Hak J. Bae
School of Computer Engineering and Information Technology
University of Ulsan

요 약

본 논문에서는 온톨로지의 범주를 재분류하는 방안을 모색하였다. 재분류 방법으로는 범주정보 단순화와 구체화를 고안하였다. 각각은 다시 표제정보와 참조정보를 이용한 방법으로 나누어 생각하였다. 또한 상이한 범주집합들 사이의 관계를 밝히는 방법도 모색하였다. 마지막으로 이러한 방법들을 활용한 예를 통해 그 유용성을 살펴보았다.

1. 서론

사람이 알고 있는 어휘의 의미와 그 어휘들 간의 결속관계의 집합이 온톨로지이다. 언어이해뿐만 아니라 우리가 접할 수 있는 여러 가지 문제나 현상의 파악에도 적절한 온톨로지가 필요함은 짐작할 수 있다. 최근 정보기술의 발달로 인하여 여러 분야에 걸쳐 온라인화 된 각종 문서의 양이 급격히 증가하고 있다. 따라서 이러한 여러 분야의 정보를 해석하는데 사용할 각각의 온톨로지가 필요하다. 하지만 하나의 온톨로지를 구축하는 경우에도 많은 시간과 인적자원이 요구되기 때문에 처리할 정보와 지식의 다양성에 비해서 온톨로지 자원이 부족한 실정이다.

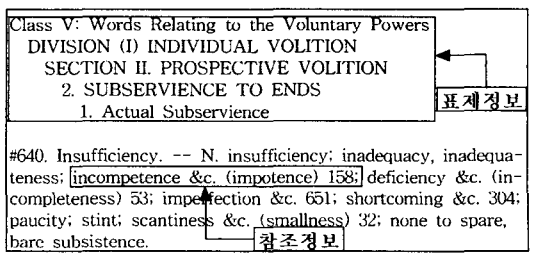
이에 본 논문에서는 기존의 온톨로지를 목적하는 바에 맞게 재활용하는 방안으로 Roget 시소러스[1] 범주정보를 재분류하는 방법론을 제시하고자 한다. 그 방법으로는 범주정보의 단순화와 구체화를 고안하였다. 이 과정에서는 Roget 시소러스의 표제정보와 참조정보를 활용하였다. 또한 상이한 범주들 사이의 관계를 밝히는 방법도 모색하였다. 마지막으로 이러한 방법들을 활용한 예를 통해 그 유용성을 살펴볼 것이다.

2. Roget 시소러스 범주 재분류

사람이 어떤 의미를 전달한다는 것은 언어를 이용해서 추상화된 개념을 전달한다는 것이다. 이런 점에서 보면 잘 짜여진 유의어 사전인 Roget 시소러스는 정보처리에 중요한 자원이라 할 수 있다. 본 논문에서는 주어진 정보를 범주화하고 그 범주들을 단순화 또는 구체화하는데 Roget 시소러스를 활용할 것이다.

2.1 Roget 시소러스

Roget 시소러스는 의미 분류를 기초로 총 6개의 강(Class)으로 구성되었다. 각 강은 하부에 부(Division), 과(Section) 등의 계층구조로 세분화되었다. 각 계층은 고유한 표제정보를 가지고 있으며 계층구조의 말단에는 총 1044개의 범주가 존재한다. 각 범주에는 품사별로 유의어 목록이 나열되어 있다. 한편, 유의어 목록에서 특정 어휘가 다른 범주를 참조하는 경우에는 "어휘 &c. (표제어) 표제번호"의 형식으로 표현한다. 이 표제정보와 참조정보를 탐색해서 대상 범주를 단순화 및 구체화하였다.



[그림 1] Roget 시소러스

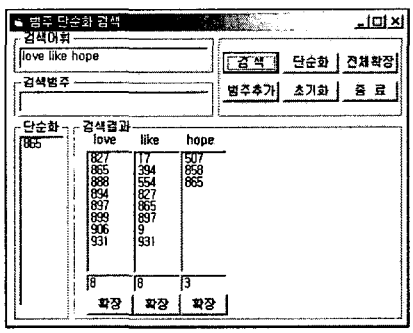
2.2 범주 단순화

범주 단순화란 여러 개의 범주들을 의미의 유사성이나 시소러스의 분류원칙을 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.

2.2.1 표제정보를 이용한 단순화

본래 Roget 시소러스의 구조가 의미체계에 기초한 계층적 분류체제를 따르고 있다. 따라서 처리 대상이 되는 정보 혹은 범주들을 시소러스 본문과 표제정보에서 검색한 후 발견되는 Roget 범주의 상위 계층을 단순화된 범주로 결정하였다.

이 단순화 방법을 자동화하기 위해서 어휘를 시소러스 본문과 표제정보에서 찾을 수 있는 검색 인터페이스를 제작하였다. [그림 2]처럼 복수개의 어휘를 동시에 검색하면 각 어휘에 해당하는 Roget 범주와 그 경우에 해당하는 단순화 범주를 계산한다. 또한 다음에서 살펴볼 참조정보를 이용한 범주 단순화에도 활용할 수 있도록 설계되었다.

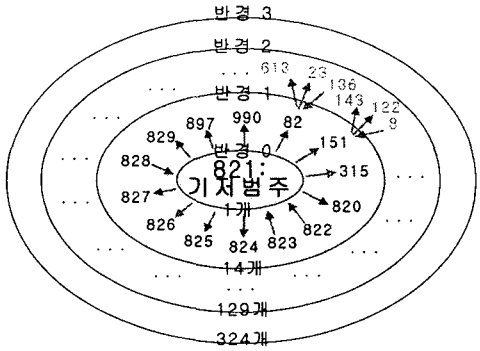


[그림 2] 범주 단순화 검색 인터페이스

2.2.2 참조정보를 이용한 단순화

다른 사전들처럼 Roget 시소러스도 어휘에 대한 부가적인 설명이나 참조가 필요한 경우, 어휘를 다른 표제어로 참조시킨다. 이러한 참조관계들의 연결인 참조 네트워크를 특정 기준으로 걸러 내거나, 참조들 사이에 관계를 맺은 후 재구성하면 범주 단순화 또는 구체화가 가능하다.

구체적인 방법은 참조 탐색의 출발점이 되는 기저범주를 결정한다. 이 기저범주의 반경 값을 0으로 하자. 그 후 이 기저범주에서 참조하는 범주들과 이 기저범주를 참조하는 범주들을 참조정보를 이용해서 추출한다. 이 범주들의 반경 값은 1로 한다. 같은 방식으로 [그림 3]처럼 모든 Roget 범주에 대해서 적당한 반경까지 확장한다.



[그림 3] 참조정보의 탐색 예

참조정보를 이용한 단순화의 경우에는 우선 복수의 정보 혹은 범주들에 대해서 각각의 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.

[표 1] 범주 단순화 활용예 : 선심소

심 소	참조정보 이용	인지적 판단
1. 신	484 Belief	484 Belief
2. 참	945 Vice	926 Duty
3. 괴	868 Fastidiousness	874 Disrepute
4. 무탐	866 Indifference	866 Indifference
5. 무진	897 Love	888 Friendship
6. 무치	518 Intelligibility	502 Sanity
7. 근	686 Exertion	686 Exertion
8. 경안	827 Pleasure	826 Inexcitability
9. 불방일	939 Probity	939 Probity
10. 행사	823 Insensibility	823 Insensibility
11. 불해	648 Goodness	618 Good

범주 단순화의 한 활용예인 유식학 선심소 분석([표 1])에서 보면, 이러한 단순화 방법의 결과와 사람의 인지적 판단에 의한 단순화 결과를 비교하면 총 11가지 범주에 대해서 서로 일치하는 경우가 4개의 범주였고 나머지 범주들에 대해서도 Roget 범주 내에서 근소한 차이를 보였다. 이 점은 Roget 시소러스의 구조를 감안하면 참조정보를 이용한 단순화 방법이 유의함을 시사한다.

2.3 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. 본 논문에서는 전처리된 Roget 시소러스[2]를

활용하여 두 가지 구체화 방법을 모색하였다. 첫 번째는 Roget 시소러스의 표제정보를 이용하는 방법이고 두 번째는 참조정보를 탐색하여 구체화하는 방법이다.

2.3.1 표제정보를 이용한 구체화

약식(informal) 온톨로지의 범주를 구체화시킬 때 기존의 온톨로지 범주 분류체계를 활용할 수 있다. 우선 구체화시킬 범주에 부합하는 기존 온톨로지의 범주들을 탐색한다. 이 결과를 약식 온톨로지의 해당 범주에 귀속시킨다. 이러한 과정을 반복하면 약식 온톨로지가 기존의 온톨로지를 통하여 구체화된다. 이점에 착안하여 Roget 시소러스의 표제정보를 이용해서 범주 구체화를 시도하였다.

모든 표제정보와 표제어를 취합한 후 각 표제정보와 표제어를 Prolog의 술어형태로 치환한다. 이 술어 형태의 표제정보에 구체화 기준을 적용해서 새로운 범주의 표제정보를 얻었다. 이 범주의 개수와 우선순위에 따라서 Roget 범주를 새로운 범주에 할당하였다.

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

2.2 참조정보를 이용한 구체화

Roget 시소러스의 범주 간 참조정보를 탐색하기 위해서 구체화 대상범주의 각 어휘를 선정한다. 대상 범주를 Roget 범주에 투영시키기 위해 선정한 어휘에 대응하는 Roget 범주를 기저범주로 정한다. 이 기저범주를 결정하기 위하여 Roget 시소러스의 표제정보를 이용하여 탐색한다. 이 경우에 표제정보에서 탐색할 수 없다면 시소러스 본문 중에 등장하는 범주들의 표제정보를 참고하여 택일한다. 결국 각 기저범주는 대상 범주를 가장 잘 나타내는 범주이다. 이 기저범주들과 나머지 Roget 범주 사이의 참조관계를 근거로 세부 범주들을 유도해낸 것이다.

모든 기저범주를 결정 한 후 참조정보를 탐색하여 세부 범주들을 찾아낸다. 그러기 위해 우선 기저범주의 반경 값을 0으로 정한다. 다음 단계에서 이 기저범주를 참조하는 Roget 범주들과 이 기저범주에서 참조하는 Roget 범주들을 취합한다. 이 범주들의 반경은 1이다. 같은 방법으로 참조정보를 탐색해서 범주의 반경을 한 단계씩 넓혀갈 수 있다.

이러한 구체화 방법은 Roget 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 교차참조가 빈번하다. 그러므로 구체화시키고자 하는 목적에 맞게 탐색반경을 한정시킨다.

3. 활용예

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

3.1 문장추상화를 위한 온톨로지 재분류

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

온톨로지를 재분류하기 위해서 Roget 시소러스 표제정보의 범주 값과 참조정보의 범주 값을 산출하였다. 산출된 각 범주 값의 가중치 평균을 근거로 범주를 결정하였다. 각각의 범주 값을 병합하는 과정에서는 표제정보와 참조정보의 중요도를 조절하기 위해 가중치를 적용하였다. [표 2]는 설화문장의 추상화를 위한 Roget 시소러스 단순화 재분류 결과이다[4].

[표 2] OfN : Ontology for Narratives

범 주	Roget 범주	개 수
등장인물	129, 130, . . . , 979, 980	15
심 상	11, 33, . . . , 999, 1000	351
사 건	40a, 60, . . . , 994	153
상 태	1, 2, . . . , 819, 965	460
공 간	180, 180a, . . . , 218, 219	35
시 간	106, 107, . . . , 136, 137	30
합 계		1044

3.2 멀티미디어 게임 흡인력 분석

인지 및 감성을 고려한 22가지 흡인요소를 Roget 시소러스의 범주정보를 바탕으로 22가지 범주로 분류하였다. 이 범주를 바탕으로 참조정보를 탐색하여 가까운 반경의 범주를 병합하였다. 이 22가지 범주를 세분화시켜서 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[5].

[표 3] OfG : Ontology for Game

	Roget 범주		개수
자극성	824	171, 392, 615, . . . , 835, 900	10
외 설	961	653, 945, 962	4
도 박	621	156, 470, 475 , . . . , 840, 945	9
폭 력	173	171, 274, 276, . . . , 900, 901	16
공 격	716	162, 276, 719, 934, 972	6
흥 분	825	173, 315, 821, . . . , 900, 901	9
비일상성	83	10, 16a, 18, . . ., 872, 964	19
현장감	1	120, 144, 151, . . . , 515, 66	12
환상감	515	1, 2, 353, . . . , 858, 984	22
도피성	671	260, 293, 295, . . . , 672, 750	10
중독성	613	104, 136, 16, . . . , 852, 871	15
공동체	712	178, 43, 696, . . . , 797, 892	10
도전감	715	742, 861, 909	4
만족감	639	102, 161, 168, . . . , 869, 953	13
긴장해소	738	460, 740, 742, 748, 925, 964	7
성취감	729	161, 292, 52, . . . , 680, 731	11
긴장감	686	171, 604, 604a, . . . , 682, 698	8
난이도	704	158, 248, 461, . . . , 706, 859	14
자유도	778	56, 709, 786	4
일체감	87	100, 43, 48, 50	5
지식획득	450	317, 451, 498, 515, 698	6
인 정	488	178, 23, 467, . . . , 762, 831	15
합 계			229

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. [표 4]는 게임의 흡인요소와 선심소의 대응관계와 각각의 선심소의 의미를 나타낸 표이다[6].

[표 4] 선심소와 게임 흡인요소 대응

선심소	심소의 의미	게임의 흡인요소
1. 신 (信)	진리를 믿고 인정하는 마음	환상감, 인정
2. 참 (慙)	성현의 가르침이나 도리를 어겼을때 부끄러워하는 마음	자극성, 도박, 폭력
3. 괴 (愧)	도덕이나 법을 어겼을 때 부끄러워 하는 마음	흥분, 폭력, 비일상성
4. 무탐 (無貪)	애탐하여 집착함이 없는 마음	중독성, 흥분, 비일상성
5. 무진 (無瞋)	괴로움에 대해서 성냄이 없는 마음	흥분, 인정, 자극성
6. 무치 (無痴)	어리석음이 없는 마음	환상감, 긴장감, 중독성, 인정, 지식획득
7. 근 (勤)	용맹하게 악을 끊고 선을 닦는 노력	긴장감, 성취감
8. 경안 (輕安)	번뇌를 멀리하여 편한 마음	인정, 자극성, 폭력, 중독성
9. 불방일 (不放逸)	번뇌를 단절하고 선을 닦음에 방탕하지 아니하는 마음	긴장감, 환상감, 중독성, 인정
10. 행사 (行捨)	치우치지 않는 평등한 마음	긴장해소, 비일상성, 중독성, 폭력, 난이도
11. 불해 (不害)	남에게 해를 끼치지 않고 자비심을 갖는 마음	성취감, 중독성, 인정, 환상감, 긴장감

4. 결론

본 논문에서는 기존의 온톨로지 정보를 활용하는 방안으로 범주 재분류에 관해서 살펴보았다. 온톨로지 정보로는 전처리된 Roget 시소러스를 이용하였다. 재분류 방법으로는 범주정보 단순화와 구체화를 고안하였다. 각각은 다시 표제정보와 참조정보를 이용한 방법으로 나누어 생각하였다. 또한 이 과정의 자동화에 사용할 범주 결정 시스템을 구현하였다.

범주 단순화 방법으로는 복수의 정보 또는 범주에 대한 대표성을 발견할 수 있었다. 또한 개별적인 정보 혹은 범주집합을 하나의 범주로 통합할 수 있었다. 범주 구체화 방법으로는 하나 또는 그 이상의 범주를 의미나 구조가 깨지지 않으면서도 원하는 목적에 맞게 나눌 수 있었다. 이를 통해 범주를 구성하는 새로운 요소들을 발견할 수 있었다. 또한 범주 단순화와 구체화를 이용해서 상이한 범주집합들 사이의 관계를 밝혀낼 수 있었다. 이 방법을 이용한다면 서로 다른 온톨로지를 통합하는데 유용할 것이다.

그 필요성에 비해서 부족한 온톨로지 자원을 새로이 구축하고자 한다면 상당한 노력이 필요할 것이다. 그렇다면 다른 방향의 접근방식을 취해야 할 것이다. 주어진 정보와 기존의 온톨로지 정보를 적절히 활용하는 방식이 그것일 것이다. 본 논문에서 시도한 범주 재분류 방법을 이용한다면 필요한 온톨로지를 보다 용이하게 얻을 수 있을 것이다.

참고문헌

[1] Roget's Thesaurus. http://promo.net/cgi-promo/pg/t9.cgi?entry=22&full=yes&ftpsite=ftp://ibiblio.org/pub/docs/books/gutenberg/.

[2] 양재군. "시소러스의 기계 가용화에 대한 연구." 울산대학교 석사학위논문, 2000.

[3] Bae J.-H. J. and Lee J.-H. "Topic Sentence Selection with Mid-Depth Understanding." Proc. of ICCPOL, pp. 199-204, 2001.

[4] 양재군, 배재학. "문장추상화를 위한 Roget 시소러스 범주 재편성." 한국인지과학회 춘계학술대회 논문집, pp. 40-45, 2002.

[5] 정혜영, 조윤경, 배재학. "온톨로지 정보를 이용한 멀티미디어 게임 흡인력 분석." 한국인지과학회 춘계학술대회 논문집, pp. 15-20, 2002.

[6] 조윤경, 손인숙, 배재학. "멀티미디어 게임 흡인요소의 순화: 유식학 응용", (출판예정), 2002.