천자문과 로제 시소러스 범주체계 비교

김상락○ 양재군 배재학

울산대학교 자동차선박기술대학원, (주)아이티스타○ 울산대학교 컴퓨터 정보통신공학부

shem0304@itstar.co.kr○ {jgyang, jhjbae}@ulsan.ac.kr

Category Comparison between Thousand-Character Text and Roget's Thesaurus

Sang-Rak Kim○ Jae-Gyun Yang Jae-Hak J. Bae

Institute of e-Vehicle Technology, University of Ulsan · ITSTAR Co., Ltd.○
School of Computer Engineering & Information Technology, University of Ulsan

요 약

본 논문에서는 천자문과 로제 시소러스의 어휘 분류체계를 분석하였다. 전처리 작업으로서 천자문과 로제 시소러스를 데이터베이스화 하였다. 그리고 분석 작업의 용이성과 연구의 효율성을 높이기 위해서 천자문 대응 로제 시소러스 검색 시스템을 개발하였다. 연구결과 로제 시소러스 총 39개 과(Section) 가운데에서 'Creative Thought' 과를 제외하고 대부분의 과가 천자문 한자와 관련성을 가지는 것으로 나타났다. 'Space in General', 'Dimensions', 'Matter in General' 3개의 과가 가장 높은 대응률을 보였다. 상관계수값도 약 0.94로 나타나 천자문 한자와 로제 시소러스의 범주 관련성이 아주 높다는 사실을 발견하였다.

Keywords : 천자문, 로제 시소러스, 온톨로지, 보편적 개념분류체계, Thousand-Character Text, Roget's Thesaurus, Ontology, Interontologia

1. 서 론

인터넷의 발달로 정보의 양이 예전과는 비교할 수 없을 정도로 많아졌다. 이렇게 많은 정보를 분류하고 관리하기 위해서는 표준화된 분류방법이 필요하다. 표준화된 분류방법은 인간의 인지적 사물 분류 능력을 통해서만 만들어 질 수가 있다. 이러한 인간의 인지능력은 한 개인의 고유한 능력이 아니다. 누구나 다 어릴 때부터 일상생활 속에서 사물을 어떻게 구분하고 정리하는 지를 배웠다. 그것을 통해 사물이나 개념을 보다 잘 이해하고 쉽게 처리할 수 있었다. 세상에 존재하는 것들은 모두 고유한 특징을 가지고 있기 때문에 특정한 범주로 구분될 수 있고, 우리는 이 모든 것들을 관련된 범주와 연관시킴으로써 이들을 보다 쉽게 이해할 수 있게 된다. 예를 들어 인간은 객체들의 공통된 특징에 따라 객체를 나무, 새, 물고기, 의자, 자동차 등의 범주로 분류한다. 이러한 범주 체계를 활용하면 객체의 특징까지도 추론할 수 있다. 의자의 범주에 속하는 객체를 보고 우리는 그것에 앉을 수 있다고 추론할 수 있게 된다. 이처럼 우리가 지각하는 세계를 각각의 특징에 따라 체계적으로 분류함으로써 정보 처리를 단순화시키고 인식하는 대상을 보다 잘 이해할 수가 있게 된다[1].

인간 사고과정에 대한 보편주의자들은 사람은 누구나 지각, 기억, 인과분석, 범주화, 그리고 추론과정에 있어서 동일하다고 말한다. 만일 차이가 난다면 그것은 문화의 차이 또는 교육의 차이 때문이지 서로 다른 인지 과정을 가지고 있기 때문은 아니라는 것이다[2]. 어휘 분류는 개념 분류의 성격을 가진다. [7~22]와 같이 현재 다양한 개념 분류체계가 존재한다. 응용 분야에 따라 다양한 개념 분류체계를 사용하고 있는 상황에서 개념 분류체계에 기반한 지능형 정보 시스템의 상호 연동에 대한 요구가 고조되고 있다. 이에 부응하여 온톨로지 사상(mapping), 병합 및 통합이나 의미적 통합에 관한 많은 연구가 진행되고 있다[3,4].

현재까지의 주된 연구방법은 공유 온톨로지의 활용이나 온톨로지 특질에서 사상을 발견하는 것이다. 그러나 준거 분류 체계로서 보편적 개념 분류체계(Interontologia)가 있다면, 개념 분류체계의 의미적 통합이 이것을 매개로 하여

보다 체계적이 되고 또 용이해질 것이다. 이에 본 논문에서는 보편적 개념 분류체계 탐구에 대한 사례 연구로서 동양의 대표적인 고전인 천자문[5,6]과 서양의 유명한 분류어휘집인 로제 시소러스(Roget's Thesaurus)[7] 간의 어휘 분류체계 관련성 연구를 하였다. 이 연구를 통해 두 비교대상의 분류체계 관련성을 분석하였다.

2. 관련 연구

천자문은 6세기경 중국 양 무제의 명을 받아서 주흥사(周興嗣)가 만들었는데, 우리나라에 일찍 전해져서 고대에 이미 보급되었던 책으로 아동용 교습서로서 널리 쓰였던 가장 대표적인 한문 교습서이다. 현재까지 한글로 훈음(訓音)이 붙여진 가장 오래된 천자문으로는 1583년에 발간된 한석봉체(韓石峰体)이다. 이보다 앞서 광주(光州)에서 1575년에 발간한 임란(壬乱) 이전판(以前板)이 있다는 기록도 있다. 천자문은 사자일구(四字一句)로 된 사언고시(四言古詩)로서 총 250句, 125对句(대구)로 이루어져 있는데, 소재는 주로 중국의 역사나 문화 등에서 취했다[5,6].

로제 시소러스[7]는 영국의 외과의사인 Peter Mark Roget에 의해 1852년에 첫판이 나왔다. 최초의 동의어 반의어 사전이다. 로제 유의어 사전은 단순한 의미 없는 알파벳 순서의 구조가 아니다. 로제 시소러스에는 어휘 지식이 체계적으로 분류되어 있다. 최상위 계층은 총 6개의 강(Class)으로 이루어져 있으며 하부에는 부(Division) 계층이 있다. 부는 다시 과(Section)로 구성된다. 각 계층별로 고유 표제정보가 있으며 계층구조의 말단에는 총 1044개의 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.

김종택, 송창선은 우리나라 대표적 고전인 천자문, 유합, 훈몽자회 등의 어휘 분류 체계를 연구하였다[5]. 이 논문에서는 현행 상용한자 1,800자 교육용 기초 한자에 대해서 무질서한 나열이라고 비판하고 있다. 반면 천자문은 구성이 잘 갖추어져있고 체계가 명백함을 주장하고 있다. 또한 천개의 글자 중 단 한 개의 글자도 중복을 허용하지 않는 분류체계의 정밀성에 대해서도 강조하고 있다. 이 논문에서는 천자문을 크게 의미단락(내용)별로 다음과 같이 분류를 하고 있다. "천문, 자연, 왕업, 수신, 충효, 덕행, 오륜, 인의, 궁전, 공신, 제후, 지세, 농사, 수학, 한거, 식사, 안락, 잡사, 기교, 경계" 등. 이 연구에서는 천자문이 의미단락별로 한자를 제시함에 있어서 기본적인 한자를 적절히 배치하였고 어휘체계 면에서나 기초한자에 대한 인식 면에서 우수하다고 결론짓고 있다.

본 논문에서 다루고 있는 어휘 분류 체계는 단어나 정보의 이용에 따라 다양하게 나눌 수 있다. 인공지능, 계산언어학, 정보통신 응용분야의 예로서, 정보 검색, 지식관리 및 경영, 정보 시스템 설계, 온톨로지 구축, 기계 번역, 사전 편찬 등을 들 수 있다. 관련 구축 사례로는 자연어 어휘자원 관련으로서 Roget's Thesaurus[7], WordNet[8], Lexical FreeNet[9], Kadokawa Thesaurus[10], EDR[11] 등이 있다. 온톨로지 구축분야에는 KR Ontology[12], CYC Ontology[13], Mikrokosmos Ontology[14], SENSUS Ontology[15], HowNet[16] 등이 있다. 응용분야로는 Enterprise Ontology[17], UMLS[18], UNSPSC[19], RosettaNet[20], ISO 2788[21], ANSI Z39.19[22] 등이 있다.

3. 연구방법 및 시스템 개발

3.1 관련성 분석 연구방법

본 논문에서는 천자문 한자와 로제 시소러스의 분류체계 관련성 연구를 위해 다음과 같이 7단계로 작업을 수행 하였다.

Step1: 천자문, 로제 시소러스 정리 작업 및 데이터베이스 구축. 천자문, 로제 시소러스 파일을 정리하여 로제 시소러스, 천자문 한자 마스터 데이터베이스 구축.

Step2: 천자문 한자와 유사한 의미를 지니는 영어단어 정리 작업 및 데이터베이스 구축. 구글 중국과 Kingsoft가 공동 출시한 중영사전 Kingsoft2008[23]와 Classical Chinese Character Frequency List[24] 를 참고하여 천자문 한자와 의미가 유사한 영어단어 데이터베이스 구축.

Step3: 천자문 한자의 의미와 가장 유사를 의미를 지니는 영어단어 맵핑 작업. 로제 시소러스 대응 천자문 한자 화면을 이용하여 천자문 한자의 의미와 가장 유사를 의미를 지니는 영어단어를 로제 시소러스 콤보박스에서 선택하여 맵핑

Step4: 로제 시소러스의 범주 정리 및 데이터베이스 구축. 로제 시소러스 범주 분류 파일을 정리하여 데이터베이스 구축.

Step5: 천자문 한자 로제 시소러스의 범주와 맵핑 작업. 로제 시소러스 대응 천자문 한자 화면을 이용하여 천자문 한자의 영어단어 상위 로제 시소러스 범주를 콤보박스에서 선택하여 맵핑.

Step6: 로제 시소러스 대응 천자문 한자 결과 분석. 로제 시소러스 과의 범주에 해당하는 천자문 한자들의 대응성을 화면에서 조회.

Step7: 로제 시소러스 과별 천자문 한자 결과 분석. 로제 시소러스 과의 범주 수와 천자문 한자 대응수를 분석하여 <그림 3>, <그림 4>와 같이 표와 그래프로 표현하고 상관계수 값을 구한 후 상관관계 분석 그래프<그림 5>로 표현함.

3.2 시스템 설계

천자문 한자와 로제 시소러스의 과(Section)가 대응하

는 관련성 분석 위주로 시스템을 설계하였다. 천자문 한자를 저장할 수 있는 테이블과 천자문 한자와 가장 유사한 의미를 지닌 영어단어를 저장하는 테이블 설계를 하였다. <그림 1>은 천자문과 로제 시소러스의 관련성 분석을 위한 개체관계도(ER Diagram)이다. 그리고 최종적으로 로제 시소러스 과와 관련성을 표현하기 위해서 로제 시소러스와 로제 시소러스 범주 설명 테이블을 추가적으로 설계하였다. 천자문 각각의 한자는 여러 가지 유사한 의미를 지니는 영어단어들이 많다. 해당 단어들을 모두 콤보박스에 채워서 천자문 한자에 가장 근접한 의미를 지니는 단어를 선택하여 맵핑할 수 있도록 설계하였다. 천자문 한자에 대응하는 영어단어의 로제 시소러스 범주도 다수 이므로 선택, 지정 작업의 용이성을 고려하여 콤보박스 처리를 하였다.

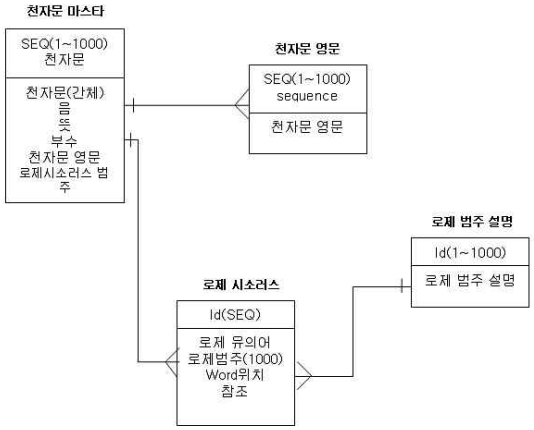

그림 108 천자문과 로제 시소러스의 관련성 분석을 위한 개체관계도

3.3 시스템 구성

본 시스템은 <그림 2>와 같이 4개의 화면으로 구성되어 있다. 초기메뉴 화면과 천자문 대응 로제 시소러스, 로제 마스타, 천자문 대응 로제 시소러스 결과분석 등이다.

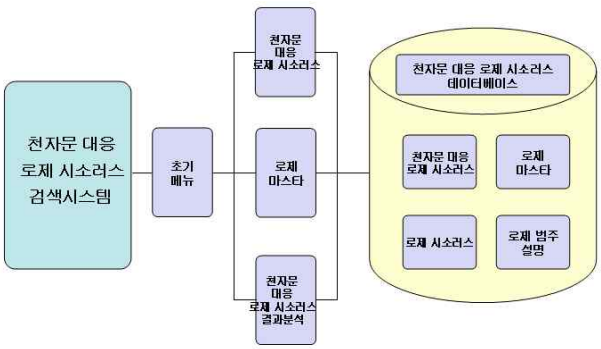

그림 109 천자문 대응 로제 시소러스 검색 시스템

4. 분석 결과

<그림 3>은 로제 시소러스 과 단위 분석 결과 데이터를 보여주고 있다. 총 39개 과 가운데에서 'Creative Thought' 과를 제외하고 대부분의 과가 천자문 한자와 대응성을 가지는 것으로 나타났다. 'Space in General', 'Dimensions', 'Matter in General' 등 3개의 과가 가장 높은 대응률을 보이고 있다. 주로 비물질적이고 형이상학적 의미를 내포하는 천자문 한자가 대응률이 높게 나타났다.

번호	강(Class)	과(Section)	로제 범주 개수	천자문 한자 개수	대응률	순위
01	1	existence	8	4	0.50	6
02	1	relation	18	9	0.50	6
03	1	quantity	34	15	0.44	15
04	1	order	27	11	0.41	19
05	1	number	23	9	0.39	23
06	1	time	36	16	0.44	13
07	1	change	13	6	0.46	11
08	1	causation	28	11	0.39	21
09	2	space in general	13	8	0.62	2
10	2	dimensions	49	33	0.67	1
11	2	form	25	11	0.44	16
12	2	motion	52	23	0.44	14
13	3	matter in general	5	3	0.60	3
14	3	inorganic matter	37	13	0.35	25
15	3	organic matter	100	44	0.44	16
16	4	operations of intellect in general	6	3	0.50	6
17	4	precursory conditions and operations	15	6	0.40	20
18	4	materials for reasoning	9	2	0.22	34
19	4	reasoning processes	4	1	0.25	30
20	4	results of reasoning	26	8	0.31	27
21	4	extension of thought	9	2	0.22	34
22	4	creative thought	3	0	0.00	39
23	4	nature of ideas communicated	9	1	0.11	38
24	4	modes of communication	26	11	0.42	18
25	4	means of communicating ideas	50	17	0.34	26
26	5	volition in general	23	5	0.22	36
27	5	prospective volition	60	28	0.47	10
28	5	voluntary action	24	11	0.46	12
29	5	antagonism	25	13	0.52	5
30	5	results of voluntary action	8	3	0.38	24
31	5	general intersocial volition	25	14	0.56	4
32	5	special intersocial volition	8	2	0.25	30
33	5	conditional intersocial volition	8	2	0.25	30
34	5	possessive relations	51	20	0.39	22
35	6	affections in general	7	2	0.29	28
36	6	personal affections	62	29	0.47	9
37	6	sympathetic affections	36	10	0.28	29
38	6	moral affections	56	14	0.25	30
39	6	religious affections	26	4	0.15	37

그림 110 천자문과 로제 시소러스의 대응률 (Section 수준)

천자문 한자와 로제 시소러스 과의 대응 그래프를 <그림 4>와 같이 나타내었다. 천자문과 로제 시소러스가 과 수준에서 비슷한 경향으로 대응하는 결과를 볼 수 있다.

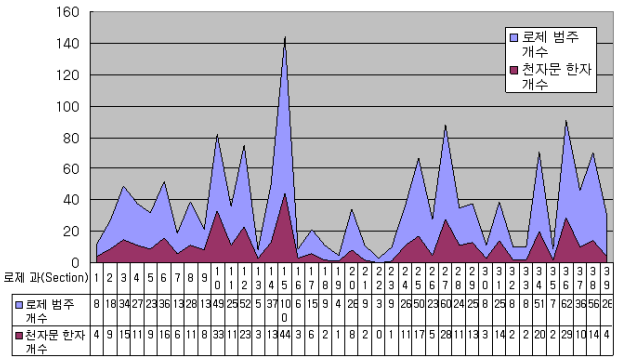

그림 111 천자문과 로제 시소러스의 대응(Section 수준)

<그림 3>의 로제 시소러스 과의 범주 개수와 천자문 한자의 대응 개수의 값을 상관분석(Correlation Analysis) 하였다. 상관계수는 로제 시소러스 과에 속한 범주 개수

의 변화에 따라 천자문 한자의 대응 개수가 변화하는 정도를 나타낸다. 상관계수를 얻기 위한 r_{xy}는 다음 (식 1)과 같다[25].

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

(식 1)

여기에서 로제 시소러스 과의 범주 개수를 X, 과에 대응하는 천자문 한자 개수를 Y, 로제 시소러스 과의 개수를 N으로 정한다. (식 1)에서 $\sum X_i$, $\sum Y_i$, $\sum X_i Y_i$, $\sum X_i^2$, $\sum Y_i^2$ 들에 대한 값은 다음과 같이 정해진다.

$\sum X_i$ = 1044, $\sum Y_i$ = 424, $\sum X_i Y_i$ = 18629, $\sum X_i^2$ = 44474, $\sum Y_i^2$ = 8252

그리고 이 값들을 (식 1)에 대입하여 계산하면 (식 2)의

$$r_{xy} = \frac{39(18629) - (1044)(424)}{\sqrt{\left[39(44474) - (1044)^2\right]\left[39(8252) - (424)^2\right]}}$$

$= 0.938156834$

(식 2)

r_{xy}는 약 0.94이다. 천자문과 로제 시소러스가 과(Section) 수준에서 상관도가 매우 높다는 것을 알 수 있다.

두 변인 간의 관계를 시각적으로 나타내기 위하여 <그림 5>과 같은 천자문과 로제 시소러스의 상관분석 그래프를 얻었다. 각 점은 천자문에 대한 로제 시소러스의 대응 과(Section)를 표현하는데, x좌표는 과 내부의 범주 개수를 그리고 y좌표는 과에 대응하는 천자문 한자 개수를 나타낸다. 이 그래프는 선형의 모습을 띠면서 매우 강한 관련성을 보이고 있다.

5. 결론 및 향후연구

지금까지 천자문 한자와 로제 시소러스의 관련성을 상관분석과 분포도 그래프를 통해 살펴보았다. 상관계수 결과 값과 그래프의 선형성으로 실험 대상의 관련성이 아주 높다는 사실을 알 수 있다. 이 실험 결과로 로제 시소러스 개념 분류체계가 천자문 한자의 분류체계와 무관하지 않다는 사실을 발견하였다. 현재 우리가 사용하고 있는 천자문 한자를 좀 더 보완한다면 로제 시소러스에 못지않은 지식 분류체계 자료로 활용이 가능함을 짐작할 수 있다.

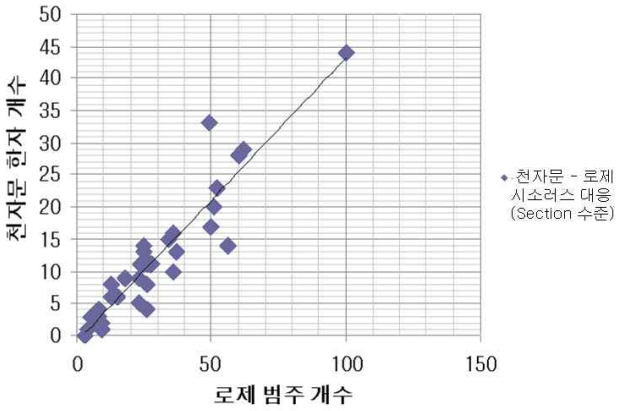

그림 112 천자문과 로제 시소러스의 상관분석

향후 연구 과제는 (1) 로제 시소러스 범주와 천자문 한자 대응 누락 부분을 교육용 한자 1800자에서 보완하여 재비교 (2) 천자문과 로제 시소러스를 범주 수준에서 비교하는 것과 과 수준에서 비교하는 것과의 차이분석 등이다. 이러한 연구들을 통해 정제된 어휘지식분류 체계용 천자문 한자를 얻게 될 것이다. 최종적으로 이러한 천자문 한자를 이용하여 지식 분류체계에 활용할 수 있는 새로운 형태의 시소러스 범주 모델을 개발할 예정이다.

<Acknowledgement>
본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다. (IITA-2008-(C1090-0801-0039))

참고문헌

[1] 노상규, 박진규, “인터넷 진화의 열쇠 온톨로지”, 가즈토이, 2007.
[2] Richard E. Nisbett (최인철 역), “생각의 지도”, 김영사, 2004.
[3] Yannis Kalfoglou and Marco Schorlemmer, Ontology mapping: the state of the art, The Knowledge Engineering Review, Vol. 18:1, 1 - 31, 2003.
[4] Natalya F. Noy, Semantic Integration: A Survey Of Ontology-Based Approaches, SIGMOD Record, Vol. 33, No. 4, December 2004.
[5] 김종택, 송창선, “천자문, 유합, 훈몽자회의 어휘 분류 체계 대비”, 한국어문학회, 어문학, 제52호, 159-192, 1991.
[6] 진태하, “『千字文』의 訓音 中 問題点”, 한글+漢字문화, 제104호, 80-82, 2008.3.
[7] Roget's Thesauri, http://www.bartleby.com/thesauri/
[8] WordNet, http://wordnet.princeton.edu/.

[9] Lexical FreeNet, http://www.cinfn.com/doc/.

[10] Ohno, S. and M. Hamanishi. 1981. New Synonyms Dictionary, Kadogawa Shoten, Tokyo. (Written in Japanese).

[11] The EDR Electronic Dictionary, http://www2.nict.go.jp/r/r312/EDR/index.html.

[12] KR Ontology, http://www.jfsowa.com/ontology/.

[13] CYC Ontology, http://www.cyc.com/.

[14] Mikrokosmos Ontology, http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html

[15] SENSUS Ontology, http://www.isi.edu/natural-language/projects/ONTOLOGIES.html

[16] How Net, http://www.keenage.com/html/e_index.html Enterprise

[17] Ontology, http://www.aiai.ed.ac.uk/project/enterprise/enterprise/ontology.html

[18] UMLS, http://www.nlm.nih.gov/research/umls/.

[19] UNSPSC, http://www.unspsc.org/.

[20] RosettaNet, http://www.rosettanet.org

[21] ISO 2788, http://www.collectionscanada.gc.ca/iso/tc46sc9/standard/2788e.htm

[22] ANSI Z39.19, www.niso.org/standards/resources/Z39-19-2005.pdf

[23] Kingsoft2008(谷歌金山詞霸), http://g.iciba.com/.

[24] Classical Chinese Character FrequencyList, http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=CL.

[25] 김영채, "사회과학의 현대 통계학", 박영사, 2005.