과학기술 문헌에 나타난 시소러스의 연관관계 유형에 관한 연구

송유화○, 최호섭
과학기술연합대학원 응용정보과학과, 한국과학기술정보연구원
painting85@kisti.re.kr, hschoe@kisti.re.kr

The type of associative relationships of Thesaurus described in literature of science and technology

Yoo-hwa Song○
University of Science and Technology

Ho-seop Choe
Korean Institute of Science and Technology Information

요 약

시소러스의 연관관계는 유형의 세분화에 관한 원칙과 방법론의 부재로 시소러스를 구축하는 기관에서 개별적인 분류를 사용하고 있다. 분류에 적용되는 패싯지시어 모형에 관한 연구는 계속 되고 있지만 그 타당성을 뒷받침 할 실증적 사례연구는 찾아볼 수 없다. 본 연구에서는 Inspec에서 구축한 시소러스 중에 일정 기준으로 선정한 우선어와 관련어를 대상으로 IEL에서 제공하는 문헌에서 두 용어가 동시에 출현하는 문장을 찾아 그 연관관계 모형을 제안한다.

주제어: 시소러스, 연관관계 유형, 관련어

1. 서론

시소러스는 전통적인 지식조직체계로서 시소러스 구축에 관한 표준인 ISO 2788에서는 시소러스에 대하여 “기능적인 면에서는 문헌, 색인자, 이용자의 자연언어를 통제가 가해진 시스템 언어로 변환시킬 때 사용되는 용어 통제표이며, 구조적인 면에서는 지식의 어떤 특수한 영역을 포함한 일반적이며 상위 개념의 용어와 하위 개념의 용어를 의미론적으로 밝힌 동적인 어휘집”이라고 정의하고 있다. 그 구성은 상하관계를 중심으로 하여 동등관계, 계층관계, 연관관계로 이루어진다. 시소러스 구축의 목적이 최종 이용자에게 참조도구로 사용되어 정보 검색의 효율을 높임에 있으나 시소러스의 구조는 개념 간에 매우 한정적인 관계를 갖기 때문에 개념 간의 관계를 충분히 반영하지 못한다는 지적을 받고 있다[1].

연관관계의 유형 분류에 대해서는 지침뿐만 아니라 참고할 만한 선행연구 또한 거의 없다. 단지 일련의 연관관계 유형 목록을 나열한 연구와 연관관계 분류를 위한 패싯을 제안한 연구가 있을 뿐이다. 즉 기존 연구들이 과연 실제 연관관계에 대하여 어느 정도의 효용성을 갖고 있는지에 대한 이론적, 실증적 기반은 없는 실정이다[2]. 따라서 본 연구는 기구축된 시소러스의 우선어와 관련어의 연관관계가 문헌에서 어떻게 나타나는지 찾아보고 관계패턴을 그룹핑하여 연관관계 모형을 제안한다.

2. 관련연구

시소러스 연관관계의 세분화에 관한 연구[3][4][5][6]는 시소러스의 구축이 연구된 시점부터 Ranganathan(1976)이 콜론분류에서 제안한 ‘PMEST’ 패싯지시어를 기반으로 세분화 모형을 제시하고 있다. 국내에서도 콜론 분류법의 5개 범주와 이와 유사한 NLM(National Library of Medicine)의 UMLS(Unified Medical Language System)의 메타시소러스를 결합시켜 새로운 연관관계를 제안하거나[7] 국내외에서 제안한 패싯지시어를 기반으로 하여 사례 분석 결과를 종합하여 통합적으로 제시한 연구가[8] 있다. 그러나 이러한 세분화 연구는 실제 시소러스를 대상으로 평가가 이루어지지 않았으며 이는 존재론적인 분류가 실제 출현 양상에 기반을 두지 않았음을 뜻한다. 백지원(2005)의 연구에서도 용어적 타당성을 주장하며

의미론적으로 관계의 설정이 가능하고 언어학적으로 분류가 가능하다고 해서 관계를 설정하거나 의미론적으로 계속적인 분류를 해 나가야하는 것은 아니라고 지적했다[8].

기존의 패싯분석을 토대로 연관관계를 설정하고 시소러스나 어휘망의 관련어를 수동 및 반자동으로 구축한 연구도 이루어지고 있다[1][9][10][11]. 과학기술 시소러스 구축 연구에서는 연관관계를 구성요소, 기능, 도구, 부속, 용도, 유형, 처리로 나누었다[12]. 생물학, 의학 등의 문헌을 바탕으로 온톨로지를 구축하는 연구에서 주제 분야에 알맞은 다양한 연관관계를 형성하기도 했다[13]. 텍스트 마이닝 분야에서도 자동 관계추출을 통해 연관관계를 설정하는 연구가 되고 있다[14].

본 논문에서는 Inspec Thesaurus에서 선정된 우선어와 그에 대응하는 673개의 관련어와의 연관관계를 문헌에서 찾아 패턴을 분석하고, 분석결과에 따른 관계모형을 제안하고자한다.

3. Inspec Thesaurus의 연관관계 분석

3.1 Inspec Thesaurus 소개 및 분석 개요

Inspec은 IEE(The Institution of Electrical Engineers)와 IET(The Institution of Engineering and Technology)가 발행하는 과학기술 문헌 데이터베이스로 Inspec이 제공하는 주제 분야는 다음과 같다.

- biomedical engineering
- biophysics
- computing
- control engineering
- electriclal and electronic engineering
- information technology
- materials science
- oceanography
- nanobiotechnology
- nuclear engineering
- physics
- power & energy
- radar

이 데이터베이스는 IEL(IEEE/IET Electronic Library)에서 서비스되고 있으며, 1969년부터 발행된 5000여 개의 저널과 2500여 종의 회의록을 제공하고 있다. Inspec은 보다 정확하고 자세한 주제 정보를 제공하기 위하여 Inspec Thesaurus를 제공한다. Inspec Thesaurus는 9,400개의 우선어를 포함한 18,000개의 어휘로 이루어져 있으며 전문가에 의해 수작업 선정되고 관리되고 있다.

Inspec Thesaurus의 우선어는 평균적으로 4개의 관련어를 갖는다. 많은 수의 관련어를 갖는 우선어가 다양한 연관관계를 포괄한다는 전제 하에, <표1>과 같이 최대 50개에서부터 27개 이상의 관련어를 갖는 우선어를 선정하였다. 이 우선어들과 이에 대응하는 673개의 관련어를 각각 쌍을 지어 IEL에서 검색되는 문헌을 대상으로 관계를 분석하였다.

관련어의 개수	우선어
50	signal processing
44	semiconductor technology
38	lasers
38	microcomputers
35	computers
35	spectra
32	safety
31	image processing
31	quality control
31	random processes
31	semiconductor devices
29	engines
29	quantum field theory
28	filtering theory
28	control system synthesis
28	quantum optics
27	integrated circuit technology
27	control system analysis
27	plasma confinement
27	reliability
27	stars

<표 1 연관관계 분석에 선정된 우선어>

검색 대상 콘텐츠는 저널, 학술대회 논문, 얼리억세스를 검색 대상으로 하였으며 용어 쌍을 검색하여 상쉬 25개의 검색결과 문헌을 분석 대상으로 지정하였다.

3.2 분석 과정

패턴을 분석하는 과정에서 우선어와 관련어의 출현 형태는 다양하게 나타났다. 우선어와 관련어가 한 문장에 동시에 출현하지 않고 한 문단에 출현하거나 서로 다른 문단에 출현하기도 한다. 심지어 동시에 출현하는 문헌이 없는 경우도 있다. 본 연구에서는 분석의 대상을 우선적으로 우선어와 관련어가 동시에 출현하는 한 문장과 지시대명사로 관계를 판단할 수 있는 이어진 복수의 문장으로 삼았다. 이에 해당하는 용어의 쌍은 673개의 용어 쌍 중 288개의 용어 쌍이었다. 그 중 우선어와 관련어가 동시에 한 문장에서 출현하였으나 관계를 판단하기가 어려운 용어 쌍이[1] 77개가 있기 때문에 이를 제외한

1) 한 문장에 나타나지만 관계를 추출하기 어려운 이유는 우선어와 관련어가 문장의 다른 요소와 관계를 갖고 있기 때문이다. 이 부분에 대한 연구는 인접해있는 다른

211개의 용어 쌍에서 관계 패턴을 추출하였다.

연관관계를 판단함에 있어 용어적 타당성을 뒷받침하는 근거를 생성하기 위하여 문헌에서 나타나는 형태를 되도록 수용하는 규칙을 따랐다. 이는 형이상학적으로 분류를 하여 일반 동사로 관계를 판단하는 전통적인 연관관계 선정뿐만 아니라 용어가 문헌에서 쓰이는 여러 형태를 반영하기 위함이다. 출현 패턴의 다양한 표현형식을 체계적으로 분석하기 위해 우선어와 관련어가 동시에 출현하는 문장의 표현형식을 <표2>와 같이 크게 세 가지로 나누었다.

형식 구분	정의 및 형식
정의문 형식	'is a', 'call' 등의 정의적 의미 동사를 갖는 문장, 전치사구에 우선어 혹은 관련어가 출현
	USE/RT + is a {something} for/in + USE/RT
비 정의문 형식	일반 동사를 갖는 문장
	USE/RT + {일반동사} + USE/RT
형태적 구조	우선어와 관련어가 함께 복합명사를 이루는 경우
	USE · RT / RT · USE

<표 2 문장의 표현형식>

문헌정보학에서 전통적으로 연관관계를 연구하는 방식은 패싯지시어를 사용하여 유형을 분류하고 분류에 맞는 일반동사를 예로 적는 존재론적 분류이다. 따라서 현재까지 연관관계의 유형의 연구대상이 되는 문장형식은 일반동사가 쓰인 비정의적 형식이다.

그러나 문헌에서는 우선어와 관련어가 동시에 출현하는 한 문장 중 정의문이 다수 등장하였다. 정의적 문장은 비록 상위어와 하위어를 찾아 계층관계를 추출하기 위해 활용되지만 판단 대상을 주어와 목적어 혹은 보어뿐만 아니라 다른 문장요소도 포함할 수 있다고 전제한다면 정의문에서 관련어도 추출할 수 있다. 예를 들어, 우선어, 'control system analysis'와 관련어, 'parameter space method'가 동시에 출현하는 문장을 검색하여 얻은 문장인 'The parameter space method **is a tool in** control system analysis and synthesis.' 를 보면 우선어가 전치사구에 나타난다. 이 문장으로부터 전치사구를 포함하여 관계패턴을 추출하면 'is a (tool) in' 라는 패턴을 얻을 수 있다.

따라서 본 연구에서는 문헌에 나타난 관계 패턴을 그대로 수용하기 위해 분석 대상의 범위를 '주어- 목적어-동사'뿐만 아니라 전치사구, 관계대명사 혹은 지시대명사로 이어지는 문장, 복합명사로 나타나는 경우도 관계패턴 분석 대상에 포함하였다.

정의문 형식은 우선어 혹은 관련어가 'is a'라는 정의적 동사를 갖는 문장의 전치사구 자리에 등장한 경우로 그 형식은 **'USE/RT is a {something} in/of USE/RT'** 이다. 그 예는 <표3>과 같다.

우선어: control system
관련어:parameter space method
The parameter space method **is a tool for in** control system analysis and synthesis.

우선어: Walsh function,
관련어: signal processing
Walsh function **is a tool for** signal processing.

<표 3 정의문 형식의 예>

비정의문 형식은 주어와 목적어가 우선어 혹은 관련어이고 일반 동사를 갖는 문장이며 그 형식은 **'USE/RT + {일반동사} + USE/RT'**로 표현할 수 있다. 그 예는 <표4>와 같다.

우선어: reliability
관련어: preventive maintenance
Simple preventive maintenance **improves** the reliability of a facility to a higher level.

우선어: signal processing
관련어: special time-varying filters
Robuston signal processing **employs** special time-varying filters.

<표 4 비정의문 형식의 예>

우선어나 관련어가 관계대명사 혹은 지시대명사로 이어진 문장은 **'USE/RT + 동사 + 관계대명사/지시대명사 + {동사}+ USE/RT'** 로 표현할 수 있으며 그 예는<표5>와 같다.

우선어: signal processing

용어들과의 상관성을 고려하여 다차원적인 접근방식으로 분석해야 하지만 향후 연구 내용으로 다루기로 한다.

관련어: digital filters
Digital filters with variable fractional group-delay are referred to as variable fractional-delay digital filters, **which are useful in** various signal processing applications.

우선어: stars
관련어: astronomy
Millimetre wave astronomy has thus opened a way to observe the interstellar matter directly in contrast with the conventional optical astronomy **which has been observing** the stars.

우선어: signal processing
관련어: expectation maximisation (EM) algorithm
The expectation maximisation (EM) algorithm is an iterative search technique for solving maximum likelihood estimation problems. **It** is an alternative to the more common gradient-based search approaches and it has proven useful in applications where gradients are difficult to compute. **It** is also well regarded for its numerical stability. **The method** has its origins in the statistics literature, but has been widely **applied in** very many other areas such as image processing, econometrics , epidemiology.

<표 5 문맥적 구조의 예>

위의 세 가지 형식을 분석할 때, 우선어 혹은 관련어가 복합명사의 첫 번째 명사로 등장하여도 경우에 따라 분석대상으로 삼았다. 복합명사를 이루는 첫 번째 단어(N_1)과 두 번째 단어(N_2) 중에 실질적 의미를 갖는 단어는 N_2이다. 예를 들어 safety protection의 주된 의미는 protection이고 echo spectra의 주된 의미는 spectra이다. 그렇기 때문에 일반적으로 문장을 분석할 때 N_1의 의미는 고려하지 않는데, 본 연구에서는 우선어나 관련어가 N_1으로 등장하고, N_2가 application, method, technique인 경우에는 분석 대상으로 삼았다. 그 이유는 technique과 application, method를 N_2로 갖는 문장에서<표6>과 같은 일정한 규칙이 나타나 하나의 관계유형으로 간주할 수 있기 때문이다.

RT [is a] USE **technique**

USE **technique** [include] RT

USE **technique** [perform] RT

RT [is a issue in] USE **application**

RT [consider as sth in] USE **application**
RT [is useful in] USE **application**
RT [is important in] USE **application**
RT [is ued in] USE **application**
RT [consider as sth in] USE **application**
RT [is interest in] USE **application**
RT [emerge as a sth for] USE **application**

RT **method** [is a tool in] USE

RT **method** [is used in] USE
USE **method** [call] RT
USE **method** [base on] RT

<표 6 특정 단어와 복합명사를 이루는 경우의 예>

형태적으로 우선어와 관련어가 함께 복합명사를 이루는 경우는 우선어와 관련어가 N_1, N_2로 등장하여 복합명사를 이루는 경우이다. 이러한 형식을 분석 대상으로 채택한 이유는 첫째, 복합명사로 등장하는 경우, 특정 문헌에서만 일시적으로 나타나는 것이 아니라 여러 차례에 걸쳐 복수의 문헌에서 복합명사 형태로 나타났기 때문이고 둘째, 패턴을 추출한 전체 문장을 의미적으로 그룹핑을 했을 때, 한 그룹을 이루는 문장의 수가 평균 5개라면 복합명사 그룹을 이루는 문장의 수는 17개로 간과할 수 없다고 생각했다. 의미적인 관계를 정의할 순 없지만 우선어와 관련어가 복합명사를 이룰 때, N_1이나 N_2가 관련어가 될 수 있다는 데 의의를 두었다. 그 예는 <표7>과 같다.

우선어: lasers
관련어: optical parametric oscillator
A new PSD measuring system was firstly built up based on the optical parametric oscillator laser.

우선어: quality control
관련어: six sigma
Six sigma quality control usually uses the achievement improvement model of DMAIC.

<표 7 형태적 구조의 예>

4. 분석 결과

211개의 용어 쌍의 관계를 그룹핑하여 43개의 관계유형을 도출하였다. 관계유형을 판단할 수 있는 문장에 출현한 일반 동사의 종류는 53개였으며 이를 <표 8>과 같이 의미적으로 33개로 그룹핑 하였다. 정의적 문장과 비정의적 문장 중, 우선어나 관련어가 복합명사 중 N_1으로 나타나더라도 일정한 규칙이 나타나는 경우를 선택적으로 지정하여 관계유형으로 설정하였다. 관계 유형 중 분야/목적은 정의적 문장에서 우선어가 문장의 전치사구에 나타나 in을 취하는 경우, 특정분야를 나타내거나 for를 취하는 경우 목적을 나타냈기 때문에 분야/목적이라고 명명하였다. D-t/a/m은 정의적 문장에서 특정단어(technique, application, method)와 복합명사를 이루는 패

형식 구분	동사구분	관계 유형	대표적인 용어 관계
정의적 문장	be 동사	동일	is a
		분야/목적	is a sth(something) in/for
		D-t (기술)	RT [is a] USE technique
		D-a (적용)	RT [is a sth in] USE application
		D-m (도구)	RT method [is a sth in] USE
	일반동사	명명	call / name / entitle / dub
		규명	identify / establish / confirm / certify / verify / dub
비 정의적 문장	일반 동사	증명	demonstrate / show
		발견	find / discover/ detect
		기원	base on / derive
		도입	introduce / establish / launch / institute
		출현	emerge / appear / emanate / stand out
		위치	locate
		구성	comprise / comprise / include / consist of
		연결	access / associate / accompany / connect
		전달	correspond to / communicate
		역할	play role in / act as / have ability
		측정	assess / calculate / measure
		평가	evaluate
		관찰	observe / watch / monitor
		수행	conduct / process / carry out / implement achieve / accomplish / perform / make with
		이용	apply / use / employ / implement
		설치	equip with / kit out
		분해	degrade / decompose
		지속	continue / maintain / sustain
		소모	consume / absorb / exhaust / expend
		습득	acquire / gain / collect / get
		생산	produce
		원인	lead to / cause / motivate /arise / effect
		고려	consider as / contemplate / deliberate
		허락	permit / allow / grant / consent
		인식	recognize / acknowledge
		동기	interest in / attract / divert / captivate
		영향	affect / involve / influence / concern / change
		개선	advance / improve / develop / enhance / increase
		제한	restrict / limit / regulate
		예상	expect / anticipate / look forward to
		V-t (기술)	USE technique [일반동사] RT
		V-a (적용)	RT [일반동사 in/of] USE application
		V-m (도구)	RT method [일반동사] USE
형태적 구조		중심-복합명사	USE·RT
		수식-복합명사	RT·USE

<표 8 연관관계의 유형>

턴이기 때문에 한 단어로 관계를 명명할 수 없어 코드를 부여하였다. 비정의적 문장의 V-t/a/m도 마찬가지로 코드를 부여하였다. 복합명사로 이루어진 형태적 구조는 관련어의 위치에 따라서 관련어가 N_2일 때에는 중심-복합명사로, N_1일 때는 수식-복합명사로 명명하였다.

5. 결론

시소러스의 연관관계는 유형의 세분화에 관한 원칙과 방법론의 부재로 시소러스를 구축하는 기관에서 개별적인 분류를 사용하고 있다. 패싯지시어 모형에 관한 연구가 계속 되고 있지만 그 타당성을 뒷받침 할 실증적 사례연구가 역시 거의 찾아볼 수 없다. 본 연구에서는 Inspec에서 구축한 시소러스의 우선어와 관련어를 대상으로 IEL에서 제공하는 문헌에서 출현문장을 찾아 실증적 연관관계를 분석하였다. 그리하여 본 연구에서는 211개의 용어 쌍에 대해 43개의 연관 관계 유형을 도출하였으며 연관관계 용어 집합을 생성할 수 있었다.

그러나 본 연구에서 다룬 연관관계 유형은 과학기술분야(특히 공학)와 관련된 분석이므로 일반화를 논의하기 위해 타 시소러스에서도 적용 가능한 것인지 평가가 이루어져할 것이다.

또한 미처 분석하지 못한 우선어와 관련어가 한 문장에 출현하지 않은 용어 쌍도 시소러스 내에서의 관계를 분석하여 패턴을 찾고, 관련어로서의 정당성 여부를 고찰할 필요가 있을 것이다. 이에 대한 미흡한 부분은 향후 연구 대상이 될 것이다.

참고문헌

[1] D. Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz., “Reengineering Thesauri for new Applications: the AGROVOC Example”, Journal of Digital Information, vol.4, pp.1-23, 2004.

[2] 백지원, “용어관계의 분류 모형 개발에 관한 연구”, 정보처리학회지, 제23권, 제1호, pp.63-81, 2006.

[3] J. Aitchison, A. Gilchrist, Thesaurus construction: a practical manual. ASLIB, 1987.

[4] W. Schmitz-Esser, "New approaches in thesaurus applications", International Classification. 18(3), 143-147, 1991.

[5] S. Jones, M. Gatford, S. Robertdon, M. Hancock-Beaulieu, J. Secker, S. Walker, "Interactive Thesaurus Navigation: Intelligence Rules OK?", Journal of the American society for Information Science, 46(1), pp. 52-59, 1995.

[6] D. Tudhope, H. Alani, C. Jones, "Augmenting thesaurus relationshios: possibilities for retrieval", Journal of Digital Information, 1(8), pp. 1-20, 2001.

[7] 전말숙. “시소러스의 연관관계 유형에 관한 연구”, 정보처리학회, 제29권, 제1호, pp. 20-39. 1998.

[8] 백지원, “용어관계의 분류 모형 개발에 관한 연구”, 이화여대 박사학위 논문, 2005.

[9] B.J. Wielinga, A. Th, Schreiber et al. "From Thesaurus to Ontology". En Proceedings of the International Conference on Knowledge Capture, ACM Press, pp. 194-201, 2001.

[10] Sin-Jae Kang and Jong-Hyeok Lee, "Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora", In Proceedings of Workshop on Human Language Technology and Knowledge Management, Toulouse, France, pp.1-8, 2001.

[11] A. Kawtrakul, A. Imsombut, et al, "Automatic term Relationship Cleaning and Refinement for AGROVOC", In 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment, Vila Real, July 25-28, pp. 1-9, 2005.

[12] 최석두, 과학기술시소러스 구축연구, 한국과학기술정보연구원, 2007.

[13] 임지희, 최호섭, 배영준, 옥철영 외 3명, “면역학 시소러스 및 온톨로지 구축”, 한국정보과학회 언어공학연구회 05 제27회 한글 및 한국어 정보처리 학술대회, pp. 21-27, 2005.

[14] 정창후, 최성필, 이민호, 최윤수, “기술용어 간 관계 추출의 성능평가를 위한 반자동 테스트 컬렉션 구축 프레임워크 개발”, 한극콘텐츠학회논문지, 제10권 제2호, pp. 481-489, 2010.