단어의 의미연상을 이용한 시소러스 설계

Thesaurus Construction Using Word Association

한승희, 서울여자대학교 문헌정보학과, hanshee@gmail.com

Han, Seung Hee, Dept. of Library and Information Science, Seoul Women's University

본 연구에서는 단어의 의미연상을 이용하여 시소러스를 작성해봄으로써 탐색 시소러스 구축에 있어 단어연상검사법의 적용가능성을 살펴보았다. 문헌정보학 분야를 대상으로 단어연상검사를 실시한 후 자극어와 반응어간의 의미관계를 파악하고 반응어와 통제어휘를 비교 분석하였다. 실험 및 분석결과, 단어연상검사를 이용하면 다양한 연관관계 용어들을 시소러스에 포함시킬 수 있으며, 통제어휘집에 나타난 하위관계와 동등관계 용어들을 어느 정도 반영할 수 있다는 것을 확인하였다. 단어의 의미연상을 이용하여 구축된 탐색 시소러스는 정보검색환경에서 질의확장에 응용될 수 있다.

1 연구의 목적 및 필요성

미국정보표준화기구(National Information Standards Organization, NISO)에서는 1993년에 발표된 *ANSI/NISO Z39.19: Guidelines for the Construction, Format, and Management of Monolingual Thesauri*를 2005년에 개정하였는데, 이 지침에 대해 개정을 시도한 이유는 바로 정보환경의 변화에서 찾아볼 수 있다.

과거의 정보 이용자들은 원하는 정보를 직접 탐색하기 어려웠기 때문에, 정보전문가가 시소러스를 구축하고 이를 정보검색에 활용하였다. 그러나 온라인 정보검색 환경이 도입되고, 네트워크로 접근 가능한 데이터베이스가 증가하면서 최종 이용자가 시소러스와 같은 통제어휘집을 이용하여 원하는 정보를 직접 탐색할 수 있게 되었다. 이러한 환경에서 정보검색에서는 탐색 시소러스에 관심을 갖기 시작했다.

그러나, 기존의 시소러스 설계지침을 살펴보면 어휘통제를 목적으로 하기 때문에 도서관 업무 중심적(library-centric)이고, 텍스트 문서 중심적(text document-centric)이며, 인쇄자료 중심적(print-oriented)이다. 즉, 기존의 시소러스 설계지침은 현재의 정보환경의 변화를 반영하고 있지 못하고 있으며, 이것은 결국 이용자 비친화적(user-unfriendly)인 시소러스를 구축하는 결과를 가져온다.

시소러스가 정보검색에 적극적으로 활용되거나 이용자 친화적인(user-friendly) 형태가 되려면 우선적으로 시소러스를 구축할 특정 주제분야 정보 이용자의 정보요구나 이들이 자주 사용하는 개념간의 관계로 표현되는 영역지식을 반영할 필요가 있다. 이러한 방식으로 시소러스를 구축하기 위해 단어연상검사법을 적용할 수 있다.

본 연구에서는 단어의 의미연상을 이용하여 시소러스를 작성해봄으로써 탐색 시소러스 구축에 있어 단어연상검사법의 적용가능성을 확인하고자 한다. 이를 위해 문헌정보학 분야를 대상으로 단어연상검사를 실시한 후 자극어와

반응어간의 의미관계를 파악하고 반응어와 통제어휘를 비교 분석하였다.

2 단어의 의미연상

2.1 단어의 의미연상과 시소러스

어떤 단어가 주어졌을 때, 그것과 관련된 사항이 머릿속에 떠오르는 심리적 작용을 연상(association)이라고 하며, 특히 단어(구) 단위의 연상을 단어연상(word association)이라 한다(임지룡 외 옮김 2004). 단어연상을 적용한 단어연상검사(word association test)는 주로 심리학과 정신분석학에서 개인의 실세계를 표현하는 방법으로 사용되어 왔으며, 이를 통해 반응자의 언어기억능력과 사고과정, 감정상태, 인성을 이해하고자 한다(김태수 2000). 이 검사에서 피험자에게 제시되는 단어를 자극어(stimulus term)라고 하고 이에 대해 피험자가 연상해낸 단어들을 반응어(response word)라고 한다.

일반적으로 단어연상검사는 자극어에 대해 연상된 의미를 식별하거나 두 자극어간의 관계를 분석하기 위해 수행된다(Deese 1962). 반응어는 자극어의 연상표현에 대한 반응어 클러스터를 생성하는데, 두 자극어에 대한 반응어간의 유사성이 높으면 두 자극어간의 관계는 유사하다. 결국 반응어 클러스터는 자극개념(stimulus concepts)과 반응어간의 적합성과 관계에 대한 피험자의 인지적 이해수준을 나타낸다고 할 수 있다(Nielsen 1998).

이용자 중심의 정보 서비스를 제공하기 위해서는 이용자의 정보이용행태와 정보요구를 이해할 필요가 있는데, 영역 분석(domain analysis)을 통해 특정 영역에서의 정보이용행태를 파악할 수 있는데 이러한 방법의 하나로 단어연상을 이용하면 특정 분야나 이용자 집단에서 나타나는 영역 특정적(domain specific) 언어나 전문용어(jargon)의 사용 패턴을 확인할 수 있다(Nielsen 2001). 그러므로 단어연상검사를 통한 어휘간의 관계 이해라는 이러한 관점은 시소러스 구축 방법론으로서의 단어연상검사법 제안을 위한 근거가 된다.

단어연상검사를 이용해 시소러스를 구축하면 특정 학문영역의 맥락과 그 영역에 속한 이용자의 정보요구를 반영할 수 있다. 즉 특정 영역에 속한 이용자 집단의 용어 사용 패턴이나 용어간에 개념을 연관짓는 방식을 확인하는 것이 가능하므로 기존의 시소러스에 비해 확장성과 유연성이 크다.

일반적으로 정보검색 시스템 이용자들의 질의어는 이용자의 직관적인 단어연상결과를 표현한다. 그러므로 시스템 내에 축적된 질의어는 연상된 단어들의 집합이라고 볼 수 있으며 이 중에서 빈번하게 함께 연상된 단어들을 활용하여 시소러스를 구축하면 효과적으로 질의확장에 응용할 수 있다. 이러한 시소러스는 불분명한 정보요구를 가진 이용자로 하여금 다른 이용자들의 단어연상 패턴을 제시함으로써 나은 검색결과를 얻을 수 있도록 한다.

2.2 선행연구

문헌정보학 분야에서 단어연상검사를 이용한 최초의 연구자인 Kiss(1975)는 로겟 시소러스(Roget's Thesaurus)와 유사한 영어 연관 시소러스를 구축하기 위해 단어연상기법을 적용하였다. Pejtersen(1991)은 연상된 반응어의 중복도 수준에 따라 디스크립터를 구축하였다. 또한 Ornager(1995)는 디지털 사진이 수록된 이미지 데이터베이스의 탐색성능을 개선하기 위해 연상검사를 적용하였으며 실제 단어연상검사를 통해 시소러스를 구축하였다.

Nielsen(2001)은 영역 분석을 기초로 하여 특정 과업을 수행하고 있는 이용자들의 정보요구와 이용행태를 반영한 시소러스를 구축하기 위해 단어연상 실험을 하였는데, 세 명의 피험자를 대상으로 24개의 자극어에 대해 각

1분의 시간을 주고 하나의 자극어 당 두 개의 반응어를 쓰도록 한 후, 자극어와 반응어간의 관계를 동등, 계층, 연관관계보다 세분화하여 분석하였다.

한편, Greenberg(2001)는 구조화된 시소러스가 갖는 동등관계, 계층관계, 연관관계 정보를 자동질의확장에 적용하였으며, 정보검색의 성능향상을 위해 용어간의 의미관계를 적용하는 것이 효과적이라는 결론을 얻었다.

3 단어연상 실험

3.1 실험방법

단어연상 검사 방법으로는 크게 두 가지가 있는데, 하나는 자유연상검사(free association test)이고 다른 하나는 통제연상검사(controlled association test)이다. 자유연상검사는 말 그대로 자극어에 대해 피험자가 자유롭게 연상하도록 하는 검사법이며, 반대로 통제연상검사는 자극어에 대한 반응어를 의미 범주나 동의어, 특정 맥락 내에 있는 용어로 제한하는 방법을 말한다. 이 외에도 자극어를 피험자에게 제시하는 방식에 따라서도 검사 방법이 달라진다(Pejtersen 1991).

본 연구에서는 문헌정보학 분야 10개 자극어에 대해 문헌정보학 분야 석사과정 이상의 연구자 16명을 실험대상으로 하여 단어연상검사를 실시하였다.

자극어의 선정은 문헌정보학분야 3개 학회지에 출현한 색인어를 10개의 군집으로 나누어 분석한 유영준(2003)의 연구를 참조하였는데, 이 때 군집의 크기를 고려하여 크기가 큰 군집에서는 복수의 색인어를, 작은 군집들은 모아서 그 중에 한 개의 색인어를 임의로 선정하였다.

단어연상방법은 통제연상검사를 적용했는데, 피험자에게 1분 동안 자극어를 보고 연상되는 단어나 구를 수에는 제한이 없되, 문헌정보학으로 한정지어 연상할 것을 요구했다.

3.2 실험결과

10개의 자극어와 반응어에 대한 각종 통계는 <표 1>과 같다.

<표 1>에서 보는 바와 같이, 자극어에 대해 피험자들이 1분 동안 연상한 반응어의 평균은 약 5개이다. 한편, 특정 자극어에 대해 피험자들이 연상한 반응어 중 일부가 중복되어 나타났는데, 중복도가 큰 반응어일수록 자극어와

<표 1> 단어연상 실험결과: 자극어와 반응어간 통계

항목 \ 자극어	목록 규칙	공공 도서관	메타 데이터	이용자	지식 관리	자동 분류	전문가 시스템	장서 개발	계량 정보학	디지털 도서관	합계	평균
반응어 총 수	81	80	71	81	70	73	67	77	102	96	798	79.8
평균	5.1	5.0	4.4	5.1	4.4	4.6	4.2	4.8	6.4	6.0	49.9	5.0
고유 단어 수	42	47	33	40	42	40	42	48	58	66	458	45.8
빈도 2 이상	10	15	10	12	13	10	11	14	18	14	127	12.7
빈도 3 이상	8	10	8	7	6	6	7	7	11	7	77	7.7
빈도 5 이상	4	2	5	5	4	4	1	1	4	3	33	3.3
최고빈도	9	7	8	9	6	12	5	5	6	6	73	7.3
최고 빈도어	MARC	사서	XML	이용자 인터페이스	정보	클러스터링	인공지능	수서	인용	전자 도서관	※ 해당사항 없음	

의 연결 강도가 세다고 할 수 있다. 실험 결과, 중복을 제외한 고유한 반응어는 평균 약 45.8개이며, 중복도 2 이상의 단어는 전체 고유 반응어의 약 28%, 중복도 3 이상의 단어는 약 17%, 중복도 5 이상의 단어는 약 7% 정도인 반면, 한 번도 중복되지 않은 단어는 약 72% 정도를 차지해, 피험자간의 반응어 중복도는 낮은 것으로 나타났다. 이것은 피험자들이 같은 분야에서 연구를 하고 있다고 해도 이들 사용하는 용어는 굉장히 다양하다는 것을 의미한다. 예를 들어 자극어 '공공도서관'에 대해 '사서'라는 단어는 7명이 연상했지만, '아웃소싱', '대학도서관', '시립도서관', '무료', '시설' 등과 같은 대부분의 단어들은 중복되지 않고 한 번씩만 연상되었다.

4 실험결과 분석

4.1 자극어-반응어의 의미관계

자극어에 대한 반응어의 의미관계 분포를 알아보기 위해 자극어와 반응어간의 관계를 동등관계(USE, UF), 상위관계(BT), 하위관계(NT), 그리고 연관관계(RT)로 나누어 분석하고 각각의 비율을 계산하였는데, 그 결과는 <표 2>와 같다.

<표 2>에서 보는 바와 같이, 자극어와 반응어간의 관계 중 연관관계가 약 83%로 가장 높았고, 그 뒤를 이어 하위관계(11.6%), 동등관계(2.6%), 상위관계(2.4%) 순으로 나타났다. 이러한 결과는 단어연상이 주로 자극어에 대한 '연관성'과 '하위개념'을 기준으로 일어나며, 동의어나 상위의 개념은 주된 연상의 대상이 아니라는 것으로 해석할 수 있다. 특히 연관관계의 비율이 높은 이유는 연관관계 자체가 갖는 특성에 기인하는데, 연관관계는 계층관계나 동등관계가 아니면서 상당한 연관이 있는 관계를 모두 포함하며 관계에 대한 정의 자체가 단순하고 포괄적이기 때문이다.

이와 관련하여 *ASIS&T Thesaurus of Information Science, Technology and Librarianship*(2005, 이하 ASIST 시소러스)에서 10개의 자극어와 해당하는 디스크립터가 어떻게 용어간의 관계를 구성하고 있는지 확인하였다. 그 결과 연관관계가 전체의 50%를, 동등관계가 19.4%, 상위관계가 17.7%, 하위관계가 12.9%인 것으로 나타났다. 단어연상 시소러스와 비교해 볼 때 연관관계가 전체 용어관계에서 큰 비율을 차지한다는 점에서는 일치하나, ASIST 시소러스는 단어연상 시소러스에 비해 연관관계를 제외한 나머지 관계가 비슷한 비율로 구성되어 있음을 알 수 있다. 이 비교를 통해 단어연상 시소러스는 다양한 연관관계 용어를 풍부하게 표현하는데 적합한 기법이라는 것을 확인할 수 있다.

4.2 통제어휘와 연상단어 비교

연상된 반응어와 ASIST 시소러스의 용어를 비교한 결과는 <표 3>과 같다. ASIST 시소러

<표 2> 단어연상 실험결과: 자극어와 반응어간의 의미 관계

자극어 / 관계	목록 규칙	공공 도서관	메타 데이터	이용자	지식 관리	자동 분류	전문가 시스템	장서 개발	계량 정보학	디지털 도서관	합계	평균	비율
동등관계 (USE, UF)	0	0	1	4	2	0	0	1	1	3	12	1.2	2.6%
상위관계 (BT)	3	0	1	0	0	0	3	4	0	0	11	1.1	2.4%
하위관계 (NT)	4	7	5	3	1	8	0	8	11	6	53	5.3	11.6%
연관관계 (RT)	35	40	25	33	39	32	39	35	46	57	381	38.1	83.4%
합계	42	47	32	40	42	40	42	48	58	66	457	45.7	100.0%

<표 3> ASIST 시소러스 용어와 반응어와의 비교

	목록 규칙	공공 도서관	메타 데이터	이용자	지식 관리	자동 분류	전문가 시스템	장서 개발	계량 정보학	디지털 도서관	합계	평균	비율
ASIST 시소러스 용어 수	3	5	10	15	2	5	5	4	8	5	62	6.2	100.0%
ASIST 시소러스와 일치한 반응어 수	1	2	3	5	1	2	2	3	4	2	25	2.5	40.3%

스의 디스크립터별 평균 용어 수는 6.2개이며, 이 통제어휘와 연상단어 간 일치한 단어 수는 2.5개로 ASIST 시소러스 용어 전체의 약 40%에 불과했다. 단어연상에 의한 시소러스가 ASIST 시소러스에 비해 디스크립터별로 훨씬 많은 용어를 포함하고 있으나 두 시소러스간의 일치도는 낮게 나타났다. 반응어 클러스터는 이용자 각각의 배경지식이나 인지구조를 바탕으로 한 자유로운 연상의 결과로 인해 생성된 것이기 때문에 통제어휘에 비해 훨씬 다양한 연관 어휘를 포함할 수 있다.

또한 ASIST 시소러스에서 디스크립터에 속한 용어의 수가 많다고 해서 연상단어와의 일치도가 높은 것은 아니라는 것을 알 수 있다. 예를 들어 '장서개발'의 경우 ASIST 시소러스에서 용어 수는 4개이지만 이 중 3개의 단어가 연상단어와 일치하여 75%의 일치율을 보였다. 반면 '이용자'는 ASIST 시소러스에서 용어 수도 가장 많고(15개) 연상 시소러스와 일치하는 용어의 수도 가장 많지만(5개) 약 33%의 일치율을 나타냈다.

디스크립터별로 ASIST 시소러스의 용어와 일치하는 연상단어의 응답 중복빈도를 분석한 결과, 고빈도에서부터 저빈도까지 고르게 분포하고 있으며, 응답 중복빈도 1(한 명의 피험자가 응답한 것을 의미함)의 단어가 오히려 8개나 나타났다. 이것은 결국 응답 중복빈도가 높다고 해서 반드시 통제어휘에 포함되는 것은 아니라는 것을 보여준다.

디스크립터별로 ASIST 시소러스의 용어와 일치한 반응어의 관계정보를 분석해 본 결과 <표 4>와 같은 결과를 얻었다.

<표 4>에서 보는 바와 같이 피험자들은 ASIST 시소러스의 용어 중 하위관계에 있는 8개의 단어를 모두 연상하였으며, 또한 동등관계에 있는 단어의 50%를 연상하였다. 동등관계가 일반적인 단어연상 패턴이 아님에도 불구하고 50%의 일치율을 보인 것은 피험자들이 자극어에 대한 동등개념을 정확하게 이해하고 있었기 때문인 것으로 해석된다. 반면 상위관계와 연관관계에 있는 용어들의 일치율은 낮았는데, 그 이유는 연관관계의 경우 단어연상 결과가 워낙 다양하여 통제어휘에서 일치하는 단어를 찾아보기가 어렵고, 또한 상위관계에 있는 어휘로는 단어연상이 잘 이루어지지 않기 때문인 것으로 풀이된다. 이것은 결국 통제어휘가 하위관계와 동등관계는 잘 표현해주고 있으나 상위관계와 연관관계는 그렇지 못하다는 것을 의미한다.

<표 4> 의미관계에 따른 두 시소러스 용어간의 일치율 분석

	ASIST 시소러스 용어 수	연상 시소러스와 일치한 용어 수	비율
동등관계	12	6	50.0%
상위관계	11	2	18.2%
하위관계	8	8	100.0%
연관관계	31	9	29.0%

<표 5>는 용어 '전문가 시스템'에 대한 ASIST 시소러스의 용어 레코드와 단어연상을 통해 중복빈도 3 이상의 반응어만을 이용하여 작성한 용어 레코드의 예이다. ASIST 시소러스의 레코드와 비교해 볼 때, 용어간의 관계

정보 뿐만 아니라 연상의 중복빈도 정보까지 알게 되면 정보검색에서 가중치를 부여하여 검색할 수 있으므로 보다 정교한 검색 결과를 얻을 수 있다.

<표 5> '전문가시스템'에 대한 ASIST 시소러스와 단어연상 시소러스 비교

디스크립터	관계	ASIST 시소러스	관계	단어연상	빈도
전문가 시스템	uf	지식기반 시스템	bt	인공지능	5
	bt	인공지능	rt	데이터 베이스	4
	rt	지식획득		지식베이스	4
		지식베이스		질의응답 시스템	4
		지식 공학		의사결정	4
				주제지식	4
				전문가	3

5 결론

본 연구에서는 단어의 의미연상을 이용하여 시소러스를 작성해봄으로써 탐색 시소러스 구축에 있어 단어연상검사법의 적용가능성을 살펴보았다. 실험 및 분석결과, 단어연상검사를 이용하면 다양한 연관관계 용어들을 시소러스에 포함시킬 수 있으며, 통제어휘집에 나타난 하위관계와 동등관계 용어들을 어느 정도 반영할 수 있다는 것을 확인하였다. 단어의 의미연상을 이용하여 구축된 탐색 시소러스는 정보검색환경에서 질의확장에 응용될 수 있다.

단어연상을 통해 이용자의 용어 사용 패턴을 반영한 이러한 유형의 시소러스는 시맨틱 웹의 폭소노미(folksonomy)나 이용자 태깅, 온톨로지의 구축에도 적용될 수 있다.

이 연구의 후속연구로 연상 시소러스에 대한 탐색 실험을 통해 시소러스와 연상 시소러스간의 질의확장 성능을 비교 분석해 볼 필요가 있다.

일반적으로 단어연상에는 문화기반마다 연상에 차이가 있다는 '문화적 연상(cultural association)'이라는 개념이 있다. 이러한 이론을 바탕으로 다국어 시소러스 구축에 단어연상검사를 적용해 볼 필요가 있다.

참고문헌

김태수. 2000. **분류의 이해**. 서울: 문헌정보처리연구회.

유영준. 2003. **문헌정보학의 지식 구조에 관한 연구**. 박사학위논문. 연세대학교 대학원 문헌정보학과.

Deese, J. 1962. Form Class and Determinants of Association. *Journal of Verbal Learning and Verbal Behavior*, 2; 79-84.

Greenberg, Jane. 2001. Automatic Query Expansion via Lexical-Semantic Relationships. *JASIS*, 52(5): 402-415.

Kiss, G. R. 1975. An Associative Thesaurus of English: Structural Analysis of a Large Relevance Network. In Kennedy, A. and Wilkes, A. ed. *Studies in Long Term Memory*. London: Wiley.

Nielsen, Marianne-Lykke. 1998. The Word Association Test in the Methodology of Thesaurus Construction. *Advances in Classification Research*, 8: 43-58.

Nielsen, Marianne-Lykke. 2001. A Framework for Work Task Based Thesaurus Design. *Journal of Documentation*, 57(6): 774-797.

Ornager, S. 1995. The Newspaper Image Database: Empirical Supported Analysis of User's Typology and Word Association Clusters. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 212-218.

Pejtersen, A. Mark. 1991. *Interfaces Based on Associative Semantics for Browsing in Information Retrieval*. Denmark: Riso Laboratory.

Redmond-Neal, Alice and Marjorie M. K. Hlava ed. 2005. *ASIS&T Thesaurus of Information Science, Technology and Librarianship*. New Jersey: Information Today, Inc.

Tsuji, Y. 편. 임지룡, 요시모토 하지메, 이은미, 오카 도모유키 옮김. 2004. **인지언어학 키워드 사전**. 서울: 한국문화사.