단어연상검사법을 이용한 탐색 시소러스 구축에 관한 실험적 연구

Searching Thesaurus Construction with Word Association Test: A Pilot Study

한 승 희(Seung-Hee Han)*

목 차

초 록

본 연구에서는 단어의 의미연상을 이용하여 시소러스를 구축하고 ,이 시소러스에 대해 탐색 시소러스로서의 기능성을 확인하기 위해 질의확장 실험을 수행하였다 . 연상 시소러스 구축을 위해 문헌정보학 분야를 대상으로 단어연상검사를 실시한 후 자극어와 반응어간의 의미관계를 파악하고 반응어와 기존 시소러스의 디스크립터를 비교 분석하였다. 실험 및 분석결과, 단어연상검사를 이용하여 시소러스를 구축하면 기존의 시소러스에 비해 연관관계 용어들을 시소러스에 다양하게 반영할 수 있으며 ,통제어휘집에 나타난 하위관계와 동등관계 용어들을 어느 정도 반영할 수 있다는 것을 확인하였다 . 또한 질의확장 실험결과 단어연상 시소러스가 기존 시소러스에 비해 비교적 우수한 성능을 보여 단어연상 시소러스가 정보검색환경에서 질의확장에 응용될 수 있음을 증명하였다 .

ABSTRACT

The purpose of this pilot study is to construct a searching thesaurus with word association test in the library and information science field and to confirm it's functionality as searching aids through query expansion experiments. The test results were analyzed to four types of relationship between stimulus words and response words, and the terms of association thesaurus were compared with descriptors of an existing thesaurus. The test results show that the word association test is a fruitful method to identify many related terms and narrower and equivalent terms in some degree to the stimulus terms. Furthermore, in the query expansion experiment, the performance of association thesaurus was better than that of an existing thesaurus, This result demonstrates that word association thesaurus can apply to query expansion.

키워드: 단어연상검사법, 탐색 시소러스, 단어연상 시소러스, 시소러스 구축, 질의확장
Word Association Test, Searching Thesaurus, Word Association Thesaurus, Thesaurus Construction, Query Expansion

* 서울여자대학교 사회과학대학 문헌정보학과 초빙강의교수(hanshee@gmail.com)
논문접수일자 2006년 8월 15일
게재확정일자 2006년 9월 15일

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, 미국 내·외에서 널리 활용되고 있는 이 지침에 대해 개정을 시도한 이유는 바로 정보이용 환경의 변화에서 찾아볼 수 있다.

과거의 정보 이용자들은 자신이 원하는 정보를 직접 탐색하기 어려웠기 때문에, 사서나 색인전문가와 같은 정보전문가가 시소러스를 구축하고 이를 정보검색에 이용하였다. 그러나 온라인 정보검색이 도입되고, 접근 가능한 웹 사이트나 데이터베이스가 증가하면서 최종 이용자가 시소러스와 같은 통제어휘집을 직접 이용하여 원하는 정보를 검색할 수 있게 되었다. 이에 따라 정보 시스템 개발자들은 최종 이용자의 정보검색을 지원할 수 있는 시소러스의 가치를 중요하게 인식하기 시작하였다. 즉 시소러스는 검색시스템에서 이용자의 질의구축이나 질의확장에 도움을 줄 수 있는 탐색도구(search aids)로서 인식되고 있다.

탐색엔진을 이용한 인터넷 정보검색에서는 기본적으로 탐색자의 주제지식을 상당부분 전제로 하며 이것은 탐색용 시소러스와 같은 도구를 필요로 함을 의미한다. 장차 온라인 정보검색이 보편화되고 메타데이터를 통한 인터넷 검색환경이 개선되리라 보이지만 현재로서는 제어어휘와 자연언어를 연결하고 이용자 스스로 탐색할 수 있는 탐색용 시소러스에 의한 검색이 효과적이다. 이 시소러스에서는 다양하고 복잡한 색인어휘를 평가하고 일상적으로 이용자가 사용하는 용어와 그 관계를 규정하는 것이 중요하다(김태수 2000).

그러나, ANSI/NISO의 Z39.19 표준뿐만 아니라, 기존의 시소러스 설계지침을 살펴보면, 그 성격이 도서관 업무 중심적(library-centric) 이고, 텍스트 문서 중심적(text document-centric)이며, 인쇄자료 중심적(print-oriented)인데, 이것은 전통적으로 시소러스의 구축이 도서관의 어휘제어를 주목적으로 하였기 때문이라고 할 수 있다. 즉, 기존의 시소러스 설계지침은 정보환경의 변화를 반영하지 못하고 있으며, 이러한 문제는 결국 비효율적인 시소러스의 구축뿐만 아니라, 이용자에게 있어 비친화적인(user-unfriendly) 시소러스의 이용을 초래한다.

이러한 관점에서 볼 때, 시소러스를 구축할 때 고려해야 할 두 개의 큰 축이라 할 수 있는 어휘의 '제어'와 이용자의 '이용' 중 무엇을 중심으로 하여 시소러스를 구축하는가에 따라 시소러스의 성격은 크게 달라질 수 있다. 시소러스가 정보검색에 적극적으로 활용되거나 이용자 친화적인(user-friendly) 형태가 되려면 우선적으로 시소러스를 구축할 특정 주제 분야 정보 이용자의 정보요구나 이들이 자주 사용하는 개념간의 관계로 표현되는 영역 지식(domain knowledge)을 반영할 필요가 있다.

시소러스가 특정 주제 분야 정보 이용자들의 영역 지식을 반영하려면 이용자들이 개념간의 관계를 어떻게 규정짓고 있는지를 확인해야 한다. 이용자들의 인지구조 속에서 특정 주제 분야를 구성하고 있는 개념간의 관계를 파악하기 위한 방법으로 단어연상검사법을 적용할 수 있다. 본 연구에서는 단어의 의미연상을 이용하

여 시소러스를 작성해봄으로써 '이용'을 위한 탐색 시소러스 구축에 있어 단어연상검사법의 적용가능성을 확인하고자 한다.

2. 단어의 의미연상과 시소러스

2.1 단어의 의미연상과 단어연상검사법

어떤 단어가 주어졌을 때, 그것과 관련된 사항이 머릿속에 떠오르는 심리적 작용을 연상(association)이라고 하며, 철학, 심리학, 정신분석의 영역에서 중시되고 있는 개념으로, 유추나 비유 등 확산적 사고(divergent thinking)의 원동력이 되는 기본적인 인지능력의 하나이다. 이 중에서 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 word)라고 하고 이에 대해 피험자가 연상해낸 단어들을 반응어(response word)라고 한다.

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 자극 개념(stimulus concepts)과 반응어간의 적합성과 관계에 대한 피험자의 인지적 이해수준을 나타낸다고 할 수 있다(Nielsen 1998).

2.2 단어연상검사법과 탐색 시소러스 구축

이용자 중심의 정보 서비스를 제공하기 위해서는 이용자의 정보이용행태와 정보요구를 이해할 필요가 있는데, 영역 분석을 통해 특정 영역에서의 정보이용행태를 파악할 수 있다. Hjørland (2002)는 영역 분석을 위한11 가지 접근법을 제시하였는데, 그 중 하나가 바로 시소러스 구축이다. 시소러스는 주로 영역 특정적인(domain-specific) 어휘로 구성되고, 어휘들을 설정하는 방법론 역시 영역 분석의 한 형태로 볼 수 있기 때문에 이를 통해 특정 주제 영역을 분석하고 그 지식구조를 규명하는 것이 가능하다.

단어연상을 이용하면 특정 분야나 이용자 집단에서 나타나는 영역 특정적 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 실제 용어 사용 패턴이나 용어 간의 개념관계를 연관짓는 방식을 확인하는 것이 가능하므로 기존의 시소러스 구축방식에 비해 확장성과 유연성이 크다.

단어연상검사를 통해 시소러스를 구축할 때 얻을 수 있는 장점은 다음과 같다(Spiteri 2005).

첫째, 이용자 주도적으로 디스크립터를 생성할 수 있다. 예를 들어, 대다수의 피험자에 의해 반응어로 연상된 단어들이 시소러스 디스크립터가 될 수 있다. 이러한 디스크립터는 영역 특정적 어휘가 되며, 이러한 어휘들이 정보검색 시스템 이용자에게 제공됨으로써 검색 효율성을 향상시킬 수 있다.

둘째, 이용자 주도적으로 디스크립터 계층을 생성할 수 있다. 자극어에 대해 많은 피험자들이 공통적으로 표현한 반응어의 속성, 자질, 특성 등과 같은 정보는 자극어와 반응어간의 관계를 설명하는 데 좋은 정보가 된다.

셋째, 이용자 주도적으로 단어간 관계 범주를 생성할 수 있다. 피험자가 규정한 자극어와 반응어간의 관계는 기존의 시소러스 구축 지침에서 보는 관계 정의와 다를 수 있으며, 이용자의 인지구조를 반영하여 자극어와 반응어간의 관계를 규정할 수 있다.

단어연상검사를 시소러스 구축 시에 적용할 때 두 가지 효과를 얻을 수 있다. 하나는 연상된 단어들을 얻게 되어 어휘를 수집할 수 있고, 이용자에 의해 규정된 단어간의 관계정보를 통해 수집된 어휘를 조직할 수 있다.

일반적으로 정보검색 시스템 이용자들의 질의어는 이용자의 직관적인 단어연상결과를 표현한다. 예를 들어 한 이용자가 '메타데이터'라는 질의어와 더불어 'XML'이라는 질의어를 추가로 입력하였다면 이것은 결국 탐색자가 '메타데이터'라는 정보요구에 대해 'XML'이라는 용어를 연상한 결과라고 볼 수 있다. 이러한 관점에서 시스템 내에 축적된 질의어쌍은 연상된 단어들의 집합이라고 볼 수 있으며, 이 중에서 빈번하게 함께 연상된 단어들을 활용하여 탐색 시소러스로 구축하면 효과적으로 질의확장에 응용할 수 있다. 이러한 시소러스는 불분명한 정보요구를 가진 이용자로 하여금 다른 이용자들의 단어연상 패턴을 제시하거나 특정 연상단어를 추천함으로써 나은 검색결과를 얻을 수 있도록 한다.

2.3 선행연구

2.3.1 정보검색과 시소러스

정보탐색 및 검색 과정에서 이용자가 직면하는 가장 큰 과제는 바로 질의를 만들어내기 위해 탐색어를 선정하는 일이라 할 수 있다. 시소러스는 질의 구축과 확장에 있어 탐색어 선정을 위한 하나의 중요한 도구이다. 탐색 시소러스는 다음과 같은 네 가지 기능을 한다 (Nielsen 1998).

첫째, 탐색 시소러스는 탐색자가 시소러스의 내용에 접근하도록 돕는다. 시소러스는 많은 양의 디스크립터를 보유하고 있는데, 이 디스크립터는 이용자로 하여금 시소러스의 내용에 접근 할 수 있도록 하는 접근점이 된다.

둘째, 탐색 시소러스는 다른 개념간의 관계를 탐색자에게 알려줌으로써 개념에 대해 명확히 이해할 수 있도록 돕는다.

셋째, 탐색 시소러스는 탐색자의 정보요구를 이해하고 표현할 수 있도록 돕는다.

넷째, 탐색 시소러스는 정보검색의 질의확장에 사용될 수 있다.

Shiri와 Revie의 연구(2001)에서는 정보검색에서의 시소러스에 대해 이용자연구의 관점을 적용하였다. 이들은 이용자가 질의어를 작성하여 탐색을 수행하는 과정에서 시소러스를

참조하는 행태에 대해 분석하여 웹 데이터베이스에서 이용자가 질의를 형성하고 확장하는 과정에서 온라인 시소러스가 어떠한 역할을 하는지에 대해 연구하였다. 일반적으로 이용자들이 정보탐색과정에서 시소러스를 활용하는 과정을 〈그림 1〉과 같이 나타냈다.

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

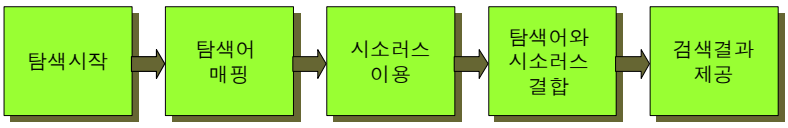
2.3.2 단어연상검사를 이용한 시소러스 구축

단어연상검사를 적용한 국내 연구의 대부분은 심리학, 교육학, 인지언어학 분야에서 이루어져 왔다. 특히 이찬규(2002)는 고등학생, 대학생, 일반인 세 집단을 대상으로 20개 일반 어휘에 대한 단어연상의 조사 결과를 분석하여 한국인의 연상 과정과 방식, 그리고 그 결과를 분석하였다. 연상 결과의 분석은 언어적 연상과 추상적 연상으로 구분하였으며, 사회·문화적 원인도 변인으로 보아 분석하였다.

문헌정보학 분야에서 단어연상검사를 이용한 최초의 연구자인 Kiss(1975)는 로젯 시소러스(Roget's Thesaurus)와 유사한 영어 연관 시소러스(associative thesaurus)를 구축하기 위해 단어연상기법을 적용하였다. 이 연구에서 그는 단어연상을 개념간의 적합성의 정도를 나타내는 척도라고 정의한 후, 연상된 단어간에 링크를 연결하여 단어연상 시소러스 네트워크(word association thesaurus network)를 구축하였는데, 단어간의 링크를 연결하기 위해 자극-반응 연관성 확률(stimulus-response associative probability)의 측정치를 반영하였다.

Nielsen(1998)의 연구는 단어연상검사를 통해 시소러스를 구축한 본격적인 연구라고 할 수 있다. 이 연구에서는 식품산업분야에 종사하는 아홉 명의 연구자들을 대상으로 단어연상검사를 실시하였는데, 피험자 간의 반응어에 대한 중복도가 평균적으로 약 31%에 그쳐 낮은 중복도를 나타냈다. 이 연구에서는 단어연상을 통한 시소러스 구축이 연관관계 어휘를 수집하는데 매우 효과적인 반면, 동등관계 어휘에 대해서는 그 효과가 적다는 결론을 얻었다. 이후의 후속연구(Nielsen, 2002)에서는 이용자의 태스크 분석을 통해 시소러스를 구축하는 과정에서 이용자 영역 지식의 분석을 위해 단어연상검사법을 적용하였다. 영역 분석을 통해 영역 특정적인 어휘나 전문용어를 확인할 수 있다는 가정하에서 진행된 이 연구에서는 세 명의 제약산업 분야 연구자들을 대상으로 24개의 자극어에 대해 단어연상검사를 실시한 후 자극어와 반응어간의 관계를 동등, 계층, 연관관계보다 세분화하여 분석하여 시소러스를 구축하였다.

Pejtersen(1991)은 연상된 반응어의 중복도



〈그림 1〉 정보검색과정에서의 시소러스 활용 단계

수준에 따라 디스크립터를 구축하였다. 또한 Ornager(1995)는 디지털 사진이 수록된 이미지 데이터베이스의 탐색성능을 개선하기 위해 연상검사를 적용하였으며, 실제 단어연상검사를 통해 시소러스를 구축하였다. 또한 Sinopalnikova (2003)은 러시아 어휘를 대상으로 워드넷(Word Net)과 같은 어휘집을 구축하기 위한 방법으로 단어연상을 적용하였다.

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 제한하거나 연상시간을 제한하는 방법을 말한다. 이 외에도 자극어를 피험자에게 제시하는 방식에 따라서도 검사 방법이 달라지는데, 실험 목적에 따라 다양한 검사방법을 적용해 본 결과, 일관적인 결과를 얻지 못해 특정 목적에 적합한 검사방법을 가려내는 데 실패했다(Pejtersen 1991).

본 연구에서는 문헌정보학 분야에서 추출한 10개의 자극어에 대해 문헌정보학 분야 석사과정 이상의 연구자 16명을 대상으로 하여 단어연상검사를 실시하였다. 자극어는 문헌정보학 분야 3개 학회지(『한국문헌정보학회지』, 『정보관리학회지』, 『한국도서관정보학회지』)에 출현한 색인어를 대상으로 클러스터링 기법을 이용하여 10개의 군집으로 나누어 분석한 유영준(2003)의 연구를 참조하여 선정하였다. 이 때 군집의 크기를 고려하여 크기가 큰 군집에서는 두 개 이상의 색인어를, 작은 군집들은 모아서 그 중에 한 개의 색인어를 임의로 선정하였다.

단어연상방법은 통제연상검사법과 자유연상검사법의 중간형태를 적용하였는데, 피험자에게 1분 동안 자극어를 보고 연상되는 명사로 된 단어나 구를 연상하여 기술하도록 하였는데, 연상한 단어의 수에는 제한이 없되, 문헌정보학으로 한정지어 연상하도록 요구하였다.

3.2 반응어 특성 분석

자극어별 반응어의 총 수와 그 평균은 〈표 1〉과 같다. 〈표 1〉에서 보는 바와 같이, 16명의 피험자들이 한 자극어에 대해 평균적으로 반응한 단어는 약 80개이며, 또한 개별 자극어에 대해 한 피험자가 1분간 연상한 반응어 수는 평균 5개로 나타났다. 가장 많은 반응어를 얻어낸 자극어는 '계량정보학'인데, 이러한 결과는, 검사 후 가진 인터뷰 결과, 피험자 중 일부가 계량정보학 분야에 관심이 많아, 다른 자극어에 비해 단어의 연상이 쉬워 많은 반응어를 연상해낸 것으로 풀이된다. 또한 가장 적은 반응어를 얻어낸 자극어는 '전문가시스템'이었는데, 인터뷰에서 피험자들 대부분이 이 자극어에 대해 연상이 가장 어려웠다고 언급하였다.

한편, 특정 자극어에 대해 피험자들이 연상한 반응어 중 일부가 중복되어 나타났는데, 중

〈표 1〉 자극어별 반응어 총 수와 평균

구분 / 자극어	반응어 총 수	평균반응어 수
목록규칙	81	5.1
공공도서관	80	5
메타데이터	71	4.4
이용자	81	5.1
지식관리	70	4.4
자동분류	73	4.6
전문가시스템	67	4.2
장서개발	77	4.8
계량정보학	102	6.4
디지털도서관	96	6
합계	798	49.9
평 균	79.8	5

복도가 큰 반응어일수록 많은 피험자들이 공통으로 연상한 단어이기 때문에 자극어와의 연결강도가 세다고 할 수 있다. 실험 결과, 〈표 2〉에서 보는 바와 같이, 연상의 중복을 제외한 고유한 반응어는 전체 반응어의 57.4%를 차지해, 고유 반응어의 비율이 전체 반응어의 절반 이상을 차지한다는 것을 알 수 있다. 또한 고유 반응어 중 두 명 이상의 피험자에 의해 중복하여 연상된 반응어의 비율은 전체 고유 반응어의 27.7%를 차지해, 피험자에 의해 중복되지 않고 한 번씩만 연상된 반응어가 전체 고유 반응어 중 약 73%를 차지하는 것으로 나타났다. 이것은 피험자들이 같은 분야에서 연구를 하고 있다고 해도 이들의 인지구조 속에 포함되어 있는 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.

〈표 2〉 고유 반응어 수와 고유 반응어 중 중복 반응어 수

반응어 특성 / 자극어	반응어 총 수	고유 반응어 수	비 율	고유 반응어 중 중복 반응어 수	비 율
목록규칙	81	42	51.9%	10	23.8%
공공도서관	80	47	58.7%	15	31.9%
데이터	71	33	46.5%	10	30.3%
이용자	81	40	49.4%	12	30.0%
지식관리	70	42	60.0%	13	31.0%
자동분류	73	40	54.8%	10	25.0%
전문가시스템	67	42	62.7%	11	26.2%
장서개발	77	48	62.3%	14	29.2%
계량정보학	102	58	56.9%	18	31.0%
디지털도서관	96	66	68.7%	14	21.2%
평 균	79.8	45.8	57.4%	12.7	27.7%

또한 중복된 반응어의 빈도분포를 살펴보면 〈표 3〉과 같다. 〈표 3〉에서 보는 바와 같이, 중복도가 높아질수록 중복도를 나타내는 단어의 비율은 낮아지는 것으로 나타났다. 세 명 이상의 피험자에 의해 중복된 반응어가 전체 고유 반응어의 16.8%를 나타내었고, 다섯 명 이상의 피험자에 의해 중복된 반응어는 전체 고유 반응어의 7.2%에 그쳤다. 이러한 결과는 피험자 간에 중복으로 연상한 단어의 비율이 전체 반응어를 기준으로 할 때 높지 않으며, 이것은 곧, 앞에서 언급한 바와 같이, 피험자 간 반응어의 다양성을 증명해주는 결과라고 할 수 있다. 실험 결과, 평균 최고중복도는 7.3이며, 〈표 3〉에서 보는 바와 같이 자극어 '자동분류'에 대해 '클러스터링'이라는 반응어가 가장 높은 중복도(12)를 보였으며, 그 뒤로 자극어 '이용자'에 대해 '이용자 인터페이스'라는 반응어가, '목록규칙'이라는 자극어에 대해 반응어 'MARC'가 중복도 9를 나타냈다.

4. 실험결과 분석

4.1 자극어-반응어의 의미관계

자극어에 대한 반응어의 의미관계 분포를 알아보기 위해 자극어와 반응어간의 관계를 동등관계(SYN), 상위관계(BT), 하위관계(NT), 그리고 연관관계(RT)로 나누어 분석하고 각각의 비율을 계산하였다. 〈표 4〉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 계층관계가 아니고 동등관계도 아니면서 상당한 연관이 있는 관계를 모두 포함하며

〈표 3〉 중복 반응어의 빈도분포와 최고 중복도를 나타낸 반응어

중복도 / 자극어	2 이상	3 이상	5 이상	최고값	최고 중복도를 나타낸 단어
목록규칙	10	8	4	9	MARC
공공도서관	15	10	2	7	사서
메타데이터	10	8	5	8	XML
이용자	12	7	5	9	이용자인터페이스
지식관리	13	6	4	6	정보
자동분류	10	6	4	12	클러스터링
전문가시스템	11	7	1	5	인공지능
장서개발	14	7	1	5	수서
계량정보학	18	11	4	6	인용
디지털도서관	14	7	3	6	전자도서관
합 계	127	77	33	73	※ 해당사항 없음
평 균 (비 율)	12.7 (27.7%)	7.7 (16.8%)	3.3 (7.2%)	7.3	

〈표 4〉 단어연상 실험결과: 자극어와 반응어간의 의미 관계

관계 / 자극어	동등관계 (SYN)	상위관계 (BT)	하위관계 (NT)	연관관계 (RT)	합계
목록규칙	0	3	4	35	42
공공도서관	0	0	7	40	47
메타데이터	1	1	5	25	32
이용자	4	0	3	33	40
지식관리	2	0	1	39	42
자동분류	0	0	8	32	40
전문가시스템	0	3	0	39	42
장서개발	1	4	8	35	48
계량정보학	1	0	11	46	58
디지털도서관	3	0	6	57	66
합 계	12	11	53	381	457
평 균	1.2	1.1	5.3	38.1	45.7
비 율	2.6%	2.4%	11.6%	83.4%	100.0%

관계에 대한 정의 자체가 단순하고 포괄적이기 때문이다.

이와 관련하여 "*ASIS&T Thesaurus of Information Science, Technology and Librarianship*"(Redmond-Neal and Hlava ed. 2005, 이하 ASIST 시소러스)을 이용하여 10 개의 자극어에 해당하는 디스크립터가 시소러스 상에서 어떻게 용어간의 관계를 구성하고 있는지 확인하였다. 그 결과, 〈표 5〉에서 보는 바와 같이, 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 적합하다는 것을 확인할 수 있다.

4.2 통제어휘와 연상단어 비교

연상된 반응어와 ASIST 시소러스의 용어를 비교한 결과는 〈표 6〉과 같다. ASIST 시소러스의 디스크립터 별 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 디스크립터 별로 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

〈표 5〉 ASIST 시소러스 디스크립터의 의미 관계 구성

관계 / 디스크립터	UF	BT	NT	RT	합계
cataloging rules	0	1	1	1	3
public libraries	1	1	0	3	5
metadata	2	1	0	7	10
users	2	1	1	11	15
knowledge management	1	1	0	0	2
automatic classification	0	2	0	3	5
expert systems	1	1	0	3	5
collection development	1	1	2	0	4
informetrics	2	1	4	1	8
digital libraries	2	1	0	2	5
합계	12	11	8	31	62
비율	19.4%	17.7%	12.9%	50.0%	100.0%

〈표 6〉 ASIST 시소러스 어휘와 반응어와의 비교

구분 / 자극어	ASIST 시소러스 용어 수	ASIST 시소러스와 일치한 반응어 수
목록규칙	3	1
공공도서관	5	2
메타데이터	10	3
이용자	15	5
지식관리	2	1
자동분류	5	2
전문가시스템	5	2
장서개발	4	3
계량정보학	8	4
디지털도서관	5	2
합계	62	25
평균	6.2	2.5
비율	100.0%	40.3%

연상단어와 일치하여 75%의 일치율을 보였다. 반면 '이용자'는 ASIST 시소러스에서 용어 수도 15개로 가장 많고, 단어연상 시소러스와 일치하는 용어의 수도 5개로 가장 많지만 약33%의 일치율을 나타냈다.

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

1) 한 명의 피험자가 응답한 것을 의미함.

또한, 디스크립터별로 ASIST 시소러스의 용어와 일치한 반응어의 관계정보를 분석해 본 결과 〈표 7〉과 같은 결과를 얻었다. 〈표 7〉에서 보는 바와 같이 피험자들은 ASIST 시소러스의 용어 중 하위관계에 있는 8개의 단어를 모두 연상하였으며, 또한 동등관계에 있는 단어의 50%를 연상하였다. 동등관계가 일반적인 단어 연상 패턴이 아님에도 불구하고 50%의 일치율을 보인 것은 피험자들이 자극어에 대한 동등 개념을 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 8〉은 '전문가 시스템'에 대한 ASIST 시소러스의 용어 레코드와 단어연상을 통해 중복도 3 이상의 반응어만을 이용하여 작성한 용어 레코드의 예이다. ASIST 시소러스의 레코드와 비교해 볼 때, 용어간의 관계정보 뿐만 아니라 연상의 중복빈도 정보까지 알게 되면 정보검색에서 가중치를 부여하여 검색할 수 있으므로 보다 정교한 검색 결과를 얻을 수 있다.

5. 질의확장 실험

피험자들의 단어연상을 통해 구축된 시소러스가 탐색 시소러스로서의 기능을 수행할 수 있는가를 확인하기 위해 질의확장 실험을 수행

〈표 7〉 의미관계에 따른 두 시소러스 어휘간의 일치율 분석

구분 / 관계	ASIST 시소러스 용어 수	단어연상 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 8〉 '전문가시스템'에 대한 ASIST 시소러스와 단어연상 시소러스 비교

디스크립터	관계	ASIST 시소러스	자극어	관계	단어연상	빈도
expert systems	UF	knowledge based system	전문가시스템	BT	인공지능	5
	BT	artificial intelligence		RT	데이터베이스	4
	RT	knowledge acquisition			지식베이스	4
		knowledge base			질의응답시스템	4
		knowledge engineering			의사결정	4
					주제지식	4
					전문가	3

하였다. 이 때, 기존의 시소러스의 질의확장 성능과 비교하기 위해 ASIST 시소러스를 이용하여 마찬가지로 질의확장 실험을 수행하였다. 단어연상 시소러스의 경우 중복도 2 이상의 단어들만을 대상으로 질의를 확장하였다.

질의확장 실험은 인터넷 검색엔진 구글(http://www.google.com)에서 실시하였다. 일반적으로 문헌정보학 관련 질의확장 실험에서 자주 사용되는 초록 데이터베이스인 LISA(Library and Information Science Abstract)가 아닌 구글을 선택한 이유는 언어적 문제에 있다. 즉, ASIST 시소러스는 영어로 되어 있고, 단어연상 시소러스는 한글로 되어 있어 LISA를 이용하면 한글 문헌을 검색할 수 없기 때문이다. 구글의 경우는 영어 문헌과 한글 문헌을 동시에 검색할 수 있으므로 본 연구에서 진행하고자 하는 실험 환경을 만족시킬 수 있다.

검색결과의 적합성 판정은 인터넷 검색엔진의 성능평가척도로 주로 사용되는 P@10을 이용하였다. 즉, 검색결과로 보여지는 상위 10개 문헌에 대한 정확률(precision)을 평가하는 것을 말한다. 성능의 평가 결과는 〈표 9〉, 〈표 10〉과 같다.

실험 결과, 전체 평균을 보았을 때, ASIST 시소러스의 성능이 49.2%, 단어연상 시소러스의 성능이 56.7%로, 단어연상 시소러스가 질의확장에 있어 더 나은 성능을 나타냈다. 또한 디스크립터 별로 그 성능을 확인한 결과, 10개의

〈표 9〉 질의확장 실험 결과: ASIST 시소러스

디스크립터	관계 / 구분	동등관계	상위관계	하위관계	연관관계	전 체
cataloging rules	적합문헌 수	해당없음	0	7	4	11
	p@10 평균	해당없음	0.0%	70.0%	40.0%	36.7%
public libraries	적합문헌 수	7	9	해당없음	6	22
	p@10 평균	70.0%	90.0%	해당없음	20.0%	44.0%
metadata	적합문헌 수	6	5	해당없음	26	37
	p@10 평균	30.0%	50.0%	해당없음	37.1%	37.0%
users	적합문헌 수	12	0	1	48	61
	p@10 평균	60.0%	0.0%	10.0%	43.6%	40.7%
knowledge management	적합문헌 수	10	3	해당없음	해당없음	13
	p@10 평균	10.0%	30.0%	해당없음	해당없음	65.0%
automatic classification	적합문헌 수	해당없음	11	해당없음	18	29
	p@10 평균	해당없음	55.0%	해당없음	60.0%	58.0%
expert systems	적합문헌 수	6	8	해당없음	16	30
	p@10 평균	60.0%	80.0%	해당없음	53.3%	60.0%
collection development	적합문헌 수	2	8	14	해당없음	24
	p@10 평균	20.0%	80.0%	70.0%	해당없음	60.0%
informetrics	적합문헌 수	10	6	12	0	28
	p@10 평균	50.0%	60.0%	30.0%	0.0%	35.0%
digital libraries	적합문헌 수	11	10	해당없음	7	28
	p@10 평균	55.0%	100.0%	해당없음	35.0%	56.0%
평 균	적합문헌 수	6.4	6	3.4	12.5	28.3
	p@10 평균	35.5%	54.5%	18.0%	28.9%	49.2%

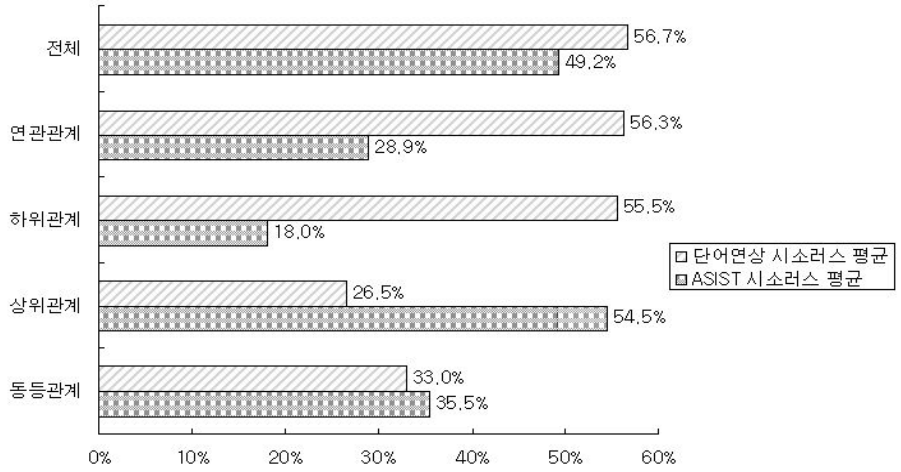
〈표 10〉 질의확장 실험 결과: 단어연상 시소러스

자극어	구분 \ 관계	동등관계	상위관계	하위관계	연관관계	전체
목록규칙	적합문헌 수	해당없음	6	13	29	48
	p@10 평균	해당없음	65.0%	60.0%	41.4%	48.0%
공공도서관	적합문헌 수	해당없음	해당없음	24	46	70
	p@10 평균	해당없음	해당없음	60.0%	63.6%	46.7%
메타데이터	적합문헌 수	6	4	13	26	49
	p@10 평균	60.0%	40.0%	65.0%	52.0%	54.4%
이용자	적합문헌 수	8	해당없음	10	35	53
	p@10 평균	40.0%	해당없음	33.3%	50.0%	44.2%
지식관리	적합문헌 수	14	해당없음	7	65	86
	p@10 평균	70.0%	해당없음	70.0%	65.0%	66.2%
자동분류	적합문헌 수	해당없음	해당없음	36	26	62
	p@10 평균	해당없음	해당없음	60.0%	65.0%	62.0%
전문가시스템	적합문헌 수	해당없음	8	해당없음	59	67
	p@10 평균	해당없음	80.0%	해당없음	59.0%	60.9%
장서개발	적합문헌 수	7	16	12	54	89
	p@10 평균	70.0%	80.0%	60.0%	60.0%	63.6%
계량정보학	적합문헌 수	해당없음	해당없음	38	40	78
	p@10 평균	해당없음	해당없음	63.3%	33.3%	43.3%
디지털도서관	적합문헌 수	18	해당없음	25	66	109
	p@10 평균	90.0%	해당없음	83.3%	73.3%	77.9%
평균	적합문헌 수	5.3	3.4	17.8	44.6	71.1
	p@10 평균	33.0%	26.5%	55.5%	56.3%	56.7%

디스크립터 모두 단어연상 시소러스의 질의확장 성능이 더욱 우수한 것으로 나타났다. 이것은 기존의 시소러스에 비해 단어연상 시소러스가 질의확장에 이용될 경우 검색시스템의 성능 향상을 가져올 수 있다는 것을 의미한다.

〈그림 2〉는 ASIST 시소러스와 단어연상 시소러스의 성능 평균을 비교한 것이다. 〈그림 2〉에서 보는 바와 같이, 단어연상 시소러스의 경우 연관관계와 하위관계로 질의를 확장했을 때 가장 좋은 성능을 보인 반면, ASIST 시소러스에서는 상위관계와 동등관계로 질의를 확장했을 때 좋은 성능을 나타냈다. 이것은 앞에서 언급한 바와 같이, 일반적으로 연상작용에서 나타나는 단어들이 주로 연관관계와 하위관계의 어휘들이기 때문이며, 반대로 동등관계와 상위관계 어휘들은 비교적 연상이 효과적으로 이루어지지 않기 때문인 것으로 이해된다.

어휘통제의 관점에서 볼 때, 일반적으로 하위관계나 연관관계의 어휘보다는 상위관계나 동등관계 어휘를 통제하는 것이 용이하다. 그렇기 때문에 기존의 시소러스는 연관관계와 하위관계 어휘들의 다양성을 잘 반영하지 못하나 상위관계나 동등관계 어휘를 비교적 정확하게 반영하고 있는 것으로 해석할 수 있다. 이러한 관점에서, 단어연상 시소러스가 기존의 시소러스에 비해 성능이 우수한 것도 결국은 연관관계 단어의 비중이 ASIST 시소러스에 비해 크기 때문인 것으로 해석할 수 있다.



〈그림 2〉 질의확장 실험 결과 : ASIST 시소러스와 단어연상 시소러스 성능 비교

6. 결론 및 제언

본 연구에서는 단어의 의미연상을 이용하여 시소러스를 작성하고, 구축된 시소러스를 이용하여 질의확장 실험을 수행하여 탐색 시소러스로서의 단어연상 시소러스의 가능성을 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. 또한 질의확장 실험에 있어서도 전체 디스크립터에 대해 기존의 시소러스에 비해 단어연상 시소러스가 더 좋은 성능을 보였으며, 특히 단어연상 시소러스는 연관관계나 하위관계 어휘를 확장할 때에 그 효과가 두드러진 것으로 나타났다. 이러한 결과로 볼 때 단어의 의미연상을 이용하여 구축된 탐색 시소러스는 정보검색환경에서 질의확장에 응용될 수 있음을 알 수 있다.

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

이 연구결과를 보다 일반화하기 위해 더 많은 피험자와 다양한 학문분야를 대상으로 단어연상검사를 실시할 필요가 있다. 또한 어휘간의 관계규정에 있어서도 기존의 시소러스에서 사용되는 네 가지 어휘관계 이외에 이용자들이 규정하고 있는 어휘관계에 대해서도 조사해 볼 필요가 있다. 한편, 단어연상에는 문화기반마다 연상에 차이가 있다는 '문화적 연상(cultural association)'이라는 개념이 있다. 이러한 이론을 바탕으로 다국어 시소러스 구축에 단어연상검사를 적용해 볼 필요가 있다.

참 고 문 헌

김태수. 2000. 『분류의 이해』. 서울: 문헌정보처리연구회.

유영준. 2003. 『문헌정보학의 지식 구조에 관한 연구』. 박사학위논문, 연세대학교 대학원, 문헌정보학과.

이찬규. 2002. 단어연상에 관한 조사 연구(I). 『어문연구』, 30(2): 5-33.

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." *Journal of the American Society for Information Science*, 52(5): 402-415.

Hjøland, Birger. 2002. "Domain analysis in information science: eleven approaches - traditional as well as innovative." *Journal of Documentation*, 58(4): 422-462.

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.

Shiri, Ali Asghar and Crawford Revie. 2001. "User-thesaurus interaction in a web-based database: an evaluation of users' term selection behaviour." Proceedings of the Infotech Oulu International Workshop on Information Retrieval [online]. [cited 2006. 5. 28]. 〈http://dlist.sir.arizona.edu/165/01/IR%5F2001.pdf〉

Sinopalnikova, Anna. 2004. "Word association thesaurus as a resource for building

wordnet." *Proceedings of GWC 2004*, 199-205.

Spiteri, Louise F. 2005. "Word association testing and thesaurus construction: a pilot study." *Cataloging & Classification Quarterly*, 40(1): 55-78.

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