인권 시소러스 구축에 관한 연구

A Study on Construction of Human Rights Thesaurus

심민석, 중앙대학교 문헌정보학과, sms@humanrights.go.kr
이두영, 중앙대학교 문헌정보학과, leety0521@hanmail.net

Min-seuk Sim, Dept. of Library & Information Science, Chung-ang University
Too-Young Lee, Dept. of Library & Information Science, Chung-ang University

인권 시소러스는 인권 관련 색인어 작성시 특정성 및 일관성을 유지하고, 다양한 이용자 계층의 정보 검색의 효율성을 증진시키고자 하는 일반적인 목적과 함께, 모호하게 사용되고 있는 인권 용어의 개념화를 통해 전문가 뿐 아니라 일반인들도 손쉽게 인권 전문정보에 접근할 수 있는 토대를 마련하고자 하였다. 이를 위해 본 연구에서는 인권 관련어로 유의미하게 사용되는 용어군을 수집한 후 실험대상군을 설정하여 어느 정도의 관련성을 가지는가를 실험한 것이다.

1. 서론

시소러스는 인터넷과 같은 대용량 디지털 정보가 급증할수록 그 필요성이 증대된다. 효율적이고 정확한 정보 및 지식 전달은 현대 사회에서의 필수적인 요소이며 정보의 커뮤니케이션 기술의 발달로 자동화되고 표준화된 어휘의 중요성이 대두되었다. 이러한 정보 환경의 변화는 기존의 전통적인 검색방법인 분류기호나 단편적인 주제명 표목만으로는 방대한 정보 내에서 다양한 이용자의 정보검색 행태를 만족시킬 수 없다. 더구나 시소러스는 자연언어 처리 등 인공지능 시스템을 실현하기 위한 지식베이스의 근간이 된다는 점에 있어서 그 어느 때보다 중요성이 부각되고 있다.

일반적으로 시소러스를 구축하는 목적은 색인어 작성시 통일성 및 일관성을 유지하고, 다양한 이용자 계층의 정보검색 접근성 및 효율성을 증진시키고자 함이다. 즉, 자연어 검색시스템의 문제점으로 지적되어온 동의어, 동음이의어, 유사어 등을 디스크립터로 통합시키고, 각 주제어간의 상하계층, 연관관계를 정의하므로써, 동일한 개념이 각기 다른 용어로 색인되는 것을 방지하고 용어간의 개념관계를 보여줌으로써 검색에서의 효율성을 도모하여 보다 효과적인 정보검색시스템을 구축하고자 한다.

2. 인권 시소러스 구축 방법

2.1 구축 목적

국내외적으로 인권관련 정보에 대한 관심과 이용이 급증하고 있기 때문에 이를 효과적으로 검색할 수 있는 최적의 도구로 활용하고, 인권관련 자료들의 색인어 선정시 통일성 및 일관성을 유지하여 체계적인 인권자료 데이터베이스 구축을 지원하며, 다양한 이용자 계층의 인권정보에 대한 오프라인 온라인 접근성을 극대화하여 국민의 알권리와 궁극적으로 국민의 인권보호 및 신장을 지원할 수 있는 지식베이스를 구축하고자 한다.

그러나, 모든 용어가 탐색가능한 온라인 시소러스에서는 우선어와 비우선어의 구별이 점점 흐려지고 있으며, 실제적으로 중요하지 않다. 중요한 것은 시소러스에 표현되는 개념이다. 그리고, 인권 관련 용어의 애매모호함은 신속한 정보 전달 및 정보 검색시스템의 장벽이 되었다. 본 시소러스 개발의 주요 목적 중의 하나는 이와 같은 애매모호성을 처리하기 위함이다.

인권 시소러스를 구축하는 구체적인 목적은 다음과 같다.

1) 색인어 작성의 일관성 및 통일성 유지
2) 이용자의 다양한 자연어 질의를 디스크립터 변환으로 검색효율 향상
3) 용어의 상하관계와 그 관련성을 안내하여 검색 용어선정에 기여
4) 구조화된 용어로 확장검색 기능 부여
5) 인권주제어 자동색인 지식베이스 구축
6) 인권용어의 개념화

시소러스는 각 주제 분야의 특성에 따라 구조화가 달라질 수 있으므로, 인권분야 주제어 및 각 어휘들의 특성을 분석하는 것이 필요하다. 인권 주제에 관련된 중요한 개념을 결정하여, 그 개념을 표현하는 가장 적절한 용어를 선택하는 것이 중요하다.

2.2 구축 절차

시소러스 구축에서의 일반적인 주요 고려사항은 용어들간의 계층적이고 결합적인 관계설정, 관련된 색인용어의 조직, 동의어, 유사동의어, 동형이의어 그리고 영문일 경우, 복수형식대 단수형식, 약어, 철자의 변수, 전조합용어 등이었음을 알 수 있다.(Louise F. Spiteri)

시소러스를 구축하는 절차는 다음과 같은 과정을 거쳐 개발되는 것이 일반적이다.

1)주제분야 결정
2)시소러스 특성 및 레이아웃 선택
3)디스크립터의 선택
4)디스크립터의 관계설정
5)표현형식 결정
6)전문가의 검증

위와 같은 구축방법은 주로 수작업에 의해 시소러스를 구축하는데 사용된다. (남영준 등 1997. pp.26)

전통적으로 시소러스에 기재될 디스크립터를 포함한 용어군은 관련 사전류와 학술서적에서 수작업으로 수집한다. 그러나 기계를 사용한 자동 추출은 추출 대상 정보원을 정한 후 명사, 명사구를 추출하여 불용어를 제거하고 빈도수를 측정하여 유의미한 용어를 선택할 수 있다. 빈도수는 어떠한 용어가 대상 정보원에서 어느 정도 출현하는지 수치화한 것으로 디스크립터(비디스크립터) 선정에 도움을 줄 수 있을 것이다.

본 연구에서는 시소러스 개발에 많은 시간이 소요되는 디스크립터를 추출시에 형태소 분석기를 이용하여 추출한 용어를 활용하는 것이 인권 분야 시소러스 구축에 유익한지를 실험하고자 한다. 형태소 분석기는 어절(구), 문장을 입력할 경우, 명사와 명사구를 분리하고 빈도수를 측정할 수 있도록 자체 개발하여 사용하였다.

2.3 실험적 방법

본 연구에서는 인권 관련어로 유의미하게 사용되는 용어군을 수집한 후, 기관에서 소장하고 있는 소장자료 서지(목차 등)데이터베이스에서 추출한 용어들을 대상으로 어느 정도 관련성을 보이는가를 실험하였다.

첫째, 실험을 위하여 유의미한 인권 관련 용어를 수집하였다. 이미 발간된 인권 사전류 및 관련 사이트를 참조하여 용어군을 수집하고자 하였으나, 작업과정에서 표제어를 추출할 만한 사전류를 충분히 확보할 수 없었기 때문에 실험을 위한 별도 명사(구) 파일을 생성하였다.

위의 파일에는 인권관련 법률 및 국가인권

위원회법, 시행령, 시행규칙에 출현하는 용어를 포함하여 법률용어사전(오세경 등 편저), 인권수첩(한상범 등 공저), 사회학사전(고영복 편), 국제인권법(토마스 버겐탈 저), 노동과 페미니즘(조순경 엮음), 등 에서 유의미한 용어를 임의로 선정하였다.

또한, 성공회대 인권평화연구소에서 웹상에 출판한 "인권평화용어사전"과 인권운동사랑방에서 제공하고 있는 "유엔 인권관련 용어"를 추가하였으며, 국립중앙도서관에서 제작한 주제명표목표에서 "인권부분"을 추가하여 1,173개의 명사(구) 파일인 실험대상(A)파일을 만들었다.

파일에서 색인어 선정의 기준으로 명사구는 명사+명사, 명사+의+명사, 명사+적+명사, 명사+성+명사를 사용하였고, 복합어가 분해되어 사용되어도 의미의 변화 없이 하나의 색인어로 사용될 경우 복합어를 분해하였다.

예) 양심과 사상의 자유
-> 양심의 자유
사상의 자유

기본적으로 복합명사는 스페이스 없이 하였으나, 명사구는 의, 적, 성 다음에 스페이스 처리를 하였다.

두번째, 기관에 수집된 인권관련 단행본 자료의 서명, 주제어, 목차정보 및 연속간행물 서명과 기사색인에 나타난 문장, 어절(구), 단어(구)로 파일(B)를 만들었다.

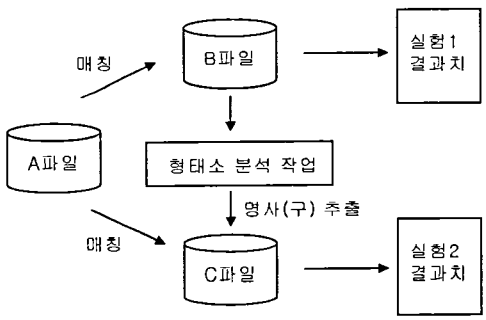
세 번째, 위의 (B)파일을 대상으로 형태소 분석기를 이용하여 명사와 명사구를 추출하여 실험대상(C)파일을 생성하였다.

소장자료를 통해 얻은 각 레코드 건수와 형태소 분석기를 통해 추출된 명사(구) 건수는 다음과 같다.

소장자료 추출 대상		레코드 수 (건)	추출명사(구) (건)
		(B)파일	(C)파일
단행본	서명	9,147	15,346
	목차	1,361	32,767
	주제어	30,517	15,565
연속간행물	서명	255	512
	기사색인	11,101	16,383

추출대상 레코드 수 및 레코드 형태가 추출 명사(구) 수치에 영향을 주었다.

네째, 실험은 그림1을 참고로 다음과 같이 하였다.



<그림 1>

실험1 : (A)파일 1,173건의 용어를 기준으로 하여 (B)파일을 매칭시킨다.

실험결과 (A)파일에 출현한 용어 1,173개 모두 (B)파일에서 출현하였다. 가장 많은 빈도수를 보인 용어는 "여성, 인권, 장애, 자유, 위원회, 장애인, 언론, 환경, 차별, 재판" 순이었다.

실험2 : (A)파일 1,173건의 용어를 기준으로 하여 (C)파일군들을 매칭시킨다. 매칭결과는 다음과 같다.

소장자료 추출 대상		추출명사(구) (건)	(A)파일과 매칭 명사(구)	비율 (%)
단행본	서명	15,346	145	0.9
	목차	32,767	260	0.8
	주제어	15,565	136	0.8
연속간행물	서명	512	17	3.32
	기사색인	16,383	95	0.6

실험결과, 명사구로 뽑은 파일(A)를 가지고 형태소 분석결과 추출된 명사(구) 파일(C)를 매칭시켰을 때 매칭률이 대부분 1%미만으로 관련도가 아주 낮았다.

실험1과 실험2의 결과치는 실험대상 파일의

다음과 같은 차이에서 비롯된다. (B)파일은 소장자료에서 추출한 레코드를 그대로 사용하였고, (C)파일은 형태소분석기를 사용하여 추출한 명사(구) 파일이었다.

인권 분야의 유의미한 단어로 선정한 (A)파일의 용어군을 살펴보면 단일명사보다는 복합명사가 많이 사용되었는데, 이는, 인권분야 관련어의 특성상 단일명사 보다는 복합명사(구)가 인권의 개념을 나타내고 있는 주제특정성에 기인하는 것으로 판단된다.

따라서, 자연어 형태의 레코드로 구성된 (B)파일에서는 (A)파일의 용어군이 모두 출현하였으나, 형태소분석기를 사용하여 추출한 명사(구)는 단일명사 위주로 추출하였기 때문에, 매칭율이 아주 낮았다고 판단된다. 따라서, 형태소분석기를 사용할 경우, 설계시 복합명사(구)를 추출할 수 있는 기능을 강화해야 할 것이다. 여기에 복합명사에 들어있는 스페이스 처리에 대한 논의가 필요하다.

또한, 실험을 위해 생성한 파일(A)의 명사(구)가 색인어로써 가치를 가지고 있는지 관련 전문가의 검증을 받아야 할 것이다.

3. 인권 시소러스 구축의 기대효과

인권 분야는 어느 독립된 학문분야로 특정 용어군이 형성되어 있다기 보다는 전 학문분야에서 포괄적으로 사용되고 있는 용어의 집합체라고 해도 과언이 아니다. 즉, 정치, 경제, 법률, 사회학, 사회복지학, 교육학, 심리학, 여성학 등 사회과학 전 분야와 많은 연관성이 있다.

이러한 특수성으로 인해 인권관련어의 모호한 용어들을 정의 내리고 동형이의어를 구별하기 위하여 한정어를 사용해야 하며 이는 향후 인권 관련어의 개념화 작업과도 연결될 수 있다.

사전 혹은 용어집의 주요 목적이 단어나 용어를 정의하거나 설명하는 것이라면 시소러스의 목적은 이용자가 용어의 의미를 알 수 있도록 도와주는 것이다. 그리고 이러한 시소러스는 역으로 사전작업의 표제어선정에 도움을 줄 수 있을 것이다.

4. 결론

인권 시소러스가 구축되면 인권분야 표목표로도 활용할 수 있으며, 더 나아가 분류체계에 확장하여 적용할 수 있을 것이다.

또한, 용어간의 관계를 나타내줌으로써 검색의 효율성을 높일 수 있고, 인권 분야 용어를 표준화하여 과거에 다양하게 사용되었던 용어들을 통합하여 전문가 집단 뿐 아니라 일반인들도 손쉽게 전문정보에 접근할 수 있을 것이다. 그러나, 무엇보다 중요한 것은 시소러스에 표현되는 개념의 정립이다.

사전이 용어는 알고 있으나, 그 뜻을 모를 경우 사용한다면, 시소러스는 뜻은 알고 있으나, 그 개념에 해당하는 적당한 용어를 찾고자 할 때 사용하는 용어집이다.

본 시소러스가 구축되면 향후 수정 보완을 거쳐 더 발전된 시소러스의 구축이 가능하게 될 것이다.

참고문헌

Louise F. Spiteri. The Use of Facet Analysis in Information Retrieval Thesauri: An Examination of Selected Guidelines for Thesaurus Construction

남영준 등. 시소러스 자동생성에 관한 실험적 연구-법학 분야를 중심으로. 「1997년도 한국정보관리학회 학술대회(제4회) 논문집」. 한국정보관리학회. 1997. pp.26.

이두영 등. 지능형정보검색에 관한 연구 -한국과학기술원보고서-. 한국통신 연구개발원. 1995. 12.

최석두. 한글 시소러스의 구축 기준에 관한 연구. 연세대학교 박사학위논문, 2002.