시소러스 자동생성에 관한 실험적 연구
-법학 분야를 중심으로-
A Study on Automatic Construction of Thesaurus in the field of Law

남영준, 전주대학교 문헌정보학과
최석두, 이화여자대학교 문헌정보학과
이두영, 중앙대학교 문헌정보학과

Nam, Young-joon (Dept. of Library & Information Science, Jeonju University)
Choi, Suk-doo (Dept. of Library & Information Science, Ewha Women University)
Lee, Too-young (Dept. of Library & Information Science, Chung-ang University)

정보의 양이 많아질수록 색인과 검색의 중요성도 증가한다. 특히, 시소러스는 법학 분야와 같이 세부주제간 구분이 가능하고 복합적인 분야의 검색에는 매우 중요한 검색도구로 사용될 수 있다. 한편, 시소러스를 개발하는 가장 합리적인 방법으로는 기존에 개발된 시소러스의 수정 및 보완이라 할 수 있다. 즉, 기존에 구축된 시소러스를 대상으로 문헌정보학전문가와 법학 전문가가 함께 연구하는 것이 가장 바람직한 시소러스 구축방법이 될 것이다. 본 연구에서는 완전한 시소러스를 자동생성하기보다는 언어분석 도구를 이용하여 반제품형태의 시소러스를 제공하므로서, 기존에 개발된 법학분야의 시소러스가 있는 것과 같은 효과를 얻을 수 있는 반제품 시소러스의 개발방법을 제시한다.

서론

각 학문분야에서는 다른 어느 시대보다 활발한 연구활동과 그에 따른 결과물들이 생산되고 있다. 도서관과 같은 정보관리기관에서는 이러한 결과물들을 효율적으로 관리하기 위해, 체계적으로 정리하여 일련의 데이터베이스를 개발하여 이용자의 정보욕구를 충족시키고 있다.

이러한 도서관 환경의 변화는 기존의 전통적인 검색방법인 분류기호나 단편적인 주제명 표목만으로는 방대한 정보군내에서 다양한 이용자의 정보검색형태를 만족시킬 수 없기에 이르렀다.

시소러스는 기존의 정보검색용 도구를 보완하고 이용자의 다양한 정보욕구를 만족시킬 수 있는 중요한 검색용 도구로서 그 효용가치가 여러 분야에서 입증되고 있다.

그러나 시소러스의 구축 및 개발, 유지관리에는 개발초기에 많은 인력과 경비, 시간이 투입되어야 하는 이유 때문에 모든 분야에서 주제별 시소러스가 개발되기보다는 상대적으로 역동적인 분야를 위주로 전문시소러스가 개발되고 있는 추세이다. 또한 각 주제별 특성에

따라 시소러스의 형태는 약간의 차이점을 보이고 있으나, 시간과 경비를 절감할 수 있는 방법은 기존에 개발된 주제전문시소러스를 수정 보완하는 것이다. 본 연구에서는 이점에 착안하여 문헌정보학분야의 전문가와 주제별 전문가가 함께 연구할 수 있는 수준의 반제품상태의 시소러스를 구축할 수 있는 방안을 제시하고자 한다. 특히, 실제 시소러스 개발에 많은 시간이 소요되는 디스크립터의 추출과 선정, 관계설정의 절차에 기계를 이용하여 법학 분야의 시소러스를 구축하고자 한다.

1 시소러스의 구축 방법

1.1 기존의 시소러스 구축 절차

시소러스를 구축하기 위해서는 크게 위원회법과 경험법, 혼합법[1]으로 구분할 수 있다. 이 가운데 가장 일반적인 방법으로는 혼합법으로서 문헌정보학 분야의 전문가와 분야별 전문가와의 협의에 의해 개발하는 방법이다. 또한, 시소러스 개발의 절차는 분야에 따라 약간의 차이는 있으나 대체적으로 다음과 같은 과정을 거쳐 개발되는 것이 일반적이다.

1)주제분야의 결정
2)시소러스특성 및 레이아웃선택
3)디스크립터의 선택
4)디스크립터의 관계설정
5)표현형식 결정
6)전문가의 검증

이 가운데 각 전문가의 역할분담은 대체적으로 다음과 같이 이루어진다. 1)에서 5)까지의 단계는 문헌정보학분야의 전문가에 의해 주로 이루어지며. 6)의 경우에 한해서 전분야별 전문가에 의해 이루어진다.

이 절차는 대체적으로 수작업에 의한 시소러스 구축 절차이다. 이때 기계의 도움을 받을 수 있는 단계는 디스크립터의 선택과정과 디스크립터의 관계설정으로 정할 수 있다. 이 단계를 좀 더 세분하면 대체적으로 다음 4개의 절차가 될 수 있다.

1) 추출정보원의 선정
2) 디스크립터의 추출
3) 디스크립터의 결정
4) 디스크립터간의 관계설정

1.2 디스크립터 추출대상원

전통적으로 시소러스에 기재될 디스크립터(비디스크립터 포함)를 추출하는 정보원으로는 크게 사전류와 학술서적으로 구분할 수 있다. 사전류는 기존에 개발된 시소러스를 포함하여, 외국어로 된 시소러스, 분류표에 나타난 상관색인집류, 해당 및 유사분야의 전문용어사전을 들 수 있다. 학술서적으로는 해당 분야의 교과서나 학술잡지, 학위논문, 전공서적등이 포함되며, 특히, 단행본류의 경우는 권말색인등이 유용한 자료원으로 활용될 수 있다. 실제 카드란(Chandran:1975)은 디스크립터 정보원으로서 백과사전과 같은 사전류를 포함하여 이상의 자료를 포함한 일반교과서외에 학술논문의 초록도 추출대상정보원으로 활용하였다.

1.3 디스크립터(후보)의 추출

디스크립터를 추출하는 것은 크게 수작업으로 이루어지는 방법과 기계의 도움을 받아 이루어지는 것으로 구분될 수 있다.

① 디스크립터의 수동추출 ,

전통적으로 디스크립터는 전부 수작업에 의해 추출되었으며, 앞에서 제시한 추출대상원을 대부분 활용하여 개발하였다. 그러나 최종적으로 디스크립터의 선정은 앞의 정보원에서 출현한 모든 키워드를 시소러스에 모두 기재하는 것은 아니며, 전문가의 결정에 따라 등재여부가 결정되었다.

② 디스크립터의 자동추출

일반적으로 기계의 도움은 후보어를 식별할 때 불용어와 같은 무의미어를 제외하는 것과 용어의 출현빈도를 계산하는 것, 검색에 사용되는 검색어와의 일치(matching)정도를 계산할 때 이루어진다. 이는 용어의 선정에 사용된다.

---

1) 최석두. 시소러스 구축 및 활용지침. 지능형 정보검색 세미나 발표자료. 제주도. 1994. p.8

1.4 법학 분야의 시소러스 구축방법

법률관계시소러스 개발은 앞에서 제시한 기존의 방법과는 커다란 차이는 없으나 시소러스가 완성된 후의 검증보다는 디스크립터의 선정과 관계설정의 단계에서 부터의 참여를 통한 개발방식을 취한다. 왜냐하면 시소러스 개발후 디스크립터의 관계 변경은 자칫하면 개발된 시소러스 체제의 혼란을 야기할 수 있기 때문이다. 이러한 문제점을 사전에 방지하기 위해서는 시소러스 개발 초기부터 법학 분야의 전문가와의 협의를 유도하여야 한다. 전문가의 협의를 실제적으로 유도하기 위해서는 문헌정보학 전문가와 법학 분야의 전문가간의 공동회의보다는 다음과 같이 개발 방법을 세분한다.

1) 법학 전문가로부터 디스크립터 정보원을 추천받는다.

2) 해당 정보원에서 적합한 정보원을 선택하여 디스크립터(후보)를 추출한다.

3) 추출된 디스크립터(후보)를 대상으로 법학 전문가가 디스크립터를 결정한다.

4) 결정된 디스크립터를 대상으로 관계를 설정한다.

5) 설정된 관계를 법학 전문가가 검증 및 수정한다.

이 때 각 단계에는 기능한한 상대 전문가의 의견이 직접 전달되지 않도록 한다. 왜냐하면, 시소러스 자체가 해당 분야의 종사자들을 대상으로 개발된 것이기 때문에 문헌정보학자와의 절충은 바람직하지 않기 때문이다.

2 시소러스 자동생성 실험

시소러스의 자동생성은 1)디스크립터(후보) 추출 2)디스크립터 결정 3)디스크립터 관계설정 절차에서 이루어질 수 있다. 법학 시소러스를 개발하기 위해서는 우선 디스크립터를 추출하기 위한 대상원을 결정한다.

2.1 디스크립터(후보) 추출

기계에 의한 디스크립터(후보)의 추출은 선정된 정보원의 기계가독형태로의 전환에서부터 시작된다. 즉, 정보원의 종류에 따라 선정될 용어의 질과 수준이 결정될 수 있다. 기본 알고리즘은 명사(구)의 추출이다. 왜냐하면, 대부분의 표제어나 디스크립터는 거의 대부분 명사나 명사구로 이루어졌기 때문에, 기존의 자동색인 연구결과에서 색인어추출 대상으로 명사(구)를 추출하는 알고리즘[2]을 도입하였다. 본 실험에서는 대표적인 명사(구)의 형태로 다음 5가지를 대상으로 하였다. 1)명사 2)명사 + 명사 3)명사+의+명사 4)명사+적+명사 5)명사+성+명사 이러한 명사구를 효율적으로 선정하기 위해 불용어 사전을 활용하였다. 디스크립터(후보) 추출과정을 위해 형태소분석기[3]를 이용하였으며, 형태소 분석의 오류는 수동으로 교정하였다. 본 연구에서 형태소 분석의 오류의 원인으로 첫째, 사전 미등록어[4]가 다른 분야보다 많았다. 예를 들면, 가산(嫁産:시가의 재산)과 같은 용어는 주요어로 처리될 수 있으나 일반 국어사전에서도 출현하지 않는 표제어이기 때문에 미등록어 오류로 처리되었다. 둘째, 마르코프 가정의 확률에 근거한 형태소 해석을 하므로서 자주 사용되지 않는 어절에 대해서 품사정보만을 사용하므로서 형태소 해석이 실패할 수 있다. 예를 들면, '빨리 커서 어른이 되고 싶다.'라는 문장이 입력되면 '어른이'의 '-이'는 보격조사임에도 불구하고 동사 앞에서 '-이'가 사용

---

2) 서은경은 "구문·통계적 기법을 이용한 한국어 자동색인에 관한 연구", 정보관리학회지, 제10권 1호, 1993, pp.97-124에서 후보색인어 구문 패턴을 58개로 하였다.

정영미는 "우리말 정보자료를 처리하는 지능형 정보검색시스템의 설계", 정보관리학회지, 제8권 2호, 1991, pp.16-22.에서 명사구패턴을 6개로 하고 관형절 및 관형구로 5개를 제시하였다.

3) 최기선 등. 한글사랑. 제3권 도구모음, 문화체육부 1996. [CD-ROM 판]

4) 형태소 분석기에서 미등록어란 기계사전에 등록되지 않은 단어를 의미한다.

된 품사는 대부분 주격조사로 간주한다. 왜냐하면 확률적으로 후자의 경우가 많기 때문이다. 이 오류는 궁극적으로 용어간의 관계설정에 오류의 원인도 될 수 있다. 셋째, 의미모호성으로서 시소러스 구축에 가장 문제가 되는 부분이다. 예를 들면, 말(馬)과 말(言語)을 들 수 있다. 후자(言語)의 경우 비동작성 보통명사이지만 전자(馬)의 경우는 동작성 보통명사임에도 불구하고 형태소 분석단계에서는 이 차이를 구별할 수 없는 점이다. 이러한 문제점 때문에 형태소 분석의 오류는 수작업으로 처리하였다. 왜냐하면, 구문해석은 형태소 해석결과의 수준에 따라 결정되며, 의미해석은 구문해석수준에 결정되기 때문이다. 즉, 자연어처리는 전적으로 형태소 해석결과에 따라 해석수준이 결정된다.

2.2 디스크립터의 선정

기계에 의한 디스크립터 선정은 추출된 (후보)색인어가운데 의미있는 명사(구)를 선정하는 작업이다. 선정의 근거는 통계적 정보를 최대한 활용하기 때문에, 선정된 (후보)디스크립터 가운데 최종적으로 디스크립터가 될 수 있는 표제어는 문헌내 출현빈도에 결정된다. 이때 고려되는 점으로는 문헌내 출현빈도를 기준으로 할 것인지 혹은 전체 분석대상이 되는 장서내 출현빈도를 기준으로 할 것인지가 결정되어야 한다. 본 연구의 분석대상이 되는 분야는 법학으로서 원론적인 분야보다는 각론적인 분야가 모여 법학이라는 하나의 커다란 분야가 결정되었다는 특성을 가지고 있다. 이러한 특성 때문에 전체 분석대상이 되는 장서군을 분석한 장서빈도를 기준으로 하기보다는 각 분야별 기본 텍스트북을 분석하는 문헌빈도를 기준으로 하였다. 특히, 불용어사전의 표제어는 명사를 최대한 생략하였으며, 기능어를 위주로 재구성하였다. 또한 명사구에 나타난 명사의 처리는 별도의 단어로 간주하였다. 예를 들면, '오스트리아와 스위스, 서독의 헌법소원'이라는 복합명사구가 출현하였을 경우, 이를 '오스트리아의 헌법소원'과 '서독의 헌법소원', '스위스의 헌법소원', '헌법소원'x3, '서독', '오스트리아', '스위스'로 빈도수가 계산되도록 하였다.

2.3 디스크립터간의 관계설정

기계에 의한 디스크립터간의 관계결정은 구문분석과 구문트리(수지도)에 의해 연결을 시도하였다. 문장의 구문구조는 문장을 구성하고 있는 단어들간에 서로 어떠한 관계로 연결되어 있는 가를 명시해 준다. 즉, 입력문장에서 단어와 단어간에 어떠한 관계를 갖고 있는가, 또한 어떠한 단어가 서로 더 밀접하게 관련되어 있는가와 같은 구조적인 표현을 나타냄으로서 문자의 의미파악을 수행하기 위한 전처리 역할을 수행할 수 있다.

특히, 법학 분야의 문헌들은 대부분 문어체로 기술되었기 때문에 격조사와 어미의 정확한 해석과 이를 바탕으로한 수지도는 시소러스의 계층관계를 설정하기 위한 중요한 기준이 되었다. 즉, 구절간 관계표시를 나타내는 것과 구절내 관계를 나타내는 것을 다음과 같은 것으로 기준을 설정하였다.

구절간 관계	격조사	주격조사, 목적격조사, 보격조사, 부사격 조사, 관형격 조사, 공동격 조사, 인용격조사, 접속격 조사, 통용보조사
	어미	대등적 연결어미, 종속적 연결어미, 관형사형 어미
구절내 관계	격조사	서술격조사, 호격조사, 종결보조사
	어미	명사형 어미, 선어말어미, 종결어미
	접사	명사파생접사, 동사파생접사, 형용사 파생접사, 파생접사

1) 구절내 관계 : 구절의 문법적 성분의 변화를 유발하거나 속성을 결정하는 기능어들이다.
2) 구절간 관계 : 이에 속하는 기능어들은 문법적 성분의 변화와 더불어 두 구성성분간의 문법적 관계를 명시하는 역할을 담당한다.

이 기준과 격틀정보를 활용하여 주어와 서술어 및 목적어 등에 출현한 명사간의 관계를 설정한다. 대표적인 격틀로 ['주어'는 '서술어'이

다], ['주어'는 '목적어'를 '목적어'에게]등 으로 90개의 기본적인 격틀을 사용하였다.

3. 시소러스 구축의 실험

이상과 같은 알고리듬을 이용하여 디스크립터의 추출과 선정, 관계설정을 실험하는 과정을 실제로 단계별로 정리하면 다음과 같다.

3.1 정보원의 선정

본 실험에서 선정한 정보원은 법학 분야의 학위논문가운데 일부이다. 색인어는 저자가 직접 제시한 것이다.

입력문 : 헌법소원은 유럽형 헌법재판소제도를 갖고 있는 국가중에서 서독, 오스트리아, 스위스 그리고 스페인만이 실시하고 있다. 이중 헌법소원의 역사에 있어서나 법계로 보나 서독, 오스트리아, 스위스의 헌법소원제도를 비교검토하는 것이 바람직하다고 사료되어 상기 3개국으로 연구의 범위를 제한하였다.

색인어 : 헌법소원, 서독의 헌법소원제도, 오스트리아의 헌법소원제도, 스위스의 헌법소원제도

3.2 후보 디스크립터의 추출

형태소 해석기를 이용하여 의미있는 명사(구)를 추출하면, 다음과 같이 형태소 및 품사가 분석된다.

헌법소원/ncx+은/jcs
~
헌법재판소제도/ncx +를/jco
~
스위스/nq + 의/jcm

이상과 같이 분석된 데이터가운데 앞에서 제시한 5가지의 명사(구)유형에 합치된 단어만을 선택하였다. 이때 복합명사는 명사로 간주한다. 관형사형 어미가 붙은 단어는 명사로 간주한다.

예) 유럽형 헌법 소원제도
(관형사+명사+명사=명사)

3.3 디스크립터의 선정

디스크립터의 선정은 문헌내 출현빈도에 근거한다. 추출된 용어가운데 출현빈도로 정열하면 다음과 같다.

국가
법계
서독(2)
서독의 헌법소원제도
스위스(2)
스위스의 헌법소원제도
스페인
오스트리아(2)
오스트리아의 헌법소원제도
유럽형 헌법재판소제도
헌법소원 (2)
헌법소원제도 (4)

3.4 관계설정

현재진행되고 있는 관계설정은 문단내 출현한 단어들간의 관계만을 설정한다. 또한, 계층관계설정에서 상(하)위어를 구분할 수없었다. 단, 관계가 계층에 해당하는 것만을 인식한다.

1) 1차 분석결과 :
헌법소원 RT 법계 RT 서독의 헌법소원제도 RT 오스트리아의 헌법소원제도 RT 스위스의 헌법소원제도
2) 2차 분석결과 (HT : hierarchical terms)
헌법소원
HT 서독의 헌법소원제도
HT 오스트리아의 헌법소원제도
HT 스위스의 헌법소원제도
RT 법계

결 론

시소러스의 구축작업에서 모든 과정을 전적으로 문헌정보학 분야의 전문가나 혹은 법학

분야의 전문가만을 대상으로 이루어지는 것은 바람직하지 않다. 왜냐하면, 실제 이용자가 법학분야 종사자이기 때문에, 용어의 선정에 있어서 양 분야의 전문가가 합의하여 용어를 선정하기 보다는 법학전문가의 의견이 최대한 반영되는 것이 필요하다. 이러한 상반된 점을 절충하기 위해서는 문헌정보학적으로 시소러스를 반제품형태로 제공하고, 이를 근거로 법학관련 전문가가 최종 검증하는 방법이 필요하다. 법학관련의 지식이 없이 객관적으로 반제품 상태의 시소러스를 개발하기 위해서는 언어학적 지식과 컴퓨터를 활용하였다.

본 연구는 컴퓨터를 이용하여 완전한 시소러스를 자동 생성하기 보다는 디스크립터의 추출과 선정, 관계설정을 통해 반제품상태의 시소러스를 생성하는 연구를 시도하였다.

형태소 해석기와 구분분석기를 이용하여 실제 시소러스를 구축한 결과는 다음과 같았다.

1. 디스크립터의 추출은 5가지의 명사(구)로 후보디스크립터의 형태를 고정하고, 형태소해석기로 이루어질 수 있었다.

2. 디스크립터의 선정은 문헌내 용어의 출현빈도에 근거한 통계정보를 이용하여 디스크립터를 결정할 수 있었다.

3. 용어간 관계설정은 기능어를 활용하여 디스크립터간 관계설정이 가능하였으나, 한 문장내 용어간 관계 설정만이 시도되었다. 또한 상(하)계층구분은 ■여직 ■ 여루어지지 ■ 않았으며, 단 계층관계 정보만은 확인할 수 있었다.

이상의 결과는 법학분야의 전문가의 최종검증이 필요한 수준이며, 정보원입수와 계층관계 설정과정에도 법학분야 전문가가 필요한 수준이다. 향후 관계설정은 문단과 문헌내 용어간 관계설정까지 확대하고, 궁극적으로는 장서군 전체를 대상으로 관계설정이 이루어진다면, 완제품에 가까운 시소러스를 구축할 수 있을 것이다. 이러한 시소러스의 자동생성후 최종점검과정에서만 법학분야전문가의 검증이 필요하게 될 것이다.

<참 고 문 헌>

Chandran, D. "Candidate therms for a thesaurus : A case study of sources of terms in the field of library and information science", Seminar on thesaurus in Information Systems, 1975. pp.A51-A61.

정진성. [단일문서 내에서의 언어 및 통계정보를 이용한 자동색인], KAIST 석사학위논문, 1992.

남영준. 색인어 형태분석에 의한 한국어 자동색인기법연구, 중앙대학교 박사학위논문. 1994.

최석두 등. - 하이텔 메뉴검색용 시소러스의 개발에 관한 연구. [정보관리학회지]. 13(1), 1996. pp.227-241

최석두 등. 시소러스 개발 지침. 서울 : 문헌정보처리연구회. 1994. (문헌정보처리연구회 시리즈 3)

최석두. 시소러스의 표시형식에 관한 연구. 「1994년도 한국정보관리학회 전국논문대회(제1회)논문집」. 1994. pp.105-108.

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