94-g-1

자연어처리를 이용한 시소러스 자동생성

Automatic construction of thesaurus using natural language processing

남 영 준(전주대학교 문헌정보학과)°
이 두 영(중앙대학교 문헌정보학과)

NAM Young-joon (Dept. of Library & Info. Sci., Jeonju Univ.)°
LEE Too-young (Dept. of Library & Info. Sci., Chungang Univ.)

시소러스를 구축하기 위해서는 해당분야의 심도깊은 이해와 지식이 필요하다. 특히, 디스크립터의 선정과 디스크립터의 관계설정은 시소러스 개발자의 주관적인 판단에 따라 이루어지게 된다. 그러나 디스크립터의 선정은 자동색인분야의 연구로서 어느 정도 객관화가 가능하지만, 디스크립터개념간의 관계설정은 개발자의 주관에 전적으로 의존하게 된다. 본 논문은 자연어처리방법과 문헌내 용어출현빈도를 근거로 기계를 이용한 디스크립터간의 관계설정방안을 제시하고 그 가능성을 조사하였다.

1. 서 론

정보검색분야에서는 이용자와 문헌간의 커뮤니케이션을 돕기 위해 여러가지 통제어도구들을 사용하였다. 그 가운데 분류표는 분류표상에 존재하지 않은 주제에 대하여 적절한 분류번호를 부여해야 하는 어려움이 있으며, 학제영역에 능동적으로 대처하지 못한다는 문제점을 가지고 있다. 한편 주제명표목표는 복합주제 형태의 주제명표목의 연거형 리스트로서 표목으로 채택한 수 없는 비디스크립터로부터 표목으로 선정한 수 있는 디스크립터 참조와 하위개념에 대한 참고가 있을 뿐, 상위개념에 대한 참조가 없이 용어개념간의 계층관계는 불분명하다. 이러한 분류표와 주제명표목표는 정보자료의 서가배열을 목적으로도 개발된 것으로 정보자료를 탐색할 수 있는 어휘통제어도구로서는 충분하다. 시소러스는 이와 같은 기존의 어휘통제방법의 단점을 보완하기 위하여 개발된 도구라 할 수 있다.

정보검색용시소러스는 색인작업시에 적절한 색인어의 선택과 색인의 통제를 위해 필요하며, 검색시에는 적절한 탐색어의 선택과 탐색어의 확장이나 축소등 탐색전략을 효과적으로 조절하는 데 필요하다. 즉, 용어간의 계층관계나 연관관계를 이용하여 포괄적인 탐색을 수행함으로서 탐색결과를 보다 적절하고 효과적으로 조절할 수 있는 것이다. 이와 같이 정보검색에 있어서 색인이와 탐색어에 대한 어휘통제는 모든 정보검색시스템의 성능향상과 검색효율의 향상에 탐색전략에 절대적인 영향을 미친다.

그러나 어떠한 용어를 어디서 어떻게 선

정할 것인 가와 선정된 용어간의 관계를 설정하는 것은 지금까지 대부분 수작업으로 이루어져왔다. 특히, 용어간의 관계설정은 검색효율에 절대적인 영향을 미치게 되나, 선정된 용어간의 개념분석과 관계설정을 명확히 정의하기는 매우 어려운 작업이다.

본 연구에서는 이러한 인간의 지적수고를 필요로 하는 작업을 언어학적 방법을 이용하여 시소러스 용어의 생성과 관계설정에 대한 알고리즘을 제시하고자한다.

II. 방법

1. 용어의 수집

일반적으로 디스크립터를 수집하는 방법으로는 주로 기존의 출간된 문헌가운데 다음과 같은 자료를 참고한다.

①시소러스 분류표등의 기존 어휘집

②백과사전,사전,어휘사전등과 같은 사전류

③전문분야의 용어집

④색인지나 초록지, 기타 출판물의 색인

⑤편람, 목록, 교재, 디렉토리, 규격등과 같은 일반자료

이밖에도 주제전문가가 이용자들의 경험과 지식을 활용하여 이들로 하여금 중요한 용어의 리스트를 작성하도록 하여 그들의 지식과 휴리스틱을 열거하여 작성한 주제질문들도 주요한 수집원이 된다.

이 가운데 ① ② ③ ④와 같은 수집원은 이미 디스크립터로 선정된 용어가 이미 선정된 것으로서 용어간의 관계설정만이 문제가 되고 있으나, ⑤와 같이 수집원 자체가 하나의 문장으로 구성이 되어 있어 관계설정이전에 용어의 선정이 필요한 수집원도 있다.

2. 용어의 선정

본 연구에서는 ⑤ 와 같은 수집원에서 디스크립터로서 가치가 있는 용어를 선정하는 방법을 설계하고자 한다.

용어선정에 있어서 기본적인 관점과 방법은 자동색인의 방법과 동일한 과정을 거치게 된다. 왜냐하면 후보색인어의 선정은 일정한 알고리즘으로 분석대상이 되는 수집원 가운데에서 주요어를 추출하는 과정이고, 디스크립터의 선정도 수집원이 되는 자료가운데에서 주요어를 추출하는 과정이기 때문이다.

자동색인의 기법은 크게 통계적 방법과 언어학적 방법으로 구분할 수 있다. 통계적 방법은 용어의 출현빈도로서 주요어를 선정하는 것이며, 언어학적 방법은 용어의 문장내 역할로서 주요어를 선정하는 것이다. 본 연구에서는 디스크립터로 사용될 수 있는 용어의 선정을 위해 언어학적 접근방법을 활용한다.

일반적으로 언어학적 분석방법은 다음 세가지 분석방법으로 분류할 수 있다.

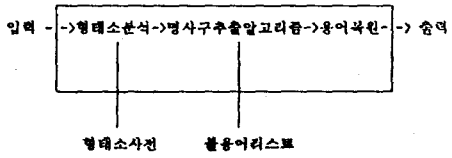
① **形態素分析** (Mophological analysis) : 일련의 문장을 어절단위로 분석하여, 각 어절을 구성하는 형태소 간의 결합 및 수식관계등을 파악하여 구조화하는 방법이다.

② **構文分析** (Syntax analysis) : 문장을 분석단위로 하여, 형태소분석의 결과를 받아 문장을 구성하는 어절간의 수식관계등 문법적 관계를 파악하고 보통은 트리형태의 결과를 추출한다.

③ **意味分析** (Sementic analysis) : 기준에 따라 다양한 형태를 보일 수 있으나, 일반적으로는 구문분석의 결과인 파싱트리(parsing tree;어절분석결과)를 입력받아, 그러한 문법관계를 갖는 문장의 구체적 의미를 파악하는 방법이다. 이 방법은 분석대상문헌과 검색문의 의미적 분석방법에 치중하는 데 이용된다.

본 연구에서는 위의 세방법가운데 특정한 방법만을 사용하지 않고 각 분석방법의 특성에 맞게 이를 사용한다. 왜냐하면 위의 세가지 해석수준은 서로 독립된 이론이 아니며, 하위 수준(형태소분석 < 구문분석 < 의미분석)의 해석결과가 바로 위의 단계로 전달되어 결국에 문장 전체의 의미를 파악하는데 필요한 정보를 제공해주기 때문이다.

우리말색인어로서 가치가 있는 것으로는 명사 혹은 명사구가 대부분이다. 그러므로 디스크립터로 사용될 수 있는 것도 명사나 명사구로 제한한다. 이러한 명사나 명사구를 추출하기 위해서는 형태소분석이 필요하고, 명사구 복원을 위해서는 간단한 구문정보가 필요하다. 즉, 후보디스크립터가 될 수 있는 용어선정을 위해서는 분석대상이 되는 수집원을 형태소분석과 구문분석이 이루어져야 한다. 본 연구의 용어선정방법은 <그림 1>과 같다.



<그림 1> 용어선정시스템

3. 관계설정

검색에 있어서 시소러스가 주제명표목과 기존의 다른 통제어휘집에 비해 훨씬 검색 효율이 높은 도구로 활용되고 있는 이유는 용어(검색용어)간의 관계표시가 설정되어 있기 때문이다. 그러나 용어간의 기본적인 관계를 설정하는 것은 매우 정교한 작업이기 때문에 해당분야의 전문적인 지식과 언어학적인 지식이 필요하 다.

용어간의 관계표시는 여러가지가 있을 수 있으나 일반적으로 1)대등관계 2)상하위관계(혹은 계층관계) 3)연관관계로 나타낸다. 이러한 관계표시구조를 갖고 있는 대표적인 시소러스는 TEST가 있으며, 그 표시는 다음과 같이 나타내고 있다.

① 대등관계(the equivalence relationship)는 USE와 UF로 나타낸다.

② 계층관계(the hierarchical relationship)는 BT와 NT로 나타낸다.

③ 연관관계(the associative relationship)는 RT로 나타낸다.

본 연구에서 사용한 관계설정방법은 의미론적 방법과 문장구조분석방법을 사용하였다.

3.1 대등관계

대등관계는 여러개의 용어가 하나의 동일한 개념을 표현할 때 표시하는 관계이다. 대등관계는 동의어와 유사동의어일 경우에 표시가 되며, 용어간의 대등관계를 자동생성하는 것은 현실적으로 어렵다. 왜냐하면, 수집원내에서 상황에 따라 우선어와 비우선어가 결정되야 하기때문이다. 예를 들면, 펭귄(penguins)과 펭귄류(sphenisciforms)는 주제분야에 따라 속명이 사용될 수도 있고, 학명이 사용될 수도 있다. 그러므로, 본 연구에서는 대등관계의 자동생성은 시도하지 않았다.

3.2 계층관계

계층관계는 개념간의 상위와 하위를 결정하는 관계이다. 상위개념은 主流 또는 全體를 나타내는 개념이며, 하위개념은 그 한 요소 또는 일부분을 나타내는 개념이다. 일반적으로 계층관계는 속관계와 계층적인 전체-부분관계, 사례관계로 구분할 수 있다.

본 연구에서는 수집원내에 출현한 용어간의 계층관계를 설정하기 위해 각 용어가 문장내에서 갖게 되는 격조사와 용언어미정보를 활용하였다.

① 주격조사와 서술격조사를 갖는 명사(구)는 계층관계를 갖는다

주어(주격조사) + 보어(서술격조사) + 동사(있다, 이다.)

정류기에는 전자관 정류기, 수은 정류
주격조사
기, 반도체 정류기 등이 있으나...
서술격조사

주어와 보어는 계층관계를 갖는다.

3.3 연관관계

연관관계는 대등관계나 혹은 계층관계보다 관계설정이 매우 어렵다. 왜냐하면 연관관계는 용어간에 유사성이나 계층성이 없으면서 상호보완성을 가지고 있는 용어의 개념을 파악해야 하기 때문이다.

본 연구에서는 다음과 같은 경우는 연관관계를 설정하였다.

① 주격조사를 갖는 명사(구)와 목적격조사를 갖는 명사(구)

주격조사:은,는,이,가,도.. 목적격조사: 을,를,에게...

② 주격 혹은 목적격조사를 갖는 명사(구)와 도구격조사를 갖는 명사(구)

도구격조사: -에 의한, 활용한, 를 이용한, 을 통한

③ 주격, 목적격조사를 갖는 명사(구)와 이유격조사를 갖는 명사(구)

이유격조사: -로/ -으로 인한

④ 접속격조사를 공유하는 명사(구)

접속격조사: 와,과,이(나), 및, 혹(은)

예) 입력문->전기철도는 증기기관차나 디젤기관차에 비해 견인력과 속도를 크게 높일 수 있어서...

전기철도	증기기관차	견인력
RT 증기기관차	RT 전기철도	RT 전기철도
디젤기관차	디젤기관차	속도
견인력		
속도		

4. 디스크립터의 관계설정 및 복원

이상과 같은 방법으로 결정된 디스크립터 연결그래프를 전체적으로 연결시키면 하나의 전체적인 그래프가 형성된다. 이때 어느 용어와도 연결되지 못한 디스크립터와 일정 수 이하로 연결된 디스크립터는 자격이 없는 용어로 간주한다. 특히, 한 문장내에서 주요어가 나타났더라도 문장내의 다른 용어들이 불용어만이 나타났다면 이러한 용어(주요어)는 디스크립터로 자격이 없는 용어로 간주하였다. 한편, 그래프가 한용어에 대해 집중적으로 연결된 용어는 주요어이며 최상위어로 간주하였다. 이상과 같은 방법으로 설정한 구조를 수작업으로 설정한 구조와 비교한 형태는 표 1과 같다.

〈표 1〉 시소러스구축결과비교

수작업결과	자동생성결과	
① 전기조명	②전기조명	④조명기구
NT 조명기구	RT 전기응용기술	RT 조명방식
RT 조명방식		NT 백열등
NT 백열등		방전등
방전등	③조명방식	
형광등	RT 조명기구	
NT 조명방식	NT 직접조명	
RT 조명기구	간접조명	
	전반조명	
	국부조명	

III. 소 결

본 연구에서는 단일문서를 대상으로 주요어를 선정하고, 용어의 격조사정보를 이용하여 용어간의 관계를 설정하고, 최종적인 디스크립터선정은 통계적 알고리즘과 인결그래프의 중복도를 활용하였다.

실험은 고등학교 공업교과서을 대상으로 실시하였다. (3장 전기공업)

향후 연구과제로서 자연어문장을 더욱 정확하게 분석할 수 있는 형태소분석기와 보다 정확한 심층격정보를 얻을 수 있는 기틀사전이 완성된다면, 시소러스의 자동생성 가능성은 훨씬 커질 것이다.

참고문헌

Rajan, T. N. Related terms in thesauri. Seminar of thesaurus in information systems. DRTC & INSDOC. 1975, pp.A13-A21.

국방과학연구소. 국방과학기술 한글시소리스. 동연구소. 1994.

금장철. 한글맞춤법 검사기, 어떻게 구현할 것인가. 마이크로소프트웨어, 1994. 2.

김명철 외. 시소러스 작성을 위한 개념 획득 도구. 인간과 기계와 언어. 제4회 한글 및 한국어정보처리 학술발표논문집. 1992, pp.39-49.

이봉구 외. 고등학교 공업. 금성교과서(주). 1993.

이주근. 인공지능. 청문각. 1988.

정영미. 정보검색론. 구미무역(주). 1993.

한국문화예술진흥원. 시소러스개발을 위한 프레임웍. 동진흥원. 1994.