로그데이터를 이용한 디스크립터의 외형적 특성 분석

Analysis of the Candidate Terms and Structure Using the Log-data

남영준, 중앙대학교 문헌정보학과, namyj@cau.ac.kr
이두영, 중앙대학교 문헌정보학과, leety0521@hanmail.net

Young-joon Nam, Dept. of Library and Information Science, Chung-ang University
Too-young Lee, Dept. of Library and Information Science, Chung-ang University

본 연구에서는 시소러스를 구축하기 위해 필요한 디스크립터 수집원으로써 이용자 로그데이터를 분석하여 후보 디스크립터의 외형적 특성을 분석하였다. 분석대상인 이용자 로그데이터는 국내 웹 검색엔진가운데 야후와 라이코스를 대상으로 하였다. 분석결과, 이용자들은 대부분 검색어로써 명사와 복합명사를 사용하였으며, 조사 '의'이외에는 다른 품사로 이루어진 검색어는 거의 존재하지 않음을 알 수 있었다. 또한 검색어로써 이용자들은 고유명사(외국어 포함)를 많이 사용함으로써, 국내외 지침에서 권고하는 고유명사의 최소한 사용지침과 실제 이용자 사이의 이용행태와의 차이를 알 수 있었다. 따라서 국내외 시소러스 개발지침을 수용하면서, 이용자 중심의 시소러스를 개발하기 위해서는 전거어나 유사어 사전을 대등관계와 연동하여 개발하는 것을 고려해야 한다.

1. 서론

도서관 현장과 문헌정보학 분야에서 최상의 목표로 하는 것가운데 이용자 중심의 이론과 방법을 개발하는 것이다. 이용자의 관점은 개발된 도구를 사용하는 관리자 입장의 사서와 실제 개발된 각종 도구를 검색에 사용하는 이용자로 구분될 수 있다. 어떤 형태의 이용자라도 해당 문헌정보학 관련 도구는 인간이 기억하기 쉽고, 사용하기 용이한 것이어야 한다. 예를 들면, 주제전개와 개념조화에 있어 분석열거식 분류체계가 분석·합성식 분류체계에 비해 상대적으로 불리하다. 이러한 불리한 조건에도 열거식 분류체계의 사용용이성과 조기성으로 오히려 많은 기관에서 활용되고 있는 것도 사용자인 인간이 수월하게 기억하고 사용할 수 있기 때문이다.

시소러스의 역할은 크게 검색을 위한 것과 색인을 위한 것으로 구분할 수 있다. 특히 시소러스는 주제명 표목표와 달리 개념간의 구조체계 원칙을 적용하여 각각의 개념을 계층관계와 함께 연관관계로 다차원적인 구조를 갖고 있다. 이는 시소러스의 주된 쓰임새가 이용자를 위한 검색지원에 비중을 두고 있음을 유추할 수 있도록 한다. 따라서 주제명 표목표는 도서관 자료 관리를 위한 색인도구이며, 시소러스는 검색 중심에 도서관 이용자의 검색도구적 성격이 강하다. ANSI/NISO에서는 디스크립터의 유지보수과정에서 새로운 용어로

대치(replacement)를 하는 과정에서 반드시 검색만을 고려해야 한다고 선언함으로써 시소러스의 역할을 보다 분명히 하고 있다(ANSI/NISO 2003).

이용자 중심의 검색과정에 많은 의미를 갖기 때문에 시소러스는 디스크립터를 사용하는 이용자들에게 친숙한 구조와 형태로 개발되어야 한다. 이 가운데 디스크립터는 시소러스를 사용하는 이용자들이 정확하게 인지하고 있는 용어로 구성되어야 한다.

한편 시소러스 개발과 관련된 국내외의 가이드라인은 시소러스에 등재될 디스크립터의 형태와 품사, 관계 설정 등과 같은 인쇄형 혹은 온라인형 시소러스 구축작업에 대한 기준이 설정되어 있다.

본 연구는 이 점에 착안하여 실제 후보 디스크립터로 추출하기 위한 기준이나 알고리듬으로 활용할 수 있는 후보 디스크립터의 외형적 특성과 관련된 기준을 제시하고자 한다. 즉, 실제 이용자들이 사용하는 검색어의 품사를 비롯하여 표준어와 외래어, 외국어의 사용패턴을 분석하고자 한다. 또한 기준을 설정하기 위한 분석자료는 야후(코리아)와 라이코스(코리아)를 통해 이용자들이 검색창에 입력한 실제 키워드를 분석하여 활용한다. 분석대상이 되는 키워드는 2001년도 상반기 6개월간에 각 홈페이지에서 제공한 주별 인기검색어 코너에서 수집하였으며, 수집대상은 입력순위에서 상위 30개의 검색어로 제한하였다.

2. 디스크립터 추출 정보원

디스크립터로 채택될 수 있는 용어는 특정 정보원에서 검증된 것으로 제한한다. 따라서 국내외 시소러스 개발 기준에서는 이에 대해 다음과 같은 기준을 갖고 있다.

2.1 국내 지침

시소러스의 개발은 크게 첫 번째로 시소러스의 특성 결정을 비롯하여, 두 번째, 용어수집, 세 번째 용어간 구조화, 네 번째 출판 및 피드백 등 4단계로 구분된다(쓰리소프트 2002). 이 가운데 용어 수집과정에서 활용한 용어원은 기존 관련 용어집과 함께 해당 분야의 주요 문서 및 문헌을 활용하였다(한국정보관리학회 1988).

한편, 대부분의 시소러스 개발은 주요 자료집과 문헌에 출현한 용어를 대상으로 문헌조사를 통해서 수집하는 것과 함께 이용자의 질문을 분석함으로써 실제로 탐색시에 사용되는 용어를 수집하는 것도 권장하고 있다(정영미 1993). 즉, 이용자 로그데이터을 분석하여 실제 이용자들이 사용하고 있는 용어를 기준으로 후보 디스크립터를 수집할 수 있다.

한국데이터베이스진흥센터의 가이드라인에서는 용어의 수집원에 대해 시소러스의 기획과 개발과정에서 용어의 검증과정을 통한 디스크립터의 채택 및 수정과정에서 다음과 같은 정보원을 사용하도록 권고하고 있다(한국데이터베이스진흥센터 2000)

- 표준기술사전과 백과사전
- 기존의 시소러스
- 분류표

2.2 국제 지침

2.2.1 ANSI/NISO-2003

ANSI/NISO에서는 후보 디스크립터의 선정 방법으로 다음 세 가지를 제안하고 있다.

- 컴퓨터를 이용하여 서명과 초록과 같은 정보원에서 후보 디스크립터를 추출하는 방법이다. 이때 불용어를 제외하고 기존의 시소러스에 등재되어 있는 단어가 포함된 명사구나 복합어를 선정한다.

- 특정자료를 색인하는 과정에서 자주 색인어로 사용되는 용어로써 시소러스에 등재되어 있지 않는 용어를 선정한다.

- 이용자 질의어와 같은 로그 파일에 자주 사용되는 검색어로써 시소러스에 등재되어 있지 않는 용어를 선정한다.

ANSI/NISO에서는 디스크립터로 사용될 수 있는 용어 형태는 명사를 포함한 명사구 등으

로 제안하고 있으며 등재된 용어는 표준어를 기입어로 사용하고 있었다. 방언이나 외국어의 경우에 대해서는 원 용어 형태를 그대로 사용한다.

2.2.2 Aslib

Aslip에서는 시소러스 구축을 위한 교육용 자료를 구성하면서 시소러스 개발에 필요한 절차를 12개의 과정으로 제안하고 있다. 이 가운데 용어수집단계에서 용어원으로 다음과 같은 자료를 기본 자료로 활용하고 있다(Aitchison, Gilchrist, Bawden 2000).

- 시소러스 및 용어 목록
- 분류표
- 백과사전, 어휘집, 사전, 용어집
- 전문용어 데이터뱅크
- 주제분야 용어관련 전문서적

이 지침에서는 디스크립터 개발에 필요한 용어원으로써 표제항이나 색인어로 통제된 용어(명사나 명사구 등)를 활용하도록 권고하고 있다.

이상과 같이 국내외적으로 시소러스를 구축하기 위한 후보 디스크립터 수집원에 대해서는 문헌자료와 함께 이용자 로그데이터을 사용하도록 권고하고 있다. 한편 ISO와 같은 국제기준에 용어 수집원에 대한 구체적인 사례를 제시하지 않고 있다(ISO 1986, ISO 1985, UNESCO 1981).

3. 디스크립터의 형태 및 품사정보

디스크립터의 형태와 품사는 국내와 외국의 기준 간에 약간의 차이가 있다. 이는 문법과 언어문화 배경의 차이에서 발생한다.

3.1 국내 지침

시소러스에 등재될 디스크립터는 대표적인 지침과 개발과정에서 명사와 명사구를 우선하고 있다. 특정 시소러스에서 형용사나 관형사가 수식하는 형태로 이루어진 명사구나 명사절을 일부 발견할 수 있다. 그러나 그 수는 전체 시소러스나 색인어집에서 극히 일부를 차지하며, 전체적인 것은 대부분 다음 형태의 명사와 명사구 형태를 유지한다(쓰리소프트 2002, 남영준 2003).

- 명사 (복합명사 포함)
- 명사 + 의(조사) + 명사
- 명사 + 적(관형사형 어미) + 명사
- 명사 + 성(관형사형 어미) + 명사

한편, 한국데이터베이스진흥센터의 개발지침에서는 디스크립터를 구성할 수 있는 품사정보로써 명사를 제안하고 있으나, 이미지나 동영상 색인에 있어 형용사형 용어를 사용할 수 있으며 이에 대한 고려도 권장하고 있다. 또한 실제 한글로 형용사와 동사로도 일련의 시소러스를 구축할 수 있다고 주장하여 일련의 예를 제시하고 있다(한국데이터베이스진흥센터 2000). 동사로 이루어진 계층에 있어 특정분야의 용어가 갖는 특수성 때문에 극히 일부에 한해서 구축될 수 있다. 또한 형용사위주로 시소러스를 구축할 수 있는 분야는 이미지나 감성분야와 같이 제한적일 밖에 없다. 디스크립터로 형용사를 사용할 경우에 이를 모두 서술형 어미인 '~다'를 부기하고 사용하여 다의적인 표현을 최소화하였다. 예를 들면, '우아한'이란 형용사는 디스크립터로는 '우아하다'로 변형하여 구조화하는 것이다. 감성관련 시소러스 구축실험에서 색상이 갖는 감성적 의미를 표현하기 위한 수단으로 이와 관련한 형용사를 사용한 시소러스 구축에서도 형용사의 활용형을 최소화하기 위해 이와 같은 방법을 사용하였다(남영준 2003). 이 실험에서 시소러스의 특징을 갖기 위해 계층과 연관관계와 같은 구조화는 색상표와 해당 색상이 갖는 의미를 연결하여 2차원적인 색상표의 구조에 용어간 거리를 대입하여 3차원 관계의 시소러스를 구축하였다.

3.2 외국의 지침

외국에서 디스크립터로 채택될 수 있는 품

사와 형태는 국내에 비해 다양성을 많이 갖고 있다. 예를 들면, 국어의 경우 복수형과 단수형의 구분이 비교적 용이하나, 영어의 경우 단·복수형에 대한 지침이 필요하다. 또한 국내 지침에서는 디스크립터의 형태로 명사와 제한된 명사구로 권고하나, 외국의 지침에서는 명사와 형용사, 부사 및 전치사 등으로 이루어진 명사구와 명사절이 디스크립터의 형태로 권고하고 있다. 기준에 제시하고 있는 주요한 내용가운데 품사정보와 형태에 관한 내용을 정리하면 다음과 같다(ISO 1986, ISO 1985, UNESCO 1981, ANSI/NISO 2003).

- 명사형: 외래어, 약어, 속어와 방언, 고유명사, 복합명사
- 명사구 : 형용사와 전치사 등이 포함된 명사구 혹은 명사절

4. 주요 웹검색엔진의 검색어 로그데이터 분석

디스크립터를 수집하는 주요 정보원가운데 이용자의 시스템에 대한 로그 파일에 수록된 검색어와 기존 색인어집에서 사용된 빈도가 높은 용어를 신규 디스크립터로 채택할 수 있다(ANSI/NISO, 정영미 1993).

4.1 실험 환경

이용자 중심의 후보 디스크립터를 선정하는 최선의 방법은 하나 이상의 시소러스를 운영하고 해당 시소러스(분류체계)에 등재된 디스크립터의 활용빈도(상관색인)를 이용하고, 신규 용어에 대해서는 이용자의 로그데이터을 분석하는 것이다.

이러한 이용자 로그 데이터 수집 방법과 가장 유사한 구조가운데 하나가 웹검색엔진에서 제공하는 디렉토리를 기준으로 이용자 로그데이터을 수집하는 것이다. 본 연구에서는 이를 위해 상용 웹검색엔진가운데 카테고리 서비스를 제공하고, 주별 단위로 인기검색어를 공개하는 야후(코리아)와 라이코스(코리아)를 분석대상으로 하였다. 야후(코리아)는 해당 웹사이트의 검색창으로 검색이 이루어진 검색어가운데 상위어 30개를 매 주단위로 공개하고 있다. 또한 라이코스(코리아)도 해당 웹사이트의 검색창으로 검색이 이루어진 검색어가운데 상위어 50개를 매 주단위로 공개하고 있다. 이들 웹검색엔진을 조사한 기간은 2001년 1월부터 6월까지 6개월간이며, 공개한 형태를 조정과정 없이 그대로 사용하였다.

4.2 실험 관점

본 실험에서는 이용자 로그데이터을 이용하여 실제 검색어에 대한 다음과 같은 관점을 중점적으로 분석하였다.

- 명사 및 명사구의 의존정도, 명사형태
- 고유명사의 의존정도
- 외국어 및 외래어의 의존정도

이러한 관점을 기준으로 유사어 사전의 범위를 범주화한다.

4.2.1 명사 및 명사구의 의존정도

분석자료는 비교를 위하여 라이코스(코리아)와 야후(코리아)의 상위 검색어 30개를 대상으로 하였다. 따라서 분석 대상이 되는 검색어의 개수는 30주를 분석하였기 때문에 각 900개이다.

① 라이코스 : 라이코스의 이용자 검색어 로그데이터가운데 형용사와 같은 수식어가 없이 '명사나 명사구의 의존정도'는 900개 가운데 37개였으며 '의'라는 조사를 포함한 것을 명사구로 간주할 경우에는 7개의 용어열이 형용사가 포함된 명사구였다.

② 야후 : 야후의 이용자 검색어 로그데이터가운데 형용사와 같은 수식어가 없이 '명사나 명사구의 의존정도'는 900개 가운데 29개였으며 '의'라는 조사를 포함한 것을 명사구로 간주할 경우에는 야후에 조사된 이용자로그파일은 모두 명사와 명사구로 이루어져 있었다.

즉, 검색어는 거의 대부분 명사와 명사구로 이루어져 있었으며, 형용사가 포함된 용어열도

특정 동영상물을 표현한 고유명사였기 때문에 실제로 이용자들이 사용한 최대한의 비명사는 조사 '의' 정도였다. 따라서 이용자 로그 파일로 디스크립터를 선정할 경우나 혹은 특정 자료에서 디스크립터를 선정할 경우에 명사(복합명사)와 조사 '의'가 포함된 명사구로 한정지을 수 있다.

4.2.2 고유명사의 의존정도

디스크립터에서 고유명사의 허용은 국내외 모든 기준에서 이루어지고 있으나, 실제 국내 시소러스 구축에서는 디스크립터로 채택되는 고유명사를 최소화하고 있다(한국정보관리학회 1998, 남영준 2002). 또한 주제명 표목표(국립중앙도서관)의 디스크립터에서 고유명사가 최소부분만을 차지하고 있다. ANSI/NISO에서도 고유명사의 경우 가능한 통제를 유도하여 고유명사로 사용하지 않고 주제에 배열하거나 별도의 전거파일을 사용하도록 권고하고 있다(ANSI/NISO 2003). 이에 비해 본 조사에서 분석된 고유명사의 의존정도는 다음과 같았다.

① 라이코스 : 라이코스 로그데이터 900개 가운데 473개의 용어가 고유명사로 조사되었다. 고유명사의 유형은 사람이름과 게임명이 대부분이었으며, 계절적 특성에 따라 정부기관명이 사용되었다.

② 야후 : 야후의 로그데이터은 900개 가운데 812개의 용어가 모두 고유명사이다. 일반명사는 게임이나 채팅과 같은 컴퓨터관련 용어로써 특정성을 갖는 용어보다 카테고리를 구성할 수 있는 범주속성을 갖는 용어로 이루어져 있다. 즉, 거의 대부분의 이용자 검색어는 고유명사로 이루어져 있다. 대부분의 고유명사는 사람이름과 게임명, 기관명(특히 언론매체)이 대부분이었다.

4.2.3 외국어 및 외래어의 의존정도

디스크립터에서 외국어는 특정한 대응어가 없을 경우를 제외하고는 사용하지 않는다. 또한 사용할 경우에도 반드시 대응외국어를 병기함으로써 한글음화에 따른 정보의 노이즈를 최소화한다. 예를 들면, 케이비에스란 용어와 케비에스란 용어는 영어대응어는 같으나 한글음화가 다르게 표현될 수 있기 때문에 kbs를 사용함으로써 한글음화에 따른 노이즈를 최소화할 수 있다.

① 라이코스 : 라이코스 로그데이터 900개 가운데 76개의 용어가 영어로 이루어진 검색어이다. 모든 외국어로 표현된 검색어는 특정 가수의 이름인 고유명사였다. 이외에도 한글이름으로 표현된 외국어(리불트 등)는 비표준어임에도 불구하고 한글로 처리된 것은 이용자들이 이미 인지하고 있는 이름(가수명이나 그룹명 등)은 문자열보다 이미지로 기억하고 있기 때문이다. 즉, 한글문자로 기억하는 것이 아니라 해당 글자를 기호로 인지하여 검색에 사용하는 것이다.

② 야후 : 야후의 로그데이터는 900개 가운데 141개 용어가 영어로 이루어진 검색어였다. 야후의 영어 로그데이터는 라이코스와 같이 사람이름과 기관명(kbs 등)이 대부분이었다.

이와 같은 분석 결과를 학술적 시소러스 구축에 그대로 적용하는 것은 무리가 있으나, 온라인 검색 환경에서 검색어의 형태적 정보는 학술적 검색과 일반적 검색에 차이가 없을 것이다. 따라서 이상의 분석 결과를 다음과 같이 요약할 수 있다.

- 온라인 검색 하에 검색어는 명사(복합명사 포함)로 대부분 이루어지고 있다. 형용사가 포함된 것은 전체조사 데이터의 0.1%이하로써 무시되어도 좋은 수준이었다.
- 명사로 이루어진 검색어의 많은 부분이 고유명사였으며, 해당 고유명사는 표준어의 준용여부를 고려하지 않는 특성을 갖고 있었다.
- 외국어로 이루어진 검색어는 모두 영어였으며 고유명사로써 특정 기관명이나 성명이었다.

이러한 분석결과에 근거하면, 온라인 검색에서 이용자 중심의 시소러스를 개발할 경우 실제 이용자들이 사용하는 용어는 비표준어와 약어 등이 사용되기 때문에 이를 수용할 수

있는 구조를 가져야 한다. 즉, 시소러스의 이용성을 극대화하기 위해서는 디스크립터의 대등관계에 배정된 유사어와 전거어 중심으로 시소러스가 개발되어야 한다.

5. 결론

온라인 환경에서 이용자 중심의 정보검색 시스템을 유지하기 위해서는 새로운 관점의 검색도구를 개발할 필요성이 있다. 이 가운데 이용자 중심의 시소러스는 전통적인 분류표의 상관색인집이나 전문용어집에 등재된 표제항과 함께 실제 이용자들이 사용하는 검색어가 반영되어야 한다. 이를 위해서는 국내외 시소러스 개발 기준에서 권장하는 디스크립터 정보원인 해당 주제분야의 문헌자료이외에 이용자 로그데이터를 분석할 필요성이 있다. 이용자 로그데이터의 분석은 새로운 주제영역을 나타내는 주요어를 도출하기 보다 기존에 존재하는 디스크립터를 대신하여 현재 온라인상에서 사용되고 있는 유사어 혹은 전거어형태의 후보어를 얻을 수 있다. 이에 따라 본 연구에서는 야후(코리아)와 라이코스(코리아)에서 검색자들이 실제 검색창을 통해 입력한 자연어 형태의 검색어 로그데이터를 분석하였다. 분석의 결과를 요약하면 다음과 같다.

- 대부분의 검색어는 명사(복합명사)로 이루어져 있었으며, 명사구로 이루어진 것은 조사'의'가 첨부된 것으로써 조사 '의'가 첨부된 명사구를 명사의 범주로 간주하면, 전체 99.9%이상의 검색어는 명사였다.
- 외국어(영어)를 검색어로 사용한 것은 모두 인명과 관련한 것이었으며, 모두 연예인의 이름이었다. 따라서 외국어로 검색한 것은 영어단어의 용어열보다는 기호적인 인식을 통한 영문기호로 사용되고 있었다.
- 비표준어로 분류될 수 있는 속어나 방언 등은 특정 게임이름과 같은 고유명사였다.

이와 같은 분석결과에 따라 온라인 검색환경에서 시소러스는 이용자 중심적 관점의 역할을 효율적으로 수행하기 위해서는 유사어나 전거어, 이형발음, 대응외국어와 같은 대등관계어가 정교하게 개발되어야 한다. 따라서 본 연구는 후보 색인어나 디스크립터를 추출하는 기초자료로 사용되어 색인어 추출기 혹은 후보 디스크립터 수집기를 개발하는 주요 알고리듬으로 활용될 수 있을 것이다.

참고문헌

남영준. 2003. 시소러스의 대등관계에 관한 연구. 「문헌정보학보」 중앙대학교 문헌정보학회. 6 : 143-175.

쓰리소프트. 2002, 「고속철도건설공단 시소러스 개발 및 구축」. [서울] : 쓰리소프트.

정영미. 1993. 정보검색론. 구미무역 출판부.

한국데이터베이스진흥센터. 2000. 「시소러스 개발지침.」 이화여자대학교. 최종연구보고서. 서울: 97~200

한국정보관리학회. 1988. 「법률분야 관련어집」. 최종보고서. 서울 : 법원도서관.

ISO. 1985. *Documentation - Guidelines for the Establishment and Development of multilingual Thesauri.* 2nd edition (ISO 5964-1985(E))

ISO. 1986. *Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri.* 2nd edition (ISO 2788-1986(E))

Jean Aitchison, Alan Gilchrist, David Bawden. 2000. *Thesaurus Construction and Use: a Practical Manual.* 4th. ed. Aslib.

ANSI/NISO. 2003. 「Guidelines for the Construction, Format, and Management of Monolingual Thesauri : ANSI/NISO Z39.19 - 2003」 the American National Standards Institute.

UNESCO. 1981. 「Guidelines for the Establishment and Development of Monolingual Thesauri-2nd revised edition」 UNESCO.