도서권말색인의 작성지침과 자동생성에 관한. 연구

A Study on the Indexing Standard and Automatic Generation of Back-of-Book Indexes

김효열, 정영미 (연세대학교 문헌정보학과)

Hyo Yeol Kim, Young Mee Chung

(Dept. of Library and Information Science, Yonsei Univ.)

본 논문은 한국어 도서권말색인의 여러 문제점들을 해결하기 위해 기존의 도서권말색인들을 분석하여 한국어 도서권말색인 작성을 위한 지침을 개발하였고, 색인 작성을 좀 더 짧은 시간에 작업하면서도 당라적인 표목을 생성하기 위해 색인 표목을 자동생성하는 시스템을 설계하고 구현하였다.

1 서 론

도서권말색인은 도서의 뒷부분에 있는 것으로, 도서 내용에 포함된 주제항목 또는 개념이 실제로 몇 페이지에 나와 있는가를 지시해 준다. 도서권말색인은 이용자가 필요한 정보를 찾기 위하여 도서 전체를 읽거나 한번 읽은 도서를 다시 읽지 않아도 되도록 소재를 정확하게 지시하여 저자와 이용자간의 인터페이스를 향상시키는 역할을 한다.

그러나 한국어 도서권말색인의 경우 색인작성에 많은 문제점이 있어 색인으로서의 역할을 제대로 수행하지 못하고 있다. 도서권말색인이 색인전문가에 의해 작성되는 경우는 거의 없으며, 대부분 저자나 저자주변 사람들에 의해 이루어지고 있다. 게다가 색인작성자는 색인작업을 하는 데 기준이 될 수 있는 정해진 형식이나 모형이 없어 어려움을 겪고 있으며, 대부분의 색인작성자는 권말색인의 구조를 이해하지 못하고 형식적인 색인을 만드는 실정이다. 또 비교적 짧은 기간에 수작업으로 색인을 작성하기 때문에 색인에 일관성과 표목의 망라성이 부족하다.

따라서 도서권말색인 작성을 위한 지침이 마련되어 색인작성자에게 색인작성의 기준으로 제공된다면 색인의 다양한 구성, 형식, 구조에 관계된 문제점들은 해결될 것이다. 또 근래의 모든 도서들이 문서편집기에 의해 쓰여지고 편집되므로 전문이 수록된 화일로부터 색인의 표목과 소재지시를 자동으로 생성할 수 있다면 수작업에서 발생되는 일관성 결여와 작업시간 부족 등의 문제점은 해결될 수 있을 것이다.

2 한국어 도서권말색인의 분석

본 논문에서는 국내에서 출판되고 있는 한글 단행본의 도서권말색인에 있어 실제로 어떤 차이가 있는가를 알아보기 위하여 색인의 구성, 내용, 편집 등을 비교기준으로 하여 문헌정보학분야에 관련되고 최근에 발행된 단행본 32권의 색인을 비교분석하였다.

1) 도서권말색인의 유형

저자색인, 지명색인, 주제색인 등으로 분리되어 있는 경우는 조사대상의 문헌에서는 보이지 않으며, 대부분 주제, 인명, 지명 등이 포함된 한글색인과 영문색인으로 구성되어 있었다.(32개 색인 중에서 25개, 78%)

2) 색인기입의 구성요소

주표목, 부표목, 소재지시, 상호참조 네가지 요소 모두가 갖추어 진 경우는 조사대상 32권의 색인 중에 7개(22%)였으며, 상호참조, 부표목이 없는 것도 많았다. 32권의 조사대상 권말색인 중에서 부표목을 가진 색인은 17개(53%)였으며, 상호참조를 가진 색인은 8개(25%)였다.

3) 표목의 선택과 표현

본문에서 주제를 표현하는 용어, 인명, 기관명, 단체명, 지명 등은 모두 표목으로 선택되고 있었다.

여러 단어로 된 표목은 대부분 본문 중에 표기된 순서대로 기입하고 있었으며, 몇몇 색인의 경우 어순을 도치시킨 경우도 보이며, 도치한 단어간에 쉼표를 찍거나 사선을 그어 구별하고 있었다.

4) 한국어 도서권말색인 표목의 구문형식

도서권말색인 작성시에 표목을 자동으로 생성할 것을 염두에 두고 실제로 표본 도서의 권말색인을 분석해 본 결과, 표목이 되는 색인어는 19개의 구문형식을 갖는 명사와 명사구로 되어 있었다.

그러나 19개의 구문형식 중 6개는 실제 한국어 도서권말색인에서 표목으로 나타나지만 표목으로 부적합한 구문형식이었다.

5) 주표목과 부표목의 관계

대부분은 하나의 색인어로 쓰이는 단어가 여러 표목내에서 반복해서 나타날 때 공통된 단어를 주표목으로, 공통된 단어를 생략한 나머지를 부표목으로 기입하고 있었다. 또 하나의 표목에 소재지시의 수가 너무 많아 주제의 특정성이 부족할 때 내용을 세분하여 부표목을 선택하고 있었다. 주표목과 부표목과의 관계는 종속, 제한, 전개, 사례, 수식, 각종 연관관계 등이 복합적으로 나타났다. 대개 여러 관계가 복합되어 주표목에 관련된 것은 모두 부표목으로 넣고 있었다.

6) 소재지시

표목으로 선택된 개념어가 어떤 한 주제에 대해 두 페이지 이상에 걸쳐 논의가 될 때 내용의 시작 페이지와 끝 페이지번호를 불임표(-)로 연결하는 색인은 32권의 조사대상 중에서 17개(53%)였다. 나머지 15개(47%)는 주제에 대해 계속되는 논의일지라도 주제에 대해 계속되지 않는 논의와 구별하지 않고 여러 페이지를 연속해서 쉼표로 연결하거나 주제의 시작 페이지만을 지시하고 있었다.

7) 상호참조의 역할과 형식

한국어 도서권말색인의 경우 조사한 32권의 권말색인 중에서 8개(25%)만이 상호참조를 가지고 있으며, 도보라참조를 가진 색인은 4개(13%)에 불과하였다.

보라참조는 기호와 문장으로 표시되는데 동의어, 약어/완전어, 두문자어/완전어, 별명/정식명, 속명/정식명, 번역어/원어, 도치형, 병렬복합어, 류와 구성요소 사이, 동일한 외래어를 상이한 철자로 표기한 경우 등의 관계에서 사용되고 있었다.

도보라참조도 기호와 문장으로 표시되었는데 유사동의어, 연관어, 반의어, 재료와 제품, 상위·하위 개념, 한국어 표기와 영문 표기 사이 등의 관계에서 사용되고 있었다.

3 도서권말색인 작성의 문제점과 개선방안

한국어 도서권말색인의 분석결과 문제점은 요약하면 다음과 같다.

첫째, 색인작성자가 색인 작성작업을 하는데 있어 기준이 될 수 있는 정해진 지침이나 모형이 없다.

둘째, 도서권말색인의 구조를 이해하지 못하는 비전문가인 저자에 의해 대부분의 도서권말색인이 작성된다.

셋째, 색인작업이 비교적 짧은 기간에 수작업으로 이루어진다. 따라서 일관성이 부족하고 표목이 망라적이지 못하다.

한국어 도서권말색인이 색인으로서의 제 기능을 발휘하기 위해서는 이러한 문제점들이 해결되어야 하며 다음과 같은 단계를 통하여 해결할 수 있을 것이다.

첫째, 도서권말색인 작성을 위한 표준이 만들어져야 한다. 정해진 하나의 지침이 있어야 누가 색인을 작성하더라도 이를 기준으로 일관성 있는 색인이 작성될 수 있을 것이다.

둘째, 근래 도서들이 문서편집기에 의해 작성되고 있다는 사실을 감안하여 화일에 저장된 텍스트로부터 도서권말색인의 표목과 소재지시를 자동으로 생성할 수 있는 시스템이 개발된다면 표목의 망라성을 확보할 수 있을 것이다.

셋째, 자동생성시스템에 의해 자동생성된 표목에서 수작업으로 통제를 가한 후에 색인작성지침에 따라 편집하면 될 것이다.

4 도서권말색인 자동생성시스템 구축

4.1 자동생성시스템 개요

본 시스템은 화일에 저장된 본문에서 도서권말색인의 표목이 될 수 있는 색인어를 자동으로 생성하는 시스템이다.

기존의 도서권말색인의 표목을 분석해 본 결과, 표목이 되는 단어나 구는 일정한 구문형식을 가지고 있다는 것을 알 수 있었다. 색인 표목이 되는 구문형식의 공통적인 특징은 명사, 대명사를 포함하는 체언이 단독으로 쓰이거나 이들이 서로 결합하여 복합명사구를 형성한다는 것이다.

본 시스템에서는 문장을 어절단위로 분리하여 체언만 남기고 나머지를 제거한 후 접속사, 후치사, 단어의 끝에 있는 조사, 관형형 어미를 기준으로 일정한 구문형식을 조합하는 방법을 구현하였다.

본 시스템에서는 색인어로 추출하는 구문형식은 앞에서 분석한 19개의 구문형식 중에서 분리 가능한 6개를 제외한 13개로 규정하였다.

4..2 시스템의 설계

1) 어절분리 모듈

문서편집기로 작성된 텍스트에서 각 페이지의 앞에 페이지번호 태그를 주고 정렬방법을 왼쪽정렬로 바꾼 다음 아스키코드로 저장한다.

2) 불용어 제거 모듈

불용어 제거 모듈에서는 체언을 제외한 나머지를 제거한다. 본 시스템에서는 구현의 편의를 위하여 3개의 불용어사전—완전일치 불용어 사전, 좌측절단 불용어 사전, 우측절단 불용어 사전—을 미리 만들었다.

좌측절단 불용어 사전은 용언의 어미와 진성부사의 어간 부분을 제외한 우측부분 등으로, 우측절단 불용어 사전은 용언의 어간, 전성부사의 어간 부분을 중심으로, 완전일치 불용어 사전은 관형사, 부사, 감탄사 등으로 이루어져 어절단위로 분리된 단어가 불용어 사전과 일치할 경우에 제거되도록 하였다.

3) 구문형식의 조합·분리 모듈

이 모듈의 목적은 분리된 어절단위에서 규정된 구문형식과 비교하여 형식이 일치하면 명사 명사구를 조합·분리하여 구문을 형성하는 데 있다. 이 중에서 표목으로 부적합한 병렬 명사구를 포함한 6개의 구문형식은 분리 알고리즘에 의해 두 개의 표목으로 분리한다.

4) 조사 제거 모듈

이 모듈은 앞 단계에서 조합된 명사·명사구의 끝에 붙어 있는 조사를 제거하도록 한다. 조사사전은 최장일치의 원칙에 의하여 구성하였다.

5) 불용어·비주제어 제거 모듈

불용어·비주제어 제거 모듈은 한 어절로 된 구문형식들 중에서 불용어와 비주제어로 보이는 단어들을 제거하는 데 그 목적이 있다. 비주제어 사전은 일반 비주제어 사전과 분야별 비주제어 사전으로 구성된다.

일반주제어 사전은 도서의 주제와 관계없이 일반적으로 주제어로서 적합하지 않은 명사와 다른 단어와 결합하지 않고서는 의미를 가지지 못하는 명사 단어로 구성된 사전이다.

분야별 비주제어 사전은 도서의 주제에 따라 구성되며 그 도서에서 주제어가 될 수 없는 명사를 모아둔 사전이다.

6) 정렬·페이지 처리 모듈

이 모듈에서는 구문들을 정렬하여 불용어나 비주제어로서 제거되어 NULL만 있는 것은 없애고 구문이 일치하는 것은 하나만 남기고 나머지 것들은 제거한다. 그리고 나서 명사·명사구의 뒤에 페이지번호를 단다.

7) 페이지별 저장 모듈

이 모듈은 본문에서 추출된 한 페이지의 표목을 임시로 저장하는 모듈이다. 앞 단계의 작업들이 한 페이지씩 반복해서 수행되기 때문에 페이지 작업 모듈의 수행 결과를 차례로 저장한다.

8) 정렬·색인어 편집 모듈

첫 페이지부터 끝 페이지가 모두 수행되고 나면 페이지별로 저장된 것을 모아 다시 정렬한다. 표목이 반복되는 것은 처음의 표목만 남기고 나머지는 페이지번호만 처음의 표목에 기입하고 표목은 삭제한다.

한 어절로 된 색인어가 다음 색인어의 앞부분에 두 번 이상 연속하여 반복해서 나타날 때 그 단어의 뒷부분이 스페이스, '의'|'에서의'|'으로부터' 등의 조사, '에 관한'|'에 대한' '에 의한' 등의 후치사 등과 일치할 때, 이들을 제거하고 나머지를 부표목으로 나타낸다. 편집 알고리즘에서는 주표목과 부표목의 관계에 있는 표목을 편집하는데 본 시스템에서는 내용전개의 관계만을 자동으로 편집한다.

4.3 자동생성 실험과 결과 분석

본 연구에서 개발한 자동생성시스템을 실험 문헌 6페이지의 본문을 대상으로 실험한 결과 앞 장에서 규정한 구문형식과 일치하는 87개의 명사, 명사구를 추출하였다.

본 연구의 자동생성시스템에 대한 평가척도는 수작업으로 추출한 색인어와 자동생성시스템으로 추출한 색인어에 대해 적합 색인어 비율과 부적합 색인어 비율을 이용하였다.

실험결과 본 시스템의 적합 색인어 비율은 98%로 적합한 색인어는 거의 모두 생성되는 것으로 나타났으며, 부적합 색인어 비율은 49%로 나타났다. 본 시스템은 거의 모든 적합 색인어를 자동으로 생성함으로써 수작업 색인에서 비롯되는 망라성의 문제를 해결할 수 있었으며, 생성된 색인어 중에서 부적합 색인어만 제거시키면 도서권말색인의 표목으로 쓸 수 있었다

본 시스템에서 자동으로 생성되는 색인어의 수는 도서권말색인의 색인어의 분량으로는 다소 많아 전부를 도서권말색인의 표목으로 사용할 수는 없으며, 저자의 수작업 통제를 거쳐야 한다는 한계점이 있다.

5 결 론

본 연구에서는 한국어 도서권말색인의 분석 결과를 기반으로 하여 한국어 단행본의 도서권말색인 작성을 위한 지침을 제시하였다. 이 지침은 도서마다 다르게 나타나는 기존의 색인의 구성, 내용, 편집형식 등에 일관성을 제시할 수 있을 것이다. 본 논문에서 제시한 색인 작성지침을 요약하면 다음과 같다.

첫째, 색인기입의 기본요소인 주표목, 부표목, 소재지시, 상호참조 각각에 대해 예를 들어 쓰임을 설명하였다. 둘째, 부표목은 한 개의 표목으로는 소재지시의 수가 너무 많아 내용을 세분하여 분리할 필요가 있을 때에 사용하며, 또 하나의 색인어가 여러 색인어에서 반복해서 나타날 때 공통된 단어를 주표목으로 공통된 단어를 생략한 나머지를 부표목으로 기입하도록 하였다. 셋째, 상호참조는 표시하는 기호로 보라참조는 '→'를, 도보라참조는 '⇒'를 사용하도록 하였다. 넷째, 주표목과 부표목의 편집에 있어 부표목의 앞부분에서 주표목과 동일하게 반복되는 어절은 생략하도록 하였다. 다만 제한의 관계일 때는 부표목의 한 어절내에서 주표목이 반복되어 생략되는 경우 그 자리에 붙임표로 생략되었음을 표시하도록 하였다. 부표목의 중간이나 뒷부분에서 주표목을 반복하는 부분은 그대로 두도록 하였다.

본 논문에서 설계·구현한 도서권말색인 자동생성시스템의 특징은 다음과 같다.

첫째, 본 시스템은 구문분석/의미분석 없이 단순한 구문형식 조합에 의해 색인어를 생성하였기 때문에 자동생성 후에 저자가 수작업 통제를 통하여 부적합 색인어를 제거하고 최종적으로 색인어를 선택하는 후통제 색인시스템이다.

둘째, 본 시스템은 기존 한국어 도서권말색인에서 나타나는 색인어 구문형식을 분석하여, 이를 바탕으로 미리 규정한 13개의 구문형식과 본문 중의 구문형식이 일치할 때 소재지시와 함께 색인어로 생성되도록 하였다.

셋째, 실험결과 본 시스템의 적합 색인어 비율은 98%로 적합한 색인어는 거의 모두 생성되는 것으로 보여지며 부적합 색인어 비율은 49%로 나타났다. 본 시스템은 거의 모든 적합 색인어를 자동으로 생성함으로써 수작업 색인에서 비롯되는 망라성의 문제를 해결할 수 있었다.

<참고문헌>

김석영, 초록 및 색인론. (서울:산업기술연구원), 1992.

최석두, "도서권말색인에 관한 연구", 이화여대 인문대 제2회 교수학술제 발표문,1994.

ANSI Z39.4-1984.

BS 3700:1988.

Diodato, Virgil, "User Prefernces for Features in Back of Book Indexes," JASIS Vol.45 No.7(1994), pp.529-636.

Mulvany, Nancy C., Indexing Books, (Chicago: The Univ. of Chicago Press), 1994.

The University of Chicago Press, The Chicago Manual of Style, 14th ed., (Chicago: The Univ. of Chicago Press), 1993.