서지마크 데이터베이스로부터의 색인어 추출과 색인어의 검색 활용에 관한 연구

- 경북대학교 도서관 학술정보시스템 사례를 중심으로 -

A Study on the Extraction and Utilization of Index from Bibliographic MARC Database

박 미 성(Mi-Sung Park)*

〈 목 차 〉

Ⅰ. 서 론
1. 연구의 필요성과 목적
2. 관련연구와 연구 내용
Ⅱ. 서지마크로부터의 색인어 추출 이론
1. 색인 태그 정의
2. 색인어 정규화 알고리즘
Ⅲ. 컨텐츠의 색인어 추출 사례
Ⅳ. 색인어 분석 및 색인어의 검색 활용도 분석
1. 색인코드 유형별 색인어 수 분석
2. 색인코드 유형별 검색 활용도 분석
Ⅴ. 결 론

초 록

본 연구의 목적은 서지정보검색시스템의 색인 정의의 중요성을 강조하고 최적 색인의 기초 자료를 마련하기 위함이다. 이를 위해 서지마크 데이터베이스로부터 색인 태그 정의 및 정규화를 통한 색인어추출이론에 대해 고찰하고, 이론에 따라 생성된 색인어의 검색 활용도를 분석하였다. 실험은 서지 2,200,488건에서 생성된 색인어 29,219,853건을 텍스트형 색인과 코드형 색인으로 나누어 이용자 웹 검색 로그에 나타난 색인 항목과 비교하여 어떤 색인 정의가 얼마나 활용되는가를 분석하였다. 결과에 따르면 서명, 저자, 출판사, 주제와 같은 텍스트형 색인어는 높은 검색 활용도를 보인 반면에 코드형 색인어는 검색 활용도가 낮아 검색에 활용되지 않는 불필요한 색인 정의들은 과감하게 제거하여 색인 정의를 최적화해야 함을 제안하였다.

주제어: 색인 정의, 정규화, 텍스트형 색인, 코드형 색인, 색인어 추출

ABSTRACT

The purpose of this study is to emphasize the importance of index definition and to prepare the basis of optimal index in bibliographic retrieval system. For the purpose, this research studied a index extraction theory on index tag definition and index normalization from the bibliographic marc database and analyzed a retrieval utilization rate of extracted index. In this experiment, we divided index between text-type and code-type about the generated 29,219,853 indexes from 2,200,488 bibliographic records and analyzed utilization rate by the comparison of index-type and index term of web logs. According to the result, the text-type indexes such as title, author, publication, subject are showed high utilization rate while the code-type indexes were showed low utilization rate. So this study suggests that the unused index is removed from index definition to optimize index.

Key Words: Index Definition, Normalization, Text-type Index, Code-type Index, Index Extraction

* 경북대학교 중앙도서관 전산관리팀 팀장(mspark@knu.ac.kr)
• 접수일: 2005년 5월 9일 • 최초심사일: 2005년 5월 30일 • 최종심사일: 2005년 5월 30일

Ⅰ. 서 론

1. 연구의 필요성과 목적

정보검색시스템의 종류는 다양한데, 검색되는 정보의 유형에 따라 데이터검색시스템, 참조정보검색시스템, 본문검색시스템, 질의응답시스템, 비디오텍스 등으로 분류할 수 있다. 이 중에서 참조정보검색시스템(reference retrieval system)은 문헌의 서지 사항과 같이 정보원에 대한 참조정보(이차정보)를 검색하는 시스템으로 문헌정보검색시스템 또는 서지정보검색시스템이라고도 한다.[1] 정보검색시스템의 종류에 따라 수행되는 작업 내용이 달라지지만 어떤 유형의 검색시스템이던지 간에 효과적인 검색을 위해 필수적으로 선행되어야 할 과정은 색인 작업이다. 색인이란 정보자료의 내용을 대표할만한 특성을 표현하는 데이터 요소를 뽑아 각 정보자료의 내용을 대표하도록 한 것으로 색인이 부여되어 있느냐 없느냐에 따라 검색 결과가 결정되어 정보검색시스템의 성능을 좌우하게 된다. 그러므로 이상적인 색인을 구축함은 최상의 검색 효율을 보장[2]할 뿐만 아니라 정보자원과 정보이용자 사이에서 원하는 선별된 정보만 제공해 줄 수 있는 커뮤니케이션 채널[3][4]과 같은 것이라 할 수 있다. 이처럼 이상적인 색인은 이용자 활용 측면에서 매우 중요한 요소로서, 색인어를 잘 정의하여 체계적으로 조직화하여 구축하는 일은 무엇보다도 중요한 일이라 하겠다. 본 연구에서는 참조정보검색시스템(reference retrieval system)으로 분류되는 대학 도서관 학술정보시스템에서 유용한 검색 도구로 사용되고 있는 색인이 어떻게 정의되며, 색인어로 생성되기까지의 추출 과정과 추출된 색인어 중 어떤 종류의 색인들이 실제 검색에 자주 활용되는지에 대해 논의해 보고자 한다.

본 연구의 목적은 첫째, 서지데이터베이스에 입력된 마크 데이터로부터 색인어 추출 과정을 이해하게 함으로 마크 데이터 입력 시 이용자 검색 활용도를 고려한 양질의 데이터 입력을 가능하게 하고, 동시에 이용자에게 좀 더 정확한 검색서비스를 할 수 있게 되리라 생각한다. 둘째, 부여된 색인이 실제 검색에서 어느 정도 활용되고 있는가를 분석하여 학술정보검색시스템에 기 정의된 색인 태그가 이상적인 색인 태그 즉 최적의 색인태그로 정의된 것인가를 평가해 볼 기초자료를 마련하기 위함이다. 셋째는 색인 태그 활용도를 분석한 기초자료를 토대로 기존 색인 태그 정의의 문제점을 발견하고, 이를 토대로 향후에 시스템의 색인 태그를 재 정의할 근거를 마련하는 것이다. 이는 불필요한 색인 정의로 생성된 색인어를 줄여나감으로 색인어 규모가 방대해 지는 것을 막을 수 있고 또한 부적합 문헌의 검색 비율(잡음율)[5]이 높아짐으로 발생될 시스템 검색 속도 저하 방지와

1) 정영미, 정보검색론(서울 : 구미무역출판부, 1993), p.11.
2) 이두영, 남영준, 인터넷 도서관과 정보검색(전주 : 전주대학교출판사, 2003), p.48.
3) 정영미, 전게서, p.52.
4) 이수상, 성공적인 개인정보관리를 위한 색인노트법(서울 : 한울, 1998), p.103.
5) 정영미, 전게서, pp.290-293.

불필요한 자료의 검색을 막아 적합한 문헌만 검색되게 함으로 정보검색시스템의 성능 향상을 도모할 수 있다.

2. 관련 연구와 연구 내용

정보검색시스템의 유형과 대상에 따라 색인어를 추출하는 방법들이 다양하게 연구되어 왔다. 전통적으로 색인 작업은 훈련된 사서나 주제전문가에 의해 수행되어 왔으나 색인해야 할 문헌의 급속한 증가와 주제지식을 갖춘 색인 경험자 부족, 컴퓨터의 발달로 수치자료 뿐만 아니라 문자 자료도 완벽하게 처리가 가능하게 되고, 인공지능의 분야로 언어 구조와 의미를 연구하는 새로운 연구분야가 발전하게 됨에 따라 컴퓨터를 이용한 다양한 자동 색인 기법들이 출현하게 되었다.[6] 자동색인의 효시를 이룬 것은 룬(H.P. Luhn)의 문헌에 나타난 단어들의 출현빈도가 문헌의 내용을 나타내는 주제어로 중요성을 측정하는 기준이 된다는 가설에서 비롯되었다.[7][8] 룬은 고빈도의 단어와 저빈도의 단어는 주제로서 가치가 없다고 판단하여 이를 제외한 중간 빈도어를 색인어로 선정하도록 하였다. 그리고 초기 룬과 함께 자동 색인의 기초를 이룬 박센데일(P.B. Baxendale)[9]은 세 가지 색인 방법을 제시하였다. 첫 번째 방법은 기능어[10]를 제외한 모든 단어를 색인어로 선정하는 방법으로 후에 대부분의 자동색인 기법의 기본적인 방법으로 사용되고 있다. 두 번째 방법은 각 문단의 첫 번째 문장과 마지막 문장에 출현한 단어를 색인어로 선택하는 방법으로 문헌의 구조적 특성을 이용한 방법이다. 세 번째 제시한 방법은 문헌을 구성하는 전치사구로부터 색인어를 선택하는 방법으로 구문분석 기법의 일종이라 볼 수 있다. 이처럼 색인어를 선정하는 기준에 따라 통계적 기법, 문헌구조적 기법, 언어학적 기법으로 나눌 수 있다. 통계적 기법은 단어의 출현빈도를 근거로 주제어로서의 중요도를 측정한 다음 색인어를 추출하는 기법으로 대표적 연구로는 마론(M.E. Maron)과 쿤스(J.L. Kuhns)[11]의 확률색인기법과 다메로우(F.J. Damerau)[12]의 상대빈도를 들 수 있다. 확률색인기법은 특정한 질문에 대해 각 문헌이 적합할 확률과 부적합할 확률을

6) H. Borko and L.B. Charles, *Indexing Concepts and Method*(New York : Academic Press, 1978), p.113.
7) H.P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Library Information," *IBM Journal of Research and Development*, Vol.1, No.4(1957), pp.309-317.
8) H.P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, Vol.2, No.2(1958), pp.159-165.
9) P.B. Baxendale, "Machine-Made Index for Technical Literature-An Experiment," *IBM Journal of Research and Development*, Vol.2, No.4(1958), pp.354-361.
10) 기능어는 대명사, 전치사, 관형사, 보조동사, 접속사, 조사, 그리고 일반적인 의미의 형용사나 부사 등 주제적 의미가 없는 단어를 총칭한다.
11) M.E. Maron and J.L.Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of Association for Computing Machinery*, Vol.7, No.3(1960), pp.216-244.
12) F.J. Damerau, "An Experiment in Automatic Indexing," *American Documentation*, Vol.6, No.4(1965), pp.283-289.

산출하여 적합 확률이 부적합 확률보다 큰 문헌을 먼저 출력되게 하는 방법으로 출력 문헌의 순위를 매기기 위해 색인정보를 문헌정보검색시스템에 사용하도록 한 것이다. 그리고 상대 빈도는 단순한 단어 빈도를 사용할 경우 색인어를 선정하는 절대적 기준을 정하기 어렵다는데 착안하여 단어빈도[13], 문헌빈도[14], 장서빈도[15] 등을 나누어줌으로써 빈도의 값을 표준화시켜 사용하는 방법이다. 1970년대에 들어서면서는 장서나 특정 문헌내의 용어의 가치를 측정하여 색인의 목적에만 적용시키던 이론을 검색의 목적과 함께 고려하게 되었다. 그 대표적인 것이 스파크 존스(J.Spark)[16]의 적은 수의 문헌에 나타난 단어에 높은 중요도를 부여하는 역문헌빈도론과 역문헌빈도를 가중치로 둔 색인방식이 효과적인 검색 결과를 가져옴을 입증한 샐튼의 문헌분리가 이론이 있다. 그리고 하터(S.P. Harter)[17]의 동일문헌에 함께 군집하여 출현하는 단어가 색인어로 유용하다던 포아송(Poisson) 모델도 있다. 그러나 1980년대 이후에는 자동색인에 언어학적 기법의 단서어기법과 구문분석기법이 주류를 이루었다. 대부분의 연구들은 이론정립보다 자동색인과 수작업 색인과의 비교를 통해 이론보다는 경험에 많이 의존하는 방식의 연구들이었다. 그리고 최근에는 기존 연구를 바탕으로 한가지 방식을 사용하는 것이 아니라 여러 가지 통계적 기법과 한국어의 언어적 특성을 복합적으로 활용하는 연구들[18][19][20][21][22]이 등장하였는데 대표적인 연구가 단일어보다 어휘 특정성이 큰 복합명사를 색인어로 가치가 있다고 판단하여 명사간의 상호 정보나 통계정보를 활용한 복합명사 색인어 추출에 대한 연구들이 등장하고 있다. 이처럼 색인을 추출하기 위한 다양한 연구들이 오랫동안 이루어져 왔다. 본 연구의 내용으로 들어가기에 앞서 우선 색인어의 종류를 간단히 살펴보면, 색인어 종류는 다양한데, 색인어의 유형에 따라 주제색인과 비주제색인[23] 또는 주제색인과 형식색인[24]으로 구분한다. 주제 색인은 정보 자원의 내용을 분석하여 주제를 나

13) 단어빈도(TF)는 색인 대상이 되는 각 문헌 i 에 특정한 단어 k가 출현한 횟수를 말한다.
14) 문헌빈도(DF)는 특정한 단어 k가 출현한 문헌의 수를 말한다.
15) 장서빈도(CF)는 특정한 단어 k가 전체 문헌 집단 내에 출현한 총 빈도를 말한다.
16) J. Spark, "A Statistical Interpretation of Term Specificity and its Application in Retrieval", *Journal of Documentation*, Vol.28, No.1(1972), pp.11-20.
17) S. P. Harter, "A Probabilistic Apporach Automatic Keyword Indexing : Part I. On the Distribution of Specialty Words in a Technical Literature," *Journal of chemical Information and Computer Sciences*, Vol.26, No.4(1975), pp.197-206.
18) 김민정, 한글 특성을 고려한 자동 색인기법(석사학위논문, 부산대학교 대학원 전자계산학과, 1993), p.35.
19) 신동욱, "복합명사의 통계적 처리에 대한 평가," 한글 및 한국어 정보처리 학술발표논문집(1997. 10), pp.36-41.
20) 김판구, 조유근, "상호 정보 기반한 한국어 텍스트의 복합어 자동 생성," 한국정보과학회 논문지, 제21권, 7호(1994), pp.1333-1340.
21) 김미진 등, "효율적인 색인어 추출을 위한 합성명사 생성방안에 대한 연구," 한국정보처리학회지, 제7권, 제4호(2000, 2), pp.1123-1127.
22) 박미성, "음성데이터베이스로부터의 효율적인 색인데이터베이스 구축과 정보검색," 한국도서관정보학회지, 제35권, 제3호(2004, 9), pp.271-291.
23) 정영미, 전게서, p.51.
24) 이수상, 전게서, p.105.

타내는 색인어를 색인 요소로 채택하는 경우를 말하며, 비주제색인(형식색인)은 저자명, 표제명, 기관명, 형태사항과 같이 주제와 직접적인 관련이 없는 색인 요소를 색인어로 채택하는 경우를 말한다. 앞서 기술한 관련 연구들은 대부분 주제색인을 컴퓨터가 자동으로 추출하는 방법들에 대한 연구라고 한다면 본 연구의 대상이 되는 서지검색시스템에서의 색인은 두 종류의 색인이 모두 포함된 색인이라 할 수 있다. 왜냐하면 사서에 의해 입력된 마크데이터 내에 주제를 분석하여 입력한 내용도 포함되어 있기 때문이다. 그리고 기존 연구에서 추출하는 색인은 문헌 내용으로부터 통계적 기법이나 문헌구조적 기법, 언어학적 기법 등을 사용하여 추출한 색인이 얼마나 가치가 있는지 평가하고 있는 반면에 본 연구에서 추출한 색인은 사서가 입력한 문헌의 참조 정보인 마크로부터 최적의 색인을 정의하고 그 정의대로 색인 대상이 추출되어 검색에 활용되기 때문에 서지정보시스템에서의 색인은 색인 태그가 얼마나 최적으로 정의되어 있느냐에 따라 색인어로서의 가치를 평가받게 된다. 그러므로 이용자 검색을 고려한 양질의 서지 입력도 중요하지만 더불어 최적의 색인 태그를 정의하기 위한 노력과 연구도 함께 이루어져야 한다.

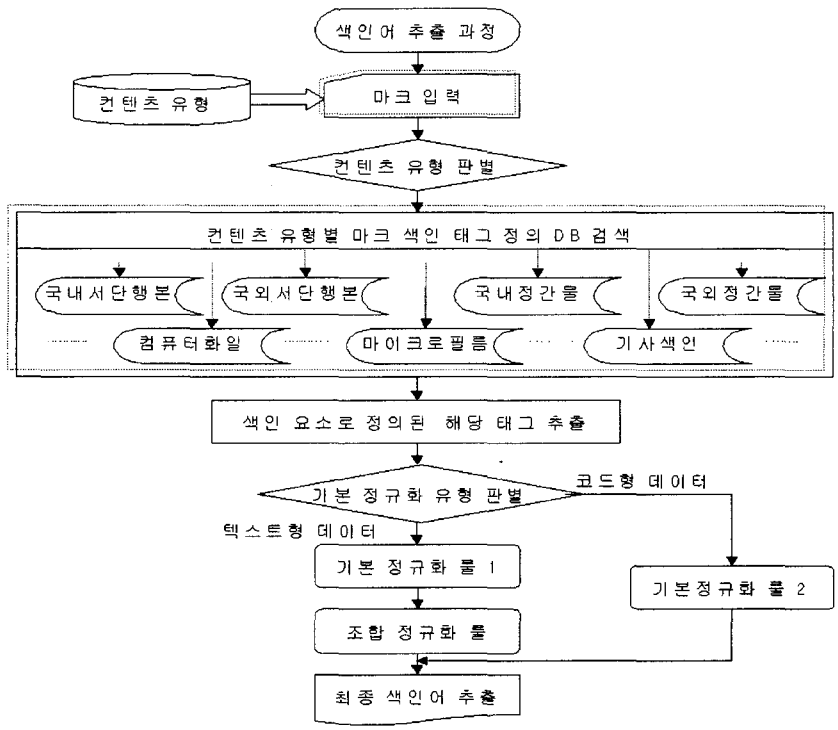
본 연구에서는 경북대학교 도서관 학술정보시스템의 MARC, MARC21, XML로 구축된 약 220만 레코드의 서지로부터 색인 태그가 어떻게 정의되어 색인 대상을 추출하는지, 추출한 색인 대상들은 어떻게 정규화 되어 색인어로 생성되는가에 대한 이론을 먼저 설명한다. 그리고 설명한 이론에 근거하여 컨텐츠의 실제 사례로부터 색인어가 생성되는 전체 과정을 보인다. 더불어 색인 정의에 따라 추출된 색인어의 비율과 이용자 검색 로그를 분석하여 어떤 유형의 색인어가 주로 검색에 활용되는지 그리고 활용되지 않는 색인어에는 어떤 것이 있는 지를 분석한 후 색인 정의의 문제점을 살펴보고 향후 연구과제로 색인 정의 최적화 방안을 논의 한 후 결론을 맺는다.

Ⅱ. 서지마크로부터의 색인어 추출 이론

대학 도서관 학술정보시스템에서 제공하는 정보자원에는 인쇄된 문헌과 비디오, 오디오, CD-ROM, DVD, E-Book, E-Journal 및 Web DB와 같은 전자정보나 멀티미디어정보 등 다양한 자원들이 있고 앞으로는 더 새로운 유형의 매체가 지속적으로 출현하게 될 것이다. 이처럼 다양한 정보자원에 일정한 질서를 부여하고 체계적으로 조직하고 정리하여 이용자가 원활하게 활용할 수 있도록 하는 서지통정(bibliographic control)은 도서관 봉사의 기본행위이다. 다시 말해 도서관 봉사의 기본 행위에는 정보자원에 대한 특징과 내용을 기록하고 특정한 형식으로 표시하는 서지정보 기술 행위와 특정 형식에 맞게 구축된 정보를 이용자가 잘 활용할 수 있도록 색인을 생산하는 행위를 포함하고 있다.[25] 이와 같은 기본 행위의 결과로 작성된 색인은 색인자의 입장에서는 색인어이지

25) 이수상, 전게서, pp.76-77.

만 검색자의 입장에서는 검색어가 되기 때문에 어떤 항목을 색인으로 정의하느냐에 따라 검색 결과가 결정되어 검색시스템의 성능을 결정짓게 된다. 그러므로 색인 작성 시 주의해야 할 사항은 불필요한 색인을 과다하게 정의하여 추출한다거나 너무 적거나 적절하지 않는 색인어를 추출하는 것은 문제가 되며, 이것은 곧 검색 효율을 저하시키는 요인이 된다는 것이다. 이처럼 검색 효율에 중요한 영향을 미치며, 이용자 검색에 활용될 색인이 도서관 서지정보검색시스템의 서지마크로부터 어떻게 정의되어 색인대상이 추출되는지 그리고 정규화 되어 최종 색인어로 생성되는지에 대한 이론을 이 장에서 먼저 고찰하고자 한다. 우선 입력된 서지마크로부터 색인어 추출 과정의 흐름도를 그려보면 〈그림 1〉과 같다.



〈그림 1〉 색인어 추출 과정

정보구축자(사서)가 문헌을 분석하여 그 유형이 국내외 단행본인지, 국내외 정기간행물인지, 아니면 컴퓨터 파일인가 하는 컨텐츠의 유형을 판별한 후 해당 컨테이너를 가져와서 문헌에 대한 마크 정보를 입력하게 된다. 입력이 완료되면 컨텐츠 유형에 따라 미리 정의된 마크 색인 태그 정의 D/B를 검색하여 색인 요소로 정의된 항목에 해당하는 태그들을 순차적으로 추출하게 된다. 그리고 추출된 색인 태그 항목들은 입력된 내용 그대로 색인어가 되는 것이 아니라 마크 기술 규칙에

포함된 다양한 부호나 공백, 탭, 한글/한자/외국어/특수문자 등을 삭제 또는 대체하는 등 여러 방식으로 가공된다. 이는 이용자가 검색어를 어떤 형식으로 입력하던지 색인어와 잘 일치되어 검색이 가능하도록 하기 위함이다. 이처럼 색인어에 다양한 룰을 적용시켜 변화를 가하는 전처리 과정을 정규화(Normalize)라 하는데, 이러한 정규화 과정을 거쳐서 적합한 최종 색인어가 생성된다. 이러한 정규화 과정은 기존 연구들의 자동 색인과는 차이가 큰 부분이다. 다음은 색인어 추출과정의 가장 중요한 두 부분인 색인 태그 정의와 정규화 알고리즘에 대해 자세히 알아보기로 한다.

1. 색인 태그 정의

모든 정보자원을 유형에 관계없이 Kormarc, Usmarc(Marc21)에 의거하여 구축하던 기존의 도서관 학술정보시스템[26]이 현재 다양한 정보 자원의 메타 데이터 관리를 수용하기 위한 통합 전자도서관시스템[27]으로 발전하면서 나타난 중요한 변화 중 하나가 색인 정의 부분이다. 기존 시스템에서는 한번 정의되어 정형화된 색인에 대해서는 필요한 색인이던지 필요하지 않는 색인이던지 상관없이 색인 구조 변경이 어렵도록 설계되어 있었다. 하지만 전자도서관시스템에서는 색인 구조를 정보구축자가 직접 정의할 수도 있고 또한 불필요한 색인 정의는 제거할 수 있도록 설계되었다. 이 점은 개발 시스템과는 독립적으로 융통성 있는 색인 정의로 이용자 검색에 필요한 색인 요소를 선별하여 자유롭게 구축할 수 있도록 하였다는 것이다. 이러한 융통성은 색인이 이용자 검색에 있어서 중요한 요소임을 강조하는 것이고, 지금까지의 거의 정형화된 색인을 자관 시스템과 이용자 검색 환경에 따라 융통성 있게 색인을 재정의 하고 구축할 수 있도록 한 것이다. 이는 지금까지 색인 정의에 대해 개발시스템에게 맡겨두었던 부분을 정보구축자로 하여금 색인 요소를 직접 선별해서 정의하게 하여 색인 구축에 대한 기본 본분을 정보구축자에게 되돌려 준 셈이다. 그러므로 정보구축자는 색인 정의의 중요성을 인식하고 최적의 색인을 정의하려고 노력해야 할 것이다.

본 연구 대상의 서지정보시스템에는 62종의 서로 다른 컨텐츠 유형의 색인정의가 존재한다. 정의 내용을 살펴보면 대부분 공통적인 색인 태그 정의를 가지며, 극히 일부 색인 태그 정의는 어떤 특정 컨텐츠에만 국한되어 있다. 이는 색인 정의가 컨텐츠 유형별 연구를 통한 색인 정의라기보다는 특별한 기준 없이 정의되어 있음을 시사한다고 볼 수 있다. 그리고 색인의 유형은 기존 연구에서 말하는 주제 색인과 비주제 색인(형식색인)을 명확하게 구분 짓기보다는 약간 섞여 있는 형태로, 색인어를 가공하는 정규화 알고리즘 적용 규칙에 따라 크게 텍스트형 색인과 코드형 색인으로 구분하여 분석하였다. 자세한 정의와 내용은 잠시 뒤에 언급하기로 하고, 우선 색인 태그 정의 예를 살펴보기로 하겠다. 〈그림 2〉는 컨텐츠유형 중 가장 많은 부분을 차지하는 국내서 단행본의 저자

26) SOLARS - SOLARS 3.0, SOLARS 4.0 SOLARS SE 버전, (주)INEK 개발 솔루션
27) SOLARS DLi, 통합 전자도서관 솔루션, (주)INEK 개발 솔루션

색인 태그 정의 내용의 일부분이다.

색인명	검색용	전시용	전거TAG	기준태그	TAG1	TAG2	TAG3	TAG4	TAG5	TAG6	TAG7
저자(AUTH)	Y	Y	100/500	100	a						
저자(AUTH)	Y	N	100/500	100	a	b	c	d	f	g	q
저자(AUTH)	Y	N	100/500	100	f	a					
저자(AUTH)	Y	Y	110/111/151/510/511/551	110	a	b	g	k			
저자(AUTH)	Y	N		110	p						
저자(AUTH)	Y	Y	111/151/511/551	111	a	e	g				
저자(AUTH)	Y	N	111/151	111	p						

〈그림 2〉 국내서 단행본의 저자 색인 태그 정의

〈그림 2〉에서 가장 좌측 항목인 "색인명" 필드는 검색 시, 이용자 검색 항목과 일치하는 항목이고, "검색용" 필드에서 "Y"는 실제 검색에 활용되는 색인임을 의미한다. 그리고 "전시용" 필드가 "Y"로 지정되면 이용자 검색 화면의 간략 서지에서 보여질 내용으로 채택된다. "전거태그" 필드는 전거 통제 시 전거의 헤딩 태그를 넣어주는 필드이다. "기준태그" 필드는 색인을 만들 기준 태그를 정의하는 것이며, 나머지 "TAG" 필드들은 색인어로 생성될 서브 필드를 정의한 것이다. 〈그림 2〉는 저자에 대한 7개의 복수 개 색인 정의를 예로 보여주고 있다. 색인명은 저자(AUTH)이고 색인 태그는 100 $a, 100 $a $b $c $d $f $g $q, 100 $f $a, 110 $a $b $g $k, 110 $p, 111 $a $e $g, 111 $p로 모두 7개이다. 그 중 100 $f $a처럼, 여러 개의 식별기호(subfield code)가 함께 나열된 것은 100 태그의 서브 필드 $f와 $a가 합쳐진 내용이 색인으로 생성되고 합쳐진 내용으로 이용자 검색이 가능하게 됨을 의미한다. 그리고 이렇게 정의된 색인 태그로부터 추출된 색인 후보들은 내용의 성향에 따라 서로 다른 정규화 룰이 적용되어 최종 색인어가 생성됨으로, 전체 색인 내용을 성향에 따라 텍스트형 색인과 코드형 색인 두 가지로 나누어 〈표 1〉, 〈표 2〉로 정리하였다.

먼저 텍스트형 색인이란 "저자", "서명", "잡지명", "출판사", "주제"와 같이 태그의 내용이 대부분 문자로 구성되어질 가능성이 높은 색인을 지칭하여 텍스트형 색인이라 정의하였다. 텍스트형 색인어 중에 색인코드 SUBJ는 정보구축자가 문헌의 주제를 분석하여 입력한 내용으로 색인 유형 구분 중 주제 색인으로 볼 수 있고, 나머지는 비주제(형식) 색인으로 볼 수 있다.

코드형 색인이란 "청구기호", "분류기호", "ISSN", "ISBN", "KDC", "DDC", "언어" 와 같이 태그의 내용이 대부분 약속된 문자나 숫자나 기호로 구성되어질 가능성이 높은 색인을 지칭하여 코드형 색인이라 정의하였다. 코드형 색인에서 분류기호 또한 주제 분석의 결과로 얻어진 주제 색인의 개념으로 볼 수 있다. 이처럼 서지검색시스템의 색인은 주제 색인과 비주제 색인이 혼합된 형태의 색인으로 간주 할 수 있다.

다음 〈표 1〉은 컨텐츠 유형별로 정의된 색인의 내용 중 텍스트형 성격을 띄는 색인을 추출하여 정리해 놓은 것으로, 기사색인의 경우는 마크로 구축하지 않고 XML로 구축하였으므로 XML[28] 엘

리먼트 이름으로 정리하였다. "저자사항-개인명-개인명" 표현은 XML문서의 계층구조 표현으로, 부모 엘리먼트인 "저자사항" 하부에 자식 엘리먼트인 "개인명"이 있고, 또 그 하부에 있는 자식 엘리먼트 "개인명"을 추출하라는 의미이다. 〈표 1〉에서 정의된 내용들은 정규화 되어 색인어로 생성될 때, 다음 장에서 설명할 기본 정규화 룰 1과 조합 정규화 룰을 적용하여 변환될 색인 태그들이다.

〈표 1〉 텍스트형 색인 정의 테이블

색인코드	색인명	색인수	태그	
AUTH	저자	26	100 [a] | [abcdfgq] | [fa] 110 [abgk] | [p] 111 [aeg] | [p] 500 [ls] 505 [d] | [e] 507 [abcdfgq]	700 [fa] | [abcdfgq] 710 [abgk] 711 [aeg] 777 [t] 900 [abcdfgq] | [fa] 910 [abgk] 911 [aeg]
			[기사색인 : XML 정의에서 추출] 저자사항-개인명-개인명 저자사항-단체명-단체명 저자사항-회의명-회의명이름	 부출표목-부출개인명 부출표목-부출단체명 부출표목-부출회의명
JTIT	저널명(기사색인)	1	[기사색인] 잡지정보-잡지명	
MJ	학위논문전공학과	1	502[c]	
PUBN	출판사	2	260[b]	[기사색인]잡지정보-출판사
SUBJ	주제	12	600 [abcdfgqxyz] 610 [fa] | [abgkxyz] | [p] 611 [aegxyz] | [p] 630 [anpxyz] | [p]	650 [axyz] 651 [axyz] 653 [a] 949 [anp]
TITL	자료명(서명)	67	130 [anp] | [p] 210 [a] 222 [a] 240 [a] | [g] | [p] 245 [a] | [b] | [abnp] | [x] 246 [a] 247 [a] 440 [anp] | [p] | [s] 500 [befghijkmr] 507 [t] 700 [t] 710 [p] | [t] 711 [p] | [t] 730 [anp] | [p] 740 [a] 760 [s] | [t] 762 [s] | [t] 765 [s] | [t]	767 [s] | [t] 770 [s] | [t] 772 [s] | [t] 773 [s] | [t] 775 [s] | [t] 776 [s] | [t] 777 [s] 780 [s] | [t] 785 [s] | [t] 787 [s] | [t] 830 [anp] | [p] 900 [t] 910 [t] 911 [p] | [t] 930 [anp] | [p] 940 [a] 949 [anp]
			[기사색인 : XML 정의에서 추출] 서명사항-기사명 서명사항-부기사명 서명사항-대등서명	 서명사항-번역서명 서명사항-기사명&부기사명 서명사항-잡제
UV	학위논문수여기관	1	502 [b]	

28) 예) 저자사항(dc:creater)-개인명(dli:person)-개인명(dli:personName)
부출표목(dli:addedEntry)-부출회의명(dli:addedEntryConference)

정보구축자가 입력한 마크/XML 데이터에 〈표 1〉에서 정의한 태그와 그의 식별기호([a], [b] 등)가 포함되어 있어야 색인 후보로 추출된다. 그리고 태그 부분에 표기된 내용은 다음과 같이 해석할 수 있다. 예를 들어, 자료명(서명) 색인의 한 경우인 245 [a] | [b] | [abnp] | [x]는 마크 내에 245태그의 $a로 색인 하나 생성, $b로 색인 하나 생성, $a + $b + $n + $p 내용을 합쳐서 색인 하나 생성, 마지막으로 $x 내용으로 색인 하나 생성하게 됨을 나타낸다. 여기서 한 가지 주의할 사항은 만약 색인이 245 [a]와 245 [abnp] 두 가지 정의에 의해 색인이 생성될 수 있다면 최장일치 결합 색인 하나만 생성하게 된다는 것이다. 왜냐하면 식별기호 [a]와 [abnp]는 [a] ∈ [abnp]라는 포함 관계가 성립함으로 굳이 [a]로 색인하지 않아도 [abnp]색인에 의해 [a]는 당연히 검색이 가능하게 되기 때문이다. 그러므로 입력한 마크의 245태그에 식별기호 $a, $b, $n, $p 모두 포함하고 있다고 가정하면 최대 3개의 색인이 생성될 수 있고, 만약 입력한 245태그에 $a, $b만 존재한다면 $b와 $a+$b 이렇게 2개의 색인을 생성하게 됨을 의미한다. 이러한 원리로 생성된 색인은 실제 검색에 활용되는데, 이용자가 "서명" 항목에 질의어를 넣고 검색을 하게 되면 〈표 1〉의 색인코드 TITL에 정의된 모든 태그로부터 생성된 색인어 내용과 질의어를 비교해서 검색 결과를 가져오게 된다.

다음 〈표 2〉는 코드형 성격을 띄는 색인어에 대해 정리한 코드형 색인 태그 정의 테이블이다. 이것은 나중에 색인어로 생성될 때 기본 정규화 룰 2에 따라 변환되어 색인어로 생성될 태그들이다. 〈표 2〉에서 부호화정보필드인 008 태그는 식별기호가 없는 고정장 40자리여서 색인어를 추출하기 위한 표현 방법을 다른 태그와 조금 다르게 표현하였다. 예를 들어 색인코드 BI(전기형식)는 008 [34]--1로 표현하였는데 이는 008 태그의 34번째 자리에서 1자리 값만 색인 대상으로 추출됨을 의미한다. 그리고 색인코드 LANG(언어)의 008[35]--3은 35번째 자리부터 3자리 값을 색인 대상으로 채택하여 추출한다는 의미이다.

〈표 2〉 코드형 색인 정의 테이블

색인코드	색인명	색인수	태그
BI	전기형식	1	008 [34]--1
CA	청구기호	2	090 [abc] 092 [a]
CLAS	분류기호	3	090 [a] 092 [a] [기사색인]ROOT-분류기호
CN	CODEN	1	030 [a]
CO	관련기관명	1	245 [ab]
CP	회의 간행물 여부	1	008 [29]--1
CT	내용형식	1	008 [24]--2
DDC	듀이십진분류기호	1	082 [a]
DT	날짜유형	1	008[6]--1
FS	기념논문집	2	008 [30]--1 010 [a]
GP	정부기관명부호	1	008 [38]--2
IG	인기(고서)	1	590 [b]

ISBN	ISBN	2	020 [a]	021 [b]
ISSN	ISSN	14	022 [a]	773 [x]
			440 [s]	775 [x]
			760 [x]	776 [x]
			765 [x]	777 [x]
			767 [x]	780 [x]
			770 [x]	785 [x]
			772 [x]	787 [x]
KDC	한국십진분류기호	1	090 [a]	
LANG	언어	2	008 [35]--3	
			[기사색인]ROOT-언어	
LC	미국의회도서관청구기호	1	050 [ab]	
LCCN	미국의회도서관제어번호	1	010 [a]	
NA	미국농학도서관청구기호	1	070 [a]	
NCCN	국립중앙도서관제어번호	1	012 [a]	
NM	미국의학도서관청구기호	1	060 [a]	
PA	판본(고서)	1	250 [a]	
PB	시작페이지	1	[기사색인]형태사항-시작페이지	
PCTR	발행국	1	008 [15]--3	
PD	형태사항	1	300 [abcde]	
PE	끝페이지	1	[기사색인]형태사항-끝페이지	
PO	출판지번호	2	260 [a]	960 [b]
PRI	가격	2	020 [c]	950 [ab]
PUBY	출판년도	1	008 [7]--4	
PYB	시작출판년도	2	008 [7]--4	
			[기사색인]잡지정보-발행년도	
PYE	종료출판년도	1	008 [11]--4	
RN	보고서번호	2	027 [a]	088 [a]
SABU	사부분류	1	092 [abcd]	
SENO	총서사항 총서번호	2	440 [v]	490 [v]
UP	대학간행물부호	1	008 [26]--2	

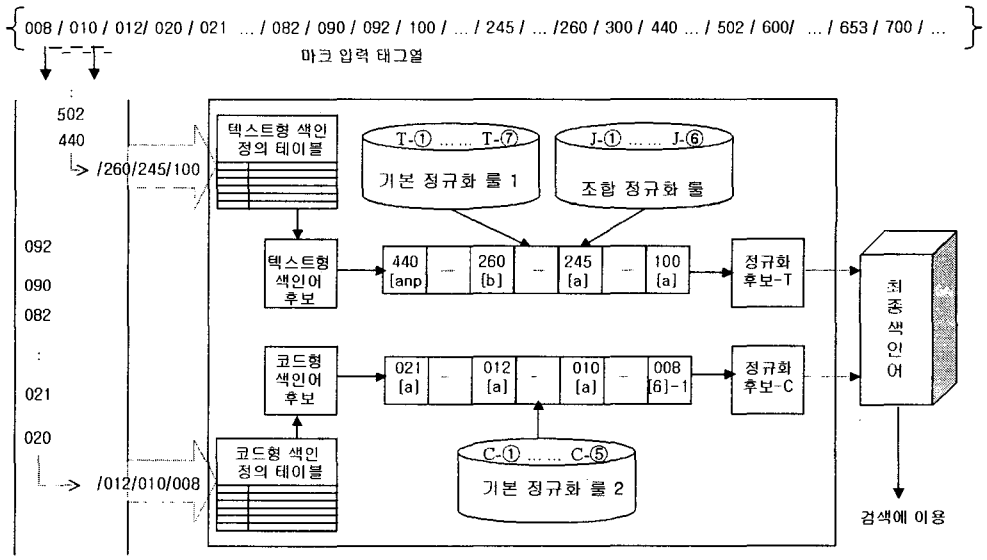
이상과 같이 서지정보검색시스템에서의 모든 색인 대상은 위에 정의한 〈표 1〉, 〈표 2〉의 색인 태그에 의거하여 서지 마크로부터 추출되고 추출된 색인 대상들은 다음의 색인어 정규화 알고리즘을 통하여 최종 색인어로 생성된다.

2. 색인어 정규화 알고리즘

색인어 정규화란 간단히 말해, 정보구축자가 기술 규칙에 의거하여 입력한 데이터 표현과 검색 시에 이용자의 예측하기 어려운 다양한 질의 입력 표현이 가능한 잘 일치될 수 있도록 색인어를 다양하게 가공하고 변환하는 것이라고 할 수 있다. 가공하고 변환한다는 것은 다양한 1byte, 2byte 특수 문자를 처리한다거나, 한자를 한글로 변환한다거나, 소문자를 대문자로 처리한다거나, 공백

처리 방법 등 여러 가지 처리 과정을 포함하고 있다.

색인어 정규화 알고리즘 적용 과정을 그림으로 그려보면 〈그림 3〉과 같은데, 먼저 정보구축자의 서지 입력이 완료되면 마크의 태그 열들이 텍스트형 색인 정의 테이블과 코드형 색인 정의 테이블을 참조하여 정의된 태그와 식별기호에 따라 색인어 후보로 추출된다. 색인어 후보 중 텍스트형의 성격을 지니는 후보들은 기본 정규화 룰 1과 조합 정규화 룰이 적용되어 정규화 된 색인어 리스트(정규화 후보-T)가 생성되고, 코드형 성격을 지니는 색인어 후보들은 기본 정규화 룰 2가 적용되어 정규화 된 색인어 리스트(정규화 후보-C)가 생성된다. 이처럼 정규화가 이루어진 정규화 후보-T와 정규화 후보-C가 모두 합쳐져 최종 색인어가 되고, 이 색인어들은 실제 검색에서 이용자 질의어와 비교하는데 활용된다.



〈그림 3〉 색인어 정규화

다음은 색인어를 가공하고 변환해주는 색인어 정규화 알고리즘에 포함된 기본 정규화 룰 1과 기본 정규화 룰 2 그리고 조합 정규화 룰의 상세한 내용을 살펴보도록 하겠다. 각 룰을 설명하는 예제에서 룰 적용 순서는 해당 예가 최종 색인어로 생성되기까지 적용된 정규화 룰의 순서이다.

1) 기본 정규화 룰 1

앞의 〈표 1〉에서 정의한 텍스트형 색인에 적용할 색인 정규화 룰은 T-①에서 T-⑦까지 7가지 기본 룰을 가지는데, 각 룰에 대한 상세 설명은 다음과 같다.

T-① 소문자를 대문자로 변환한다.

T-② Tab은 공백으로 처리한다.

T-③ KSC5601(128byte기준)로 표현할 수 없는 문자의(독일어, 프랑스어, 스페인어, 러시아 등 각 나라 표현 문자(129-155byte사이 표현))경우 문자 보정 루틴 을 수행하여 삭제한다. 예를 들어 독일어에서 움라이트 문자 ë를 e@ 표현하게 되는데 이 룰에 의해 @을 삭제하여 색인어를 생성하게 된다. 일본어(^), 프랑스어(<, >, *), 중국어(#, ')에서도 마찬가지로 적용된다.

예) TITL(서명) : 245 $aBu@rgerlicher Realismus
-> BURGERLICHER REALISMUS
(룰 적용 순서 : T-① -> T-③)

예) AUTH(저자) : 700 $aMurakami, Yukizo^ -> MURAKAMI YUKIZO
(룰 적용 순서 : T-① -> T-③ -> T-④ -> T-⑤)

예) AUTH(저자) : 100 $aHsu#, Hung-tsu -> HSU HUNG TSU
(룰 적용 순서 : T-① -> T-③ -> T-④ -> T-⑤)

예) TITL(서명) : 245 $aLa critique litte>raire franc*aise du XXe sie>cle
-> LA CRITIQUE LITTERAIRE FRANCAISE DU XXE SIECLE
-> CRITIQUE LITTERAIRE FRANCAISE DU XXE SIECLE
(룰 적용 순서 : T-① -> T-③ -> J-⑥)

T-④ 1byte 특수문자(&와 ( ) 제외한 자판의 특수문자)는 공백으로 처리한다.
=> ~ ' ! @ # $ % ^ * _ - + = | \ [ ] { } ; : ' " , < > . ? /

예) PUBN(출판사) : 260 $bE $ FN Sopn -> E FN SOPN
(룰 적용 순서 : T-① -> T-④ -> T-⑤)

예) TITL(서명) : 245 $a국어의 {-와}에 대한 연구
-> 국어의와에대한연구
(룰 적용 순서 : T-④ -> T-⑤ -> T-⑥)

2byte 특수문자는 삭제한다.
=> 「 」 『 』 《 》 " " ~ · ' ' … × 등

예) TITL(서명) : 245 $a『대공황 전후 유럽경제』
-> 대공황전후유럽경제
(룰 적용 순서 : T-④ -> T-⑥)

T-⑤ 두 개 이상의 공백은 하나의 공백으로 처리한다.

T-⑥ 2byte 문자(한글, 한자 등)의 전후 공백은 삭제한다.

T-⑦ 맨 앞뒤 공백은 삭제한다.

2) 기본 정규화 룰 2

앞의 〈표 2〉에서 정의한 코드형 색인 태그에 적용할 색인 정규화 룰은 C-①에서 C-⑤까지 5가지 기본 룰을 가지는데, 각 룰에 대한 상세 설명은 다음과 같다.

C-① 소문자는 대문자로 변환한다.

예) PCTR(발행국) : 008[15]--3 tgk -> TGK

C-② Tab은 공백으로 처리한다.

C-③ 특수문자 중 -와 / 는 삭제한다.

예) ISSN : 022 $a1228-3436 -> 12283436

C-④ 두 개 이상의 공백은 하나의 공백으로 처리한다.

C-⑤ 앞 뒤 공백은 삭제한다.

3) 조합 정규화 룰

기본 정규화는 모든 경우에 기본적으로 적용되는 정규화인 반면에, 조합 정규화는 텍스트형 정규화에만 적용되는 룰로 해당되는 조건이 존재할 경우만 적용되며, J-①에서 J-⑥까지 6개의 룰이 있는데, 이 중 적용해야 할 경우의 수가 n개라면 최대 2n개 색인까지 생성될 수 있다. 조합 정규화에서 J-③과 J-④에 해당하는 2byte 특수 문자란 모든 2byte 특수 문자가 대상이 되는 것은 아니고, 주로 그리이스 문자, 특수 유럽 문자, 일본어, 러시아 문자, 로마 숫자, 화폐 단위 등이 해당된다.

J-① 한자는 한글로 변환한다.

예) AUTH(저자) 100 $a朴美星 -> 박미성

(룰 적용 순서 : J-①)

J-② 괄호가 있을 경우 ( ) 속의 내용을 넣고 색인 하나 생성하고, ( ) 속의 내용을 빼고 색인 하나 생성한다.

예) TITL(서명) 245 $a(쉽고 실용적인)Xml 무작정 따라하기

-> XML무작정따라하기

-> 쉽고실용적인XML무작정따라하기

(룰 적용 순서 : T-① -> J-② -> T-⑥)

J-③ 2byte 특수 문자 (Α, Β, Γ 등)은 알파, 베타, 감마 등 한글로 색인 하나를 생성한다.

예) TITL(서명) : 245 $b特히 β-γ bridging에 關하여
-> 특히베타감마BRIDGING에관하여
(룰 적용 순서 : T-④ -> T-⑤ -> J-① -> J-③ -> T-⑥)

J-④ 2byte 특수 문자 (Α, Β, Γ 등)은 ALPHA, BETA, GAMMA 등 영어로 색인 하나를 생성한다.

예) TITL(서명) : 245 $b特히 β-γ bridging에 關하여
-> 특히BETAGAMMABRIDGING에관하여
(룰 적용 순서 : T-④ -> T-⑤ -> J-① -> J-④ -> T-⑥)

J-⑤ &는 AND로 치환한다.

예) TITL(서명) 260 $b이론 & 활용 --> 이론 AND 활용
(룰 적용 순서 : J-⑤)

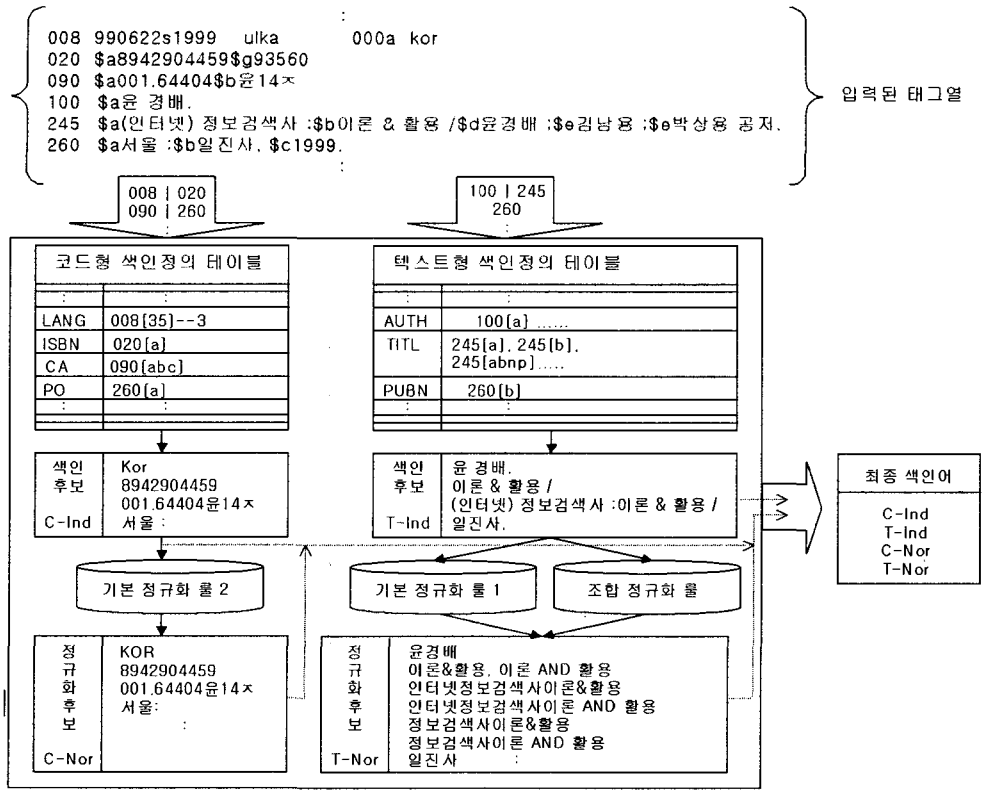
J-⑥ 정관사, 부정관사(a, an, the, le, la 등)가 있을 경우 정관사 붙여서 색인 하나를 생성하고 정관사를 빼고 색인 하나를 생성한다.

예) TITL(서명) : 245 $aLa critique litte>raire franc*aise du XXe sie>cle
->LA CRITIQUE LITTERAIRE FRANCAISE DU XXE SIECLE
->CRITIQUE LITTERAIRE FRANCAISE DU XXE SIECLE
(룰 적용 순서 : T-① -> T-③ -> J-⑥)

이상으로 Ⅱ장에서는 색인 태그 정의 과정과 색인어 정규화 알고리즘 적용 과정에 대한 이론적인 설명을 하였고, 다음 Ⅲ장에서는 색인어 추출 이론에 근거하여 실제 컨텐츠 사례를 가지고 색인어 추출 예를 보이도록 하겠다.

Ⅲ. 컨텐츠의 색인어 추출 사례

실제 입력된 국내 단행본 컨텐츠를 중심으로 색인어가 추출되고 정규화 되는 전반적인 과정을 그림으로 표현해 보면 <그림 4>와 같다.



〈그림 4〉 마크로부터의 색인어 추출 흐름도

다음 〈그림 5〉에서는 실제로 입력된 마크에 대한 색인 후보 추출과 추출된 색인 후보가 정규화 될 때 다양한 룰이 적용되어 최종 색인어가 생성되는 과정을 보인다.

〈그림 5〉의 내용은 도서관 시스템에서 프로그램을 작성하여 색인되고 정규화 되는 내부 과정의 내용들을 데이터베이스로부터 추출한 것으로, 그림으로 상세하게 표현하였으므로 따로 설명하지 않아도 이해되리라 생각한다.

이상으로 Ⅱ장과 Ⅲ장에서 색인 태그 정의, 색인어 정규화 알고리즘 그리고 이를 활용한 색인어 추출 사례까지 살펴보았다. 이러한 과정을 통해 생성된 색인어가 실제 이용자 질의어와 비교되어 일치한 문헌이 검색 결과로 출력되고, 그 결과에 따라 검색시스템의 성능을 평가받게 된다.

```
[001]|  | 0000503401▲
[005]|  | 20041015134100▲
[008]|  | 990622s1999    ulka          000a kor  ▲
[020]|  | ▼a8942904459▼g93560▲
[090]|  | ▼a001.64404▼b윤14ㅈ▲
[100]|1 | ▼a윤 경배.▲
[245]|20| ▼a(인터넷) 정보검색사 :▼b이론 & 활용 /▼d윤경배 ;▼e김남용 ;▼e박상용 공저.▲
[260]|  | ▼a서울 :▼b일진사,▼c1999.▲
[300]|  | ▼a611 p. :▼b삽도 ;▼c27 cm.▲
[500]|00| ▼a정보검색사 자격인증시험대비▲
[700]|1 | ▼a김 남용▲
[700]|1 | ▼a박 상용▲
[999]|  | ▼a771▼cD01031▲
```

코드형 색인 태그: 300, 250, 090, 020, 008

텍스트형 색인 태그: 700, 700, 260, 245, 100

코드형 색인정의테이블		색인후보
1 DT	008[6]--1	s
2 PYB	008[7]--4	1999
3 PUBY	008[7]--4	1999
4 PCTR	008[15]--3	ulk
5 CP	008[29]--1	0
6 FS	008[30]--1	0
7 LANG	008[35]--3	kor
8 ISBN	020[a]	8942904459
9 CLAS	090[a]	001.64404
10 CA	090[abc]	001.64404윤14ㅈ
11 KDC	090[a]	001.64404
12 PO	260[a]	서울 :
13 PD	300[abcde]	611p. :삽도 ;27cm.

텍스트형 색인정의테이블		색인후보
1 AUTH	100[a]	윤 경배.
2 AUTH	700[a]	김 남용
3 AUTH	700[a]	박 상용
4 TITL	245[ab]	(인터넷) 정보검색사 이론 & 활용
5 TITL	245[b]	이론 & 활용
6 PUBN	260[b]	일진사,

정규화 적용 규칙	정규화 결과
1 C-①	S
2	1999
3	1999
4 C-①	ULK
5	0
6	0
7 C-①	KOR
8	8942904459
9	001.64404
10	001.64404윤14ㅈ
11	001.64404
12 C-⑤	서울:
13 C-①,C-⑤	611P.:삽도;27CM.

정규화 적용 규칙	정규화 결과
1 T-④, T-⑥	윤경배
2 T-⑥	김남용
3 T-⑥	박상용
4 T-⑥, J-②	인터넷정보검색사이론&활용
	정보검색사이론&활용
4 T-⑥, J-②, J-⑤	인터넷정보검색사이론 AND 활용
	정보검색사이론 AND 활용
5 T-⑥	이론&활용
5 J-⑤	이론 AND 활용
6 T-④, T-⑦	일진사

〈그림 5〉 마크부터의 색인 후보 추출 및 정규화 룰 적용 과정

그렇다면 이 시점에서 색인어와 검색 결과의 관계에 대해 깊이 생각해 보아야 할 것이 있다. 현재 시스템에 정의된 색인 태그가 과연 얼마나 적정한 수준의 색인인가 하는 문제이다. "분담목록시

스템을 위한 데이터 표준화(II)"[29]나 현재 개발된 대부분의 학술정보시스템처럼 거의 모든 태그와 식별자에 대해 무조건적 색인 생성이 바람직한 것인지, 아니면 어떤 적정 기준을 정하여 선별적인 색인 태그 정의와 색인 생성이 바람직한 것인가 하는 문제이다. 이 문제를 거론해야 하는 이유는 앞에서도 언급했듯이 색인정의 작업은 검색시스템의 성능 및 시간과 밀접한 관계가 있기 때문이다. 처음에 색인을 어떻게 정의하느냐에 따라 검색 결과가 달라지는데, 색인 요소가 많이 정의될수록 많은 색인어가 생성되고, 많은 색인어가 생성되면 자연히 검색된 문헌 수가 증가하여 재현율[30]은 높아진다. 반면에 검색된 문헌들 가운데 부적합한 문헌 또한 많이 포함하게 되어 정확률[31]은 떨어지고 검색 시간도 길어지게 된다. 여기서 재현율과 정확률을 고려한 최적의 색인을 정의하는 문제는 결코 쉬운 문제는 아니다. 하지만 시스템에 부담을 주지 않으면서도 이용자들의 질의에 만족을 줄만한 적합한 검색 결과가 나올 수 있는 최적 규모의 색인 생성에 대한 연구는 다각도로 이루어져야 할 것으로 본다. 이 문제는 앞으로 검색 대상이 많아지면 많아질수록 더욱 중요한 문제로 부각되리라 생각되며, 무조건적 색인 생성은 시스템의 성능 저하와 너무 많은 결과 출력으로 이용자가 원하는 근접한 검색 결과를 얻기까지 많은 시간을 소비하게 됨은 너무 당연한 결론이다. 그러므로 현재 정의된 색인 태그가 최적의 색인 태그 정의인지에 대한 검증은 다각도로 연구해 볼만 한 가치가 있는 향후 과제라 생각된다.

그래서 다음 Ⅳ장에서는 최적의 색인 정의를 위한 기초 자료 마련을 위한 한 방안으로 현재 시스템에서 정의한 색인코드 유형별로 색인어 수를 조사하고, 이를 이용자 검색 로그에 나타난 색인 항목과 비교하므로 검색에 활용되는 색인어와 검색에 활용되지 않는 불필요한 색인어를 분석해 보았다.

Ⅳ. 색인어 분석과 색인어의 검색 활용도 분석

색인어 분석과 색인어의 검색 활용도 분석을 위해 먼저 서지 레코드 2,200,488건을 대상으로 생성된 색인 유형별 색인어 수를 조사해 보았고, 다음은 이용자 웹 검색 로그 분석을 통해 생성된 유형별 색인어가 실제 검색에 어떻게 활용되고 있는지를 조사해 보았다.

1. 색인코드 유형별 색인어 수 분석

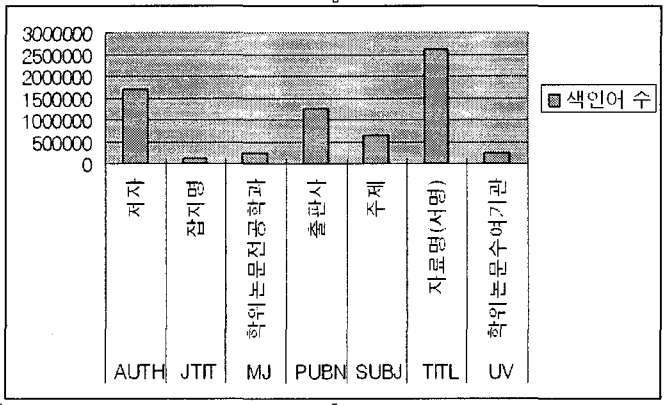
추출된 전체 색인어 수는 29,219,853건이었다. 그 중에서 텍스트형 색인어 수는 6,831,492건으로

29) 조수, 조순영, "분담목록시스템을 위한 데이터 표준화(Ⅱ)," 국립대학도서관보, 제13권(1995), pp.126-140.
30) 재현율(recall ratio)은 소장한 전체 문헌 중에서 검색된 적합 문헌의 비율을 말한다.
31) 정확률(precision ratio)은 검색된 문헌 중에서 적합 문헌의 비율을 말한다.

〈표 3〉과 같이 나타나 전체 색인어의 약 23.4%를 차지하였고, 코드형 색인어 수는 22,388,361건으로 〈표 4〉와 같이 나타나 전체 색인어의 76.6%를 차지하였다.

〈표 3〉 텍스트형 색인 코드 유형별 색인어 수

색인코드	색인명	색인어 수	백분율(%)
AUTH	저자	1714676	25.09958293
JTIT	잡지명	120657	1.766188118
MJ	학위논문전공학과	235550	3.448002281
PUBN	출판사	1251103	18.31375928
SUBJ	주제	650399	9.5205996
TITL	자료명(서명)	2630230	38.50154549
UV	학위논문수여기관	228877	3.350322301
합계		6831492	100



먼저 〈표 3〉을 보면 전체 색인어 중 자료명(서명) 색인 비율이 38.5%로 가장 높았고, 그 다음이 저자(25%), 출판사(18.3%), 주제(9.5%), 학위논문전공학과(3.45%), 학위논문수여기관(3.35%), 잡지명(1.76%) 순으로 색인들이 생성되었음을 알 수 있었다.

〈표 4〉의 코드형 색인어의 경우에는 텍스트형 색인어에 비해 색인 코드 종류도 많고 색인어 수도 3배 이상 많았다. 코드형 색인어의 유형별 색인어 생성 비율을 보면 분류기호, 언어, 시작출판년도, 형태사항, 날짜유형, 발행국, 출판지번호, 기사색인의 시작페이지, 끝페이지, 청구기호 등의 순으로 많은 색인이 생성되었음을 알 수 있었다.

2. 색인코드 유형별 검색 활용도 분석

다음은 색인어의 검색 활용도를 분석해 보기 위해 2개월간의 웹 검색 로그[32] 파일 62개를 분석

32) 경북대학교 도서관 웹서버의 자료 검색 일자별 로그(2004년 12월 1일 - 2005년 1월 31일)

해 보았다. 생성된 색인어 가운데 어떤 색인어들이 이용자 검색 항목으로 주로 활용되는가를 알아보기 위해 로그 파일에 나타난 색인 코드와 일치하는 검색 항목의 검색 건수를 조사해 보았더니 〈표 5〉과 같은 결과가 나왔다.

〈표 4〉 코드형 색인 코드 유형별 색인어 수

색인코드	색인명	색인어수	백분율
CLAS	분류기호	2200128	9.827106147
LANG	언어	2099641	9.378270254
PYB	시작출판년도	2023532	9.038321296
PD	형태사항	1642808	7.337776982
DT	날짜유형	1640884	7.329183231
PCTR	발행국	1639363	7.322389522
PO	출판지번호	1575317	7.036321239
PB	시작페이지(기사색인)	1403133	6.267243055
PE	끝페이지(기사색인)	1401742	6.261030006
CA	청구기호	1340328	5.986717831
PUBY	출판년도	814387	3.637546313
FS	기념논문집	674926	3.014628896
CO	관련기관명	572633	2.557726311
KDC	한국십진분류기호	533084	2.381076489
CP	회의 간행물여부	520333	2.324122789
ISBN	ISBN	472399	2.11002047
DDC	듀이십진분류기호	320211	1.430256552
LC	미국의회도서관청구기호	284494	1.270722765
UP	대학간행물부호	187054	0.835496623
LCCN	미국의회도서관제어번호	173578	0.775304633
PA	판본(고서)	169694	0.757956333
SABU	사부분류	162942	0.727797805
ISSN	ISSN	123448	0.551393646
PRI	가격	103020	0.460149807
CT	내용형식	89125	0.398086309
GP	정부기관명부호	50202	0.224232582
PYE	종료출판년도	48576	0.21696988
NCCN	국립중앙도서관제어번호	46847	0.209247117
BI	전기형식	37366	0.166899221
NM	미국의학도서관청구기호	22232	0.099301597
SENO	총서사항 총서번호	6889	0.030770453
CN	CODEN	3297	0.014726402
RN	보고서번호	1797	0.008026492
TDTM	석박사구분	1673	0.007472633
NA	미국농학도서관청구기호	1094	0.004886468
IG	인기(고서)	184	0.000821856
합계		22388361	100

〈표 5〉 색인 코드 유형별 검색 활용 건수

검색항목	색인 유형	12월	1월	합계	백분율
TITL(서명)	텍스트형 색인	160858	151130	311988	67.30246936
AUTH(저자)	텍스트형 색인	32636	32808	65444	14.11766736
PUBN(출판사)	텍스트형 색인	8870	8742	17612	3.799284237
ISBN(국제표준도서번호)	코드형 색인	2624	5549	8173	1.76309051
SUBJ(주제)	텍스트형 색인	1809	1979	3788	0.817152435
PUBY(출판년도)	코드형 색인	1095	2050	3145	0.678443614
LANG(언어)	코드형 색인	638	792	1430	0.308481516
KDC(한국십진분류기호)	코드형 색인	186	218	404	0.087151421
ISSN(국제표준연속간행물번호)	코드형 색인	137	239	376	0.081111224
기타(컨텐츠유형,컨텐츠번호,기관코드등)		31601	19600	51201	11.04514832
합계		240454	223107	463561	100

웹 검색 프로그램[33]에서 검색항목으로 나온 요소는 서명, 저자, 출판사, 컨텐츠유형, 한국십진분류기호, 청구기호, 주제, 등록번호, 목차, 컨텐츠번호, 분류기호, 별치기호, ISBN, ISSN으로 14가지인데, 청구기호와 등록번호, 목차, 분류기호, 별치기호 5가지 유형에 대해서는 로그 기록이 누락되어 분석할 수 없었고 검색 화면의 제한 항목에 사용되면서 색인어로 생성된 출판 년도, 언어는 분석 대상에 포함시켰다. 그리고 색인과는 무관하지만 검색 화면의 제한 항목으로 있는 컨텐츠유형과 컨텐츠번호, 기관코드 등은 기타에 포함시켜 분석하였다. 그러므로 〈표 5〉에서는 색인 코드 9종류와 기타로 나누어 검색 활용 건수를 나타내었다. 〈표 5〉의 검색 항목별 건수와 〈표 3〉, 〈표 4〉의 색칠된 부분의 색인 생성 비율과 비교하여 살펴보면 텍스트형 색인어 중 서명, 저자, 출판사, 주제의 검색 활용 순위가 비교적 상위에 있음을 알 수 있는데, 이 순위는 색인어 생성 순위와도 같음을 알 수 있다. 이것은 텍스트형 색인 정의가 활용도와도 근접하게 잘 정의되어 있다고 판단 할 수 있다. 반면에 코드형 색인어의 경우 전체 생성된 색인어의 76.6%를 차지하고는 있으나 검색 활용 순위도 비교적 낮고, 활용 순위가 색인어 생성 순위와도 무관하게 나타났다. 이는 코드형 색인 정의에 대해서는 어떤 형식이던 색인 정의의 최적화가 필요하다는 결론을 내릴 수 있었다.

끝으로 본 연구를 통해 서지정보검색시스템의 색인 정의를 최적화하고 검색 활용도를 높이기 위한 향후 연구 과제로 다음 몇 가지 내용을 제안하고자 한다. 첫째, 많은 비중을 차지하고 있는 코드형 색인 정의에서 실제 검색에 활용되지 않는 색인은 과감하게 제거하여 시스템 내의 색인어 수를 줄이고, 시스템 검색 속도를 향상시키자 것이다. 둘째, 이용자들이 서명 검색을 가장 많이 활용하고 있는 것으로 나타났는데, 서명 검색 시에 서명에 나타난 단어의 내용 일치 검색은 기본으로 지원하되, 서명에 나타난 단어들의 동의어나 유의어에 대한 서명 색인 시소러스를 구축하여 이용하면 좀 더 효과적인 그리고 어느 정도 의미가 부여된 서명 검색이 가능하리라 생각된다. 셋째, 현재 컨텐츠 유형별 구분 없이 거의 동일한 색인 정의를 단행본, 연속간행물, 비도서류, E-Journal, E-Book과 같은 전자자료 등 몇 개의 그룹으로 분리하여 컨텐츠 유형별 특징에 맞는 색인 정의 방법을 찾는다면 좀 더 효율적인 검색시스템을 구축할 수 있을 것으로 판단되었다.

V. 결 론

지금까지 경북대학교 도서관 학술정보 시스템 사례를 중심으로 서지정보검색시스템에서는 색인 태그가 어떻게 정의되어 색인 대상이 추출되는지, 추출된 색인 대상들은 어떻게 정규화 되어 색인어로 생성되는가에 대한 이론을 고찰하고, 그 이론에 근거하여 컨텐츠 사례로부터 색인어가 실제

33) 경북대학교 도서관 웹 검색시스템, 〈http://155.230.44.36/dliweb/pagehub.aspx?menuid=131〉 [cited 2005. 1. 31].

생성되는 과정을 살펴보았다. 그리고 색인 정의에 따라 추출된 색인코드 유형별로 생성된 색인어 수와 이용자 검색 로그에 나타난 색인 항목을 비교 분석함으로 어떤 유형의 색인어가 주로 검색에 활용되는지 그리고 활용되지 않는 색인어는 어떤 것이 있는 지를 분석해 보았다. 본 연구의 결과로 정보구축자는 색인 정의의 중요성을 인식하고 이용자 검색 활용에 적합한 최적의 색인을 구축하려고 노력해야 할 것을 강조하였다. 그리고 색인 정의의 문제점으로 발견된 활용되지 않는 색인 정의에 대해서는 과감하게 삭제하는 것이 바람직하고 앞으로 구축할 정보량이 증가할수록 최적의 색인 정의가 검색시스템 성능 향상의 주요 열쇠가 됨을 강조하였다. 그리고 끝으로 서지정보검색시스템의 색인 정의를 최적화하고 검색 활용도를 높이기 위한 세 가지 방안도 제시하였다.

향후 연구과제로 본 연구에서 다룬 색인 정의 및 생성 방법에 대한 기본 이론을 기초로 생성된 색인과 로그에 나타난 색인 항목의 비교 분석을 통한 검색 활용 결과를 토대로 하여 컨텐츠 유형별로 좀 더 구체적인 분석을 하거나 이용자 피드백 또는 다양한 시뮬레이션을 통해 컨텐츠 유형별 최적 색인 태그 기준을 마련하는 것이다. 그리고 서명에 나타난 단어들의 동의어나 유의어에 대한 서명 색인 시소러스를 구축하여 활용도가 높은 서명 검색의 효율을 향상시키는 것이다.

〈참고문헌은 각주로 대신함〉