KWIC索引과 Descriptor索引의 檢索 效率性

A Study on the Retrieval Effectiveness of KWIC Index versus Descriptor Index

최 상 기*

초 록

본 연구는 자동색인법에 의해 작성된 KWIC색인과 수작업색인 방법에 의해 작성된 Descriptor색인의 검색효율성을 비교하는데 그 목적이 있다. 실험의 절차와 방법은 먼저 281건의 원자력분야의 연속간행물의 논문기사를 표본으로 하여 KWIC색인과 Descriptor색인방법에 의해 색인한 다음 컴퓨터에 각각의 서지및 색인파일을 만들어 놓고 10건의 이용자 질문서를 근거로 검색을 수행하는 것으로 이루어졌다. 실험결과, KWIC색인과 Descriptor색인의 평균 재현율은 각각 54.89%와 64.42%로 나타났다.

ABSTRACT

The purpose of this study is to compare the retrieval effectiveness of KWIC index by automatic indexing method with Descriptor index by manual indexing method. The number of documents and requests used in this experimental study are 281 journal articles and 10 user requests in the area of nuclear engineering. The results of experiment show an average recall ratio of 54.89% for KWIC index and 64.42% for Descriptor index.

緖 論

과학 및 산업 기술의 급속한 발달과 연구 개발 업무에 종사하는 인구의 급증으로 인하여, 과학 기술 지식의 양과 출판된 情報 資料의 발생량은 매년 가속적으로 증가하는 추세에 있다. 특히 학술 연구 분야에서 이용되는 다양한 정보 자료 가운데서 그 발생량과 이용도 면에서 볼 때 가장 큰 비중을 차지하고 있는 것은 連續刊行物이다.[2] 또한 세계 各國의 情報 利用者가

* 국방품질검사소
접수일자 : 1985.4.24.

항상 접하고 있는 2次文獻이나 데이터 베이스 (Data Base) 도 情報源으로 연속간행문의 論文 記事를 가장 많이 수록하고 있다. 따라서 국내의 경우, 특히 과학기술 분야의 專門 圖書館들은 막대한 經費를 들여 외국에서 발간되는 연속 간행물들을 구독하고 있는 실정이다.

이와 같이 귀중한 연속간행물의 논문기사를 이용자로 하여금 보다 효율적으로 活用토록 하기 위해서는 재래식 情報 시스템을 개량하거나 또는 새로운 전산화된 情報 시스템으로 변환시키는 것이 요구된다.

일반적으로 索引은 檢索의 접근점을 제공하는 기능을 가지고 있으므로 索引作成은 情報 檢索 과정에서 핵심적 위치를 차지하고 있다. 따라서 어떤 索引語를 선택하고 색인 작성을 어떻게 하느냐 하는 문제는 檢索 시스템의 효율에 큰 영향을 미치는 요인이므로 시스템 設計 전에 반드시 정해져야 한다. 현재까지 기존 情報 檢索 시스템에서 가장 많이 사용되어온 索引 方法은 Descriptor 索引法이다. 이 색인법은 主題 專門家가 정보 자료의 주제를 분석한 후, 키워드를 디소러스를 사용하여 조정하는 절차를 거쳐 색인 작성이 이루어 지며 통제 언어가 색인어로 선정되는 통제 언어 시스템이다. Descriptor 색인법의 장점으로는 검색의 효율성이 높다는 점을 들 수 있으나 색인 작성을 위해 많은 경비와 시간이 소요된다는 단점이 있다.

반면 자동 색인법의 하나인 KWIC (Key-Word-in-Context) 색인법은 컴퓨터가 논문기사의 표제에서 키워드를 자동적으로 추출하여 색인을 작성하는 색인법이며 자연 언어가 색인어로 선정되는 자연 언어 시스템이다. 이 색인법의 장점으로는 색인 작업의 신속성과 경제성을 들 수 있다.

일반적으로 검색 시스템을 평가하는데 있어서 경제성과 신속성못지 않게 중요한 요소는 는 검색의 효율성이다. 따라서 本稿는 과학 기술 분야의 정보 자료를 색인하는 과정에서 만약 자연언어 시스템인 KWIC 색인법과 통제언어 시스템인 Descriptor 색인법이 검색의 효율면에서 커다란 차이가 없다면 경제성과 신속성을 고려하여 KWIC 색인법을 적용하는 것이 바람직하다는 가정 아래 이 두 색인법의 검색 효율성을 살피는데 그 목적을 두었다.

과학 기술 분야에서는 자연언어 시스템이 통제언어 시스템만큼의 검색 효율을 기대할 수 있다는 사실이 여러 실험 결과[1]를 통해 입증되었다. 한편 KWIC 색인과 Descriptor 색인의 비교 연구에 관한 선행 연구 문헌은 거의 없고 1968년에 Rosenberg와 Blocher가 발표한 보고서[2] 단 1편 뿐이며 이 보고서는 KWIC 색인과 Descriptor 색인 용어의 적합성(Relevance)을 측정하는 방법을 다루었다.

실험의 범위에 있어서는 검색의 효율성을 측정하는 방법 가운데서 가장 대표적인 재현율에 의한 측정 방법을 택하였다. 실험의 방법으로는

---

1) ①C. Cleverden, "The Cranfield tests on index language devices," Aslib proceedings, Vol.19, June 1967.

② G. Salton. "A comparison between manual and automatic indexing method," American documentation, Vol. 20, Jan. 1969, p. 61-71.

③ W.A. Van der Meulen and P.J.F.C Janssen, "Automatic versus manual indexing," Information Processing & Management. Vol. 13, 1977, p. 13-21.

2) Kenyan C. Rosenberg and Charles L.M Blocher, "A comparison of the relevance of Key-Word-in-Context versus Descriptor indexing terms," American Documentation, Vol. 19, Jan. 1968, p. 27-29.

먼저 原子力工學관계의 연속간행물 논문 기사를 표본으로 선정하고, 실험에 필요한 데이터 베이스를 설계하였으며 2가지 색인법에 의한 색인 작업을 통하여 각각의 색인을 작성하였다. 그 다음 이용자 질문서(2장 3항 참조)를 검토한 뒤에 작성한 탐색식을 통해 탐색한 측정결과를 분석·평가하였다.

I. KWIC索引法과 Descriptor索引法

A. KWIC索引

1. 發生 및 原理

IBM社에 근무하던 Hans Peter Luhn 은 1958년에 컴퓨터에 의해 생산되는 自動順列標題索引인 KWIC索引을 개발하였다. Marguerite Fischer는 1959년에 Luhn이 발표한 "Key-word-in-Context Index for Technical Literature (KWIC Index)"를 KWIC색인에 관한 古典的인 論文(classical paper)이라 하였고 이 논문은 표제에 근거를 두고 기계에 의해 생산되는 순열표제색인에 관한 개념과 계획을 소개하였다[3]고 하였다. 그리고 그녀는 이보다 조금 더 일찌기 1958년의 과학 정보에 관한 국제회의(International conference on Scientific Information)에서 Luhn과 Ohlman은 각자가 독자적으로 개발한 기계에 의해 생산되는 순열색인의 사본들을 배포하였다[4]고 밝히고 있다.

한편 1960년에 출판된 Law Library Journal의 한 비평란에는 KWIC색인이 미국화학회에서 발간하는 Chemical Titles를 작성하는데 사용되었다고 記述[5]되어 있으나 실제로 KWIC색인은 1961년에 Chemical Titles에서 처음 사용되었다.

KWIC색인은 키워드를 索引對象標題의 문맥 사이에 두고 그 左右에 표제로 연결시키는 것으로 문장 형식으로 된 논문의 표제에서 단어를 추출하여 키워드로 추출할 수 없는 단어 즉 시스템에 Stopword로 정해진 단어를 제외한 用語들이 키워드로 나타나게 한 것이다. Stopword들은 索引시스템 製作 전에 선정되며 그 목록은 컴퓨터에 입력되어야 한다. 그리고 이 Stopword들은 영어에서는 보통 전치사, 관사 및 접속사 등과 기타 키워드로 될 수 없는 단어들이 포함되는데, Stopword List는 KWIC색인을 만드는 제작자에 따라 다르게 작성된다.

KWIC색인의 일반적인 작성 원리는 다음과 같다. 표제와 Stopword List를 대조하여 키워드가 선정되면 각 키워드를 알파벳 순서로 배열하여 문맥의 정해진 위치에 고정시키고, 키워드의 左右에 Stopword를 포함한 다른 단어들을 문장형으로 나열한다. 그러나 KWIC색인의 出力時 컴퓨터의 인쇄기(Line Printer)는 한 줄에 인쇄할 수 있는 길이가 제한되어 있기 때문에 긴 표제는 표제 전체를 모두 한 줄에 인쇄할 수 없다. 이런 경우에는 키워드의 右側部分이 넘게 되면 左側으로 이동하여 나머지 부분을 인쇄하고 이동해서도 초과되는 부분은 생략된다. 이와 같이 표제가 키워드를 중심으로 해서 左右側으로 계속되거나 생략될 때에 사용하는 기호는 시스템에 따라 다르나, 기존 시스템에서는 보통 ·, 二, +, /, *, 등의 기호를 선택해서 사용한다.

---

3) Marguerite Fisher, "KWIC Index Concept: A Retrospective View," American Documentation. Vol. 17, 1966, p. 57.

4) loc. cit.

5) loc. cit.

도표 1 KWIC索引의 作成 例

B		A	B
Retrieval Systems =	+	Design	and Evaluation of Information
Systems = + Design and		Evaluation	of Information Retrieval
+ Design of Information		Information	Retrieval Systems =
Evaluation of Information		Retrieval	Systems = + Design and
Information Retrieval		Systems	= + Design and Evaluation of

A. 키워드
B. 문 맥

예를 들어 "Design and Evaluation of Information Retrieval Systems" 이라는 표제를 KWIC색인으로 작성한다면 보통 Stopword로 정해지는 전치사와 접속사는 키워드에서 제외되므로 여기에서 키워드는 Design, Evaluation, Information, Rotrieval, Systems이 추출되며 이 각각의 키워드 좌우에 표제가 인쇄된다. 예를 든 표제를 간단하게 표시하면 도표 1과 같다.

2. 構成 및 特徵

Fisher에 의하면 KWIC색인의 구성 요소는 색인어(Index Word), 문맥 (Context) 및 참조코우드(Code)로 이루어진다[6]고 한다. 이 가운데 색인어와 문맥은 대개의 경우 문헌의 표제에서 추출된다. 여기에서의 색인어란 보통의 키워드로서 KWIC색인에서 문맥의 중심부에 위치하는 용어이고, 그것은 검색의 접근점이 되며 KWIC색인의 요소중 가장 큰 비중을 차지한다. 문맥은 키워드의 좌우에 위치하며 이용자로 하여금 키워드가 포함된 표제의 전체 내용을 쉽게 이해하도록 해주는 기능을 가지고 있다. 그리고 참조코우드는 보통 긴 표제명이 잘릴 경우 서지사항에 관한 정보를 제공하기 위해 사용되며 잡지명코우드, 券號, 面數로 구성되거나 일련번호 혹은 분류번호로 구성되는 경우도 있다. 코우드를 작성하는 방법으로 이와 같은 것 외에 Luhn이 만든 코우드가 있는데 이 코우드는 최초로 만들어졌고 가장 널리 사용되는 코우드이며 다음과 같은 요소로 구성되었다.

1. 著者의 姓이나 發行機關
2. 出版年度
3. 文獻의 標題[7]

Luhn의 코우드는 전체가 11개의 文字로 구성된다. 즉, 처음 6개는 저자나 발행 기관의 이름에서 따오고, 다음 2개는 출판년도의 끝에 있는 2字에서 따오며, 마지막 3개는 표제로 부터 취한다.[8] 이 Luhn의 코우드를 최초로 사용한 索引誌는 Chemical Abstracts 이며 1962년 부터 사용하였다.

KWIC색인의 體制는 한 줄에 키워드, 문맥 및 참조코우드로 구성되며 각 구성 요소의 작성 방법과 배열은 색인에 따라 제각기 다르며, 한 줄의 전체 자리수는 보통 60자와 100자로 된 것이 많이 사용된다.

KWIC색인은 원래 신속성 (timeleness)의

6) ibid. p. 63.
7) loc. cit.
8) ibid. p. 67

문제를 극복하기 위해서 만들어 졌다.[8] 즉 이 색인은 최신정보주지(Current Awareness)의 도구로 생각되었고, 주로 그와 같은 용도로 사용되고 있다. 따라서 KWIC색인은 처음에는 Biological Abstracts와 같은 초록지나 색인지의 주제별 색인으로 사용되었다.

John H.Veyett와 A.Resnick은 각각 KWIC 색인을 정보 검색의 유형(Pattern)으로 적용시켰고 정보 검색의 측면을 강조하면서 지지하였다.[9]

B. Descriptor 索引

Librarian's Glossary에 의하면 情報 檢索에서 Descriptor는 한 주제를 식별하기 위해 사용되는 기초용어(elementary term)이며, 하나의 주제로 사용되는 單一語나 句로 설명되어 있다.[10] Campbell은 어떤 하나의 개념을 위해 색인된 문헌 아래에 있는 기호집합(collection of symbols)을 Descriptor[11]라고 하였고, 색인에 사용되거나 혹은 사용될 수 있는 Descriptor의 세트는 Descriptor 言語로 일컬어진다[11]라고 한데 반해 Jester는 Descriptor와 Uniterm 사이에는 차가 있을 수 없다[11]고 하였다.

원래 Descriptor의 유래는 Moores가 개개 용어들의 범위(scope)가 정의된 Descriptor라고 하는 매우 적은 어휘들의 사용을 주창하고, 그리고 정보검색이라는 용어를 소개하는 동시에 정보검색에 Descriptor라는 용어를 적용하면서 부터 시작된 것이다.[12]

Soergel은 색인언어와 디소러스에 관한 그의 저서에서 Descriptor의 同義語는 Keyword, Clueword, 색인용어(Index Term) 및 또 다른 용어들이 있고 主題名標目은 특정한(specific) 형태의 Descriptor[13] 라고 하였다. 이와 같이 Descriptor는 알파벳순 주제목록이나 인쇄색인(printed index)에서도 이용된다. 그러나 몇몇 사람들은 Descriptor를 조합색인법(Combination Indexing)을 사용한 정보 축적 및 검색시스템에만 관련지어 사용하고 있는데 이러한 사용법에서 Descriptor는 단일개념(elemental concepts)을 표현하는 반면에 주제명표목은 복합개념(compound concepts)을 표현하고 있는 것이다.

Descriptor와 디소러스의 관계 및 Descriptor와 색인과의 관계는 다음과 같이 記述될 수 있다. 디소러스는 현대의 정보검색시스템에서 사용되는 중요한 어휘조정(Vocabulary Control)의 도구이고 색인언어나 분류표와 동일한 개념을 가지고 있는 System Vocabulary로 구성되며, Descriptor는 디소러스의 안에 있는 색인용어이다. 그리고 색인언어는 Descriptor List나 System Vocabulary, 혹은 Lexicon을 포함한다.[14] 따라서 이상의 내용을 종합하면 Descriptor는 색인언어인 것이다.

本稿의 실험 대상의 하나인 Descriptor 索

---

9) ibid. p. 57.

10) L.M. Harrod, The Librarian's Glossary: A reference book, 4th ed. Worcester, Andre Deutch, 1977, p.269.

11) Alan Gilchrist, The thesaurus in Retrieval, London, Aslib, 1971, p. 8.

12) F.W. Lancaster, Advances in Librarianship, Vol. 7, 1977, p. 4

13) Dagobert Soergel, Indexing Language and Thesaurus; Construction and Maintenance. Los Angeles, Melvile Publishing Company, 1974, p. 31-34.

14) ibid. p. 27-28.

International Nuclear Information System

引法은 索引 作成時 디스크립스에 있는 Descriptor를 색인언어로 사용하고, 검색 단계에서도 Descriptor 색인법은 색인작성 및 검색 단계에서 어휘의 조정이 행해지는 후조합색인법에 속하며 이것은 Uniterm 방법에 의해 직접 영향을 받았다. 그리고 Descriptor 색인법은 색인 작업에 어떠한 일관성이 없는 Uniterm 의 단점을 보완하기 위해서 사용되었고, 궁극적으로 주제색인이 후조합방식에서 완벽한 어휘조정에 대한 필요성의 인식 및 정보검색 디스크립스의 출현과 더불어 사용된 색인법인 것이다.

Ⅱ. 實 驗

A. 實驗環境

1. 人員構成

實驗에 참여한 인원은 본 연구의 연구자 외에 韓國에너지硏究所 技術情報室에 근무하는 INIS[15] 주제전문가 3명, 電子計算室에 근무하는 프로그래머 1명과 키펀치 2명 및 이용자 그룹인 각 연구실의 연구원들로 구성되었다. 이들이 수행한 각각의 업무를 기술하면 다음과 같다. 첫째, 연구자는 전체 실험의 계획과 설계 탐색 질문식의 작성, 탐색 및 결과 분석의 업무를 수행하였고, 둘째, INIS 주제전문가들은 원자력관계 문헌의 주제분석, 색인작성 및 검색에 많은 경험을 가진 사람들로서 본 실험에서는 Descriptor 색인작성, 이용자 질문서의 주제분석과 색인용어의 변환 및 탐색 질문식 작성의 업무를 수행하였고, 세째, 프로그래머는 실험에 필요한 入力, 出力 프로그램 및 檢索 프로그램을 작성하였고, 네째, 키펀치들은 入力 데이터와 모든 프로그램을 穿孔하였으며, 다섯째, 이용자 그룹인 연구원들은 기술정보실에서 실행하고 있는 INIS SDI 검색 써비스의 신청자들로 구성되었고, 이들은 본 실험에서는 검색된 문헌의 적합성 평가의 업무를 수행하였다.

2. 標本文獻과 入力

a. 標本文獻

본 연구의 실험을 위해서 선정한 標本文獻의 주제는 原子力工學이고, 표본 문헌의 주제를 원자력공학으로 선택한 이유는 다음과 같다. 첫째, 한국에너지연구소의 主 硏究 분야는 原子力工學이고, 이 주제의 자료가 다른 주제에 비해 이용율이 높기 때문이었으며, 둘째, Descriptor 索引 作業과 探索에 필요한 이용자 질문서 및 검색 결과의 적합성 평가에 관련된 문제를 고려할 때 동일한 주제 분야에 종사하는 주제전문가와 이용자 그룹이 필요한데 이와 같은 문제점들은 시간적 공간적으로 가까이 있는 본 연구소의 원자력공학 분야의 주제전문가와 연구원들의 협력을 통해 해결할 수 있었기 때문이었고, 세째, 실험에 필요한 原文獻의 획득이 용이했기 때문이었다.

수집된 표본문헌은 본 연구소에 소장된 연속간행물 중에서 이용율이 높은 자료를 정하여 조사한 결과, 국내 자료보다 이용율이 훨씬 높은 외국의 자료가 선정되었다. 자료의 이용율 순위는 1982년에 기술정보실이 연구소에 제출한 "1981年度 運營報告書 原子力 情報 資料運營"[16]에 나타난 학술잡지의 이용 빈도수와 신착학술잡지 목차 제공 신청 순위표를 참고로 하였다. 그리고 標本對象刊行物의 種數는 10

15) INIS 는 국제원자력정보시스템 (International Nuclear Information System) 의 약자이다.

16) 韓國에너지硏究所, 1981年度 運營報告書 原子力情報資料運營. 서울, 同硏究所, 1982.

표 1. 데이터入力項目 및 자리수

항 목	문서번호	표 제	저자명	간행물명	권	호	페이지	출판월	출판년	키 워 드 (13)
자리수	3	150	50	50	3	3	9	4	4	30

種으로 이용율 순위가 높은 자료를 標本對象으로 하였고, 최근 1980년과 1981년도에 간행된 논문 기사를 표본으로 하였다.

그리고 본 실험을 위해 입력된 표본문헌의 數는 281件 이었다. 실제로 檢索效率性 測定에 관한 실험에서 사용되는 표본의 數는 일정하게 정해져 있지 않다. 일반적으로 실험에서 사용된 표본의 수가 많을수록 그 실험 결과에 대한 일반성과 객관적 타당성을 보다 더 입증할 수 있지만 가능한 한 최소의 시간과 비용으로 실험에 적절한 표본수를 정하는 것이 효과적인 것이다. 본 연구에서는 연구의 기간과 Descriptor색인 작성 및 檢索 結果의 適合性評價단계에서 소요되는 인력과 시간을 고려하여 색인 작성한 결과, 색인작성자 1명이 하루에 색인할 수 있는 평균 양은 2건 이었고 색인 작성 기간으로 정한 4개월 동안 3명의 주제전문가가 각자 담당하고 있는 주 업무 이외의 시간에 색인한 양은 281건 이었다. 그리하여 281건을 본 실험의 표본문헌으로 하였으며, 실제의 검색 실험[17]에서 사용했던 표본의 수로 100-200건 정도를 사용한 예도 있었으므로 281건의 표본수는 타당성이 있는 것이다.

b. 標本文獻의 入力레코드

入力레코드의 항목은 실험의 목적과 결과에 큰 영향을 미치는 요소가 아니므로 기존 데이터베이스의 레코드의 구조를 참고로 하여 연속간행물 논문 기사의 검색에 필요한 書誌事項만 포함시켰다. 그리고 入力레코드의 배열순서는 "學術情報 媒體의 標準化에 관한 指針"[18] 을 기준으로 해서 정하였다. 레코드의 항목은 고정장필드로 정하였고 항목 내용과 자리수는 표 1과 같다.

표 1의 키워드 항목에서 Descriptor 색인의 수는 13개까지 입력할 수 있게 하였다. 일반적으로 검색시스템에서 文獻當 색인어의 수는 12-13개 정도로 하는 것이 적절한 수준이고 본 연구소의 기술보고서 검색시스템[19] 에서는 색인어의 수를 13개까지 입력할 수 있게 하여 사용하고 있다.

c. KWIC索引의 키워드 抽出

KWIC색인은 입력된 문헌의 표제에서 stop word를 제거하여 키워드를 추출하는 과정을 거쳐 작성되기 때문에 KWIC색인 작성을 위해 가장 먼저 해야 할 일은 Stopword List를 작성하는 것이다.

17) Charles P. Bourne, "Evaluation of Indexing System," Annual Review of Information Science & Technology, Vol. 1, 1966, p. 171-190.

18) 鄭勝燮 編譯, 學術情報 媒體의 標準化에 관한 指針, 서울, 韓國圖書館協會, 1978.

19) 韓國에너지 硏究所, 1981年度 硏究報告書 原子力 關係技術報告書 檢索시스템, 서울, 同硏究所, 1982.

본 실험에서 사용된 KWIC색인의Stopword List는 다음과 같은 방법으로 만들었다. 標本文獻의 주제는 原子力工學이지만 이 분야에서 사용되는 KWIC 색인의 stopword List가 없기 때문에 현재 과학 기술 분야에서 KWIC 색인을 사용하고 있는 Chemical Titles 의 KWIC색인 stopword List에 있는 용어 중에서 INIS 디소러스[20]에 Descriptor로 나타난 용어는 본 실험의 KWIC색인 Stopword List에서 제외시켰다. 이와 같은 방법을 통해 작성한 본 실험의 KWIC색인 Stopword 의 수는 1,296개이다.

d. Descriptor 索引用語 抽出

Descriptor색인 작성을 위한 Descriptor 색인 용어 추출 작업은 먼저 문헌의 주제를 분석하여 키워드를 뽑아내고, 디소러스를 참고로 하여 다시 키워드를 색인용어를 변환하는 과정으로 이루어 지는데, 이 과정에서 키워드를 Descriptor로 변환하는 작업이 중요하다.

본 실험에서 사용된 색인 작성의 도구는 원자력 분야의 대표적 디소러스인 INIS 디소러스로써 1982년에 출판된 21판이었으며, 이 책에 수록된 Descriptor의 수는 16,211개이다.

Descriptor색인 작업은 주제전문가 1명이 1건의 논문 기사를 주제 분석부터 색인 작성까지 일괄적으로 처리하였고, 각 논문기사에 대한 주제 분석 부분은 표제와 저자 초록의 내용만을 한정하였다.

c. 入力

Descriptor색인과 KWIC색인을 위한 입력 작업은 作業用紙(Worksheet)한 장으로 처리되었다. 작업용지에 기재된 각 논문 기사의 서지사항과 Descriptor색인 용어는 키펀처에 의해 천공되어 입력되었다. 입력된 데이터는 데이터 서지 파일과 Descriptor색인 파일, 그리고 KWIC색인의 키워드 파일로 각각 나누어졌다.

3. 檢索에 사용된 利用者 質問書

檢索에 사용된 利用者 質問書는 본 연구소의 기술정보실에서 제공하고 있는 INIS SDI 검색 써비스에 접수된 질문서 중 원자력공학 분야의 것을 추출하였고, 추출된 件數는 10건이었다. 본 연구의 주제와 동일한 실험은 아니었지만 索引시스템 評價의 실험을 위해 추출된 질문서 건수로 가장 적게 사용한 예는 University of Texas에서 컴퓨터공학의 검색 기법에 관한 실험에서 사용한 4건이고[21] U.S.Bureau of Ships에서 再現率測定에 관한 실험을 위에서 조합색인법과 링크(links) 및 로울(role)을 사용한 색인법 비교 실험에서 추출한 10건이다.[21]

4. 하드웨어 및 소프트웨어

본 연구의 실험을 위한 데이터 入力·出力 프로그램의 작성에 사용된 컴퓨터의 하드웨어 및 소프트웨어는 다음과 같다. 하드웨어는 主記憶장치, 入出力장치, 補助記憶장치 및 穿孔機로 구성되며, 그 기능과 성능은 다음과 같다. 주기억장치는 CYBER 174-16으로 주기억 용량은 262KW (1 word = 60 bite)이고 명령의 처리 속도가 2 MIPS (Millions of Instruction for Second)며 기억 변환도가 100ns당 1word이다. 入·出力장치는 1분당 1,200매의 천공카드(Punched Card)를 판독할 수 있는 카드판독기(Card Reader)와 1분당 1,200줄을 인쇄하는 인쇄기(Line Printer)가 사용되었다. 보조기억장치로는 844-21디스

20) IAEA, INIS Thesaurus, IAEA-INIS-13 (Rev. 21), Vienna, IAEA, 1982.

21) Charles P. Bourne, op. cit., p. 177-179.

크유니트(Disk Unit)와 磁氣테이프(Magnetic Tape)를 사용하였으며, 천공기로는 IBM 029를 사용하였다. 그리고 소프트웨어는 "시스팀 2000"이란 多目的用 DBMS(Data Base Management System) 패케지(Package)를 사용하였고, "시스팀 2000"은 IBM 360/370, UNIVAC 1100, CDC 6000 및 CYBER 시리즈에서 작동되는 범용의 DBMS이다.

B. 實驗方法

본 연구의 실험 방법 가운데서 표본문헌의 선정, KWIC색인과 Descriptor색인의 색인 작성 및 데이터 입력 방법은 앞 節에서 논하였으므로 본 節에서는 探索方法, 檢索效率測定 및 結果分析에 관해서만 논하기로 한다.

1. 부울린 論理(Boolean Logic)에 의한 探索

첫째, 探索式의 작성을 위해 탐색자와 주제 전문가가 함께 이용자 질문서의 주제를 분석한 후 키워드를 추출하였다. 둘째, 주제 분석을 통해 추출한 키워드를 KWIC색인의 키워드파일과 Descriptor색인의 키워드 파일에 있는 색인 언어와 동일한 형태의 색인 언어로 변환하였다. 즉, KWIC색인의 探索式 작성을 위해서는 자연어(Natural Language)를 사용하였고, Descriptor색인의 탐색식 작성을 위해서는색인 작성시와 같이 키워드를 INIS 디소러스에 있는 Descriptor로 변환하였다. 세째, 색인 언어의 변환이 끝난 후 탐색 전략에 따라 탐색 용어를 선택하여 탐색식을 작성하였다. 그리고 탐색식 작성시 복합 주제를 포함하고 있는 문헌의 검색을 위해서 색인어의 상호 관계를 부울린 論理 방법으로 표시하였고, 효과적인 검색을 위해 색인 언어를 절단하는 방법(Term Truncation)[22]중에서 우측 절단(Right Truncation)과 좌측 절단(Left Truncation) 방법을 사용하였다. 네째, 작성된 탐색식을 통해 각각의 색인 파일에서 탐색을 수행한 후 검색된 문헌의 적합성을 평가하였다. 이 과정에서 탐색 작업은 컴퓨터와 연결된 CRT 터미널(Cathode Ray Tube Terminal)에서 수행하였고 1件의 탐색식에 대하여 3回의 탐색을 반복하였다. 그리고 탐색 작업에서 사용한 명령어는 SEARCH, FIND(A, B), REVIEW, COMBINE, DISPLAY 및 END등이고, 실제로는 각 命令語의 약자인 SE, FIN(A, B), REV, COM, DISP 등을 사용하였으며, 각각의 기능은 다음과 같다.

(1) SEARCH : 찾고자 하는 키워드가 포함되어 있는 문헌을 찾는 명령어이고 결과는 집합 번호, 해당 문헌 건수와 그리고 찾고자하는 키워드를 다시 보여준다.

(2) FIND : 용어의 절단 방법을 사용하여 키워드들을 하나의 집합으로 만드는 기능을 가지고 있으며 FIND(A)는 우측 절단 그리고 FIND(B)는 좌측 절단을 위해 사용하였다. 결과는 SEARCH 명령과 같다.

(3) REVIEW : SEARCH, FIND 및 COMBINE 명령에 의해 하나의 집합으로 정해진 결과를 일목요연하게 보여주는 기능을 가지고 있다. 결과는 집합 번호, 해당 문헌 건수 그리

22) Term Truncation은 질문식 작성시 색인어를 기입할때 동일개념을 나타내는 용어중에서 語順, 語幹, 語尾등이 유사한 것은 그중 공통되는 부분만 기입하고 나머지 부분은 절단하여 한 용어만 기입·입력하면 탐색시 공통되는 부분과 일치하는 글자가 포함된 모든 용어가 탐색되는 방법으로, 그 종류는 non truncation, right truncation, left truncation 및 both truncation이 있다.

고 SEARCH, FIND 명령에 의한 키워드나 COMBINE 명령의 결과를 보여준다.

(4) COMBINE : SEARCH나 FIND명령에 의해 만들어진 집합들을 부울린연산자(Boolean Operator)를 사용하여 다른 집합으로 만들고자 할 때 사용하는 명령어로 연산자는 다음의 특수문자를 대신하여 AND는 '*', OR은 '+'를 사용하였다.

(5) DISPLAY : DISPLAY 명령은 검색의 결과를 보여주는 기능을 가지고 있으며, DISP 명령 다음에 집합 번호를 입력하면 포함하고 있는 문헌 번호와 서지사항들을 보여준다.

(6) END : 탐색 작업의 종료를 컴퓨터에 알리는 기능을 가지고 있다.

적합성 평가는 이용자 그룹에 의해 수행되었으며, 평가 방법은 각각의 이용자가 검색된 문헌과 전체 표본문헌을 이용자 자신의 질문서와 대조하여 그 주제에 적합한 문헌을 체크하는 것이었다. 끝으로 이용자에 의해 평가된 적합성 판정 결과를 근거로 하여 再現率산출 공식에 의한 再現率을 측정하였다.

2. 再現率에 의한 檢索效率測定

검색 효율 측정 방법 중에서 가장 많이 사용되는 것은 再現(Recall)과 精度(Precision)이다. 여기에서 再現은 검색된 적합 문헌의 量(proportion)을 의미하며, 精度는 적합 문헌이 검색된 문헌에 들어있는 量을 의미한다. 즉, 再現이란 특정한 문헌이 검색되는가, 검색되지 않는가의 측정, 혹은 원하는 문헌의 검색이 발생하는 程度를 뜻하는 반면에 精度는 보통 어떤 종류의 검색시스템에서 雜音(Signal to Noise) 비율의 측정에 언급된다. 이 둘의 관계를 볼때 再現과 精度는 서로 역관계의 성격을 띠고 있다. 부연하면 再現을 높이기 위해 탐색의 범위를 넓히면 精度는 감소하고 반대로 精度를 높이기 위해 탐색의 범위를 좁히면 再現이 감소하는 경향이 있다.

대체로 精度率(Precision Ratio) 측정은 정보 검색 과정의 출력 단계에서 소모한 시간과 비용의 간접적인 측정을 위해 사용되고 있으며 이용자가 소비한 시간의 비용 요인(cost factor)의 유형으로 간주되므로 본 연구의 목적인 색인언어와 색인법의 평가에는 부적합한 검색 효율 측정 방법이다. 그리고 탐색문헌의 수가 많지 않은 경우에 정보 이용자들은 주제 검색시 가능한 한 많은 양의 적합 문헌이 검색되기를 바라는 경향이 있는데 이것은 이용자들이 일단은 많은 양의 적합 문헌이 검색되기를 원하는 개념으로 해석될 수 있다. 따라서 이와 같은 점을 고려한 결과 본 실험에서는 검색 효율 측정 방법으로 再現率을 선정하였다.

보통 再現의 개념은 完全性(Completness)으로 나타내기도 하며 再現의 同義語로는 感度(Sensitivity)가 사용되기도 한다. 再現率의 정의는 앞에서 밝힌 바와 같이 검색된 적합 정보의 양이고, 가장 많이 사용되는 再現率의 정의는 질문서에 답하기 위해 실제로 검색된 파일 내에서의 적합 문헌의 비율이며, 再現率산출 공식은 다음과 같다.

$$\text{再現率} = \frac{\text{검색된 적합문헌의 수}}{\text{전체 적합문헌의 총수}} \times 100$$

3. 結果分析

10件의 探索式을 통해 검색한 적합 문헌의 再現率은 검색 효율 측정에서 제시한 再現率 산출공식에 의해 계산되었으며 再現率 測定 結果는 표 2와 같다. 표 2에 나타난 再現率의

표 2. 再現率 測定結果

탐색번호	KWIC 색인 A/C	KWIC 색인 재현율	디스크립터색인 A/C	디스크립터색인 재현율	차 재현율
# 1	4/8	50 %	5/8	62.5 %	12.5 %
# 2	3/6	50 %	4/6	66.7 %	16.7 %
# 3	4/7	57.2 %	3/7	42.9 %	− 14.3 %
# 4	3/5	60 %	4/5	80 %	20 %
# 5	2/5	40 %	3/5	60 %	20 %
# 6	7/11	63.6 %	9/11	81.8 %	18.2 %
# 7	5/8	62.5 %	4/8	50 %	− 12.5 %
# 8	5/9	55.6 %	7/7	77.8 %	22.2 %
# 9	3/5	60 %	3/5	60 %	0 %
# 10	4/8	50 %	5/8	62.5 %	12.5 %
합 계		548.9 %		644.2 %	95.3 %
평 균		54.89 %		64.42 %	9.53 %

A : 검색된 적합문헌 수
B : 전체 적합문헌 수

평균치는 KWIC색인의 54.89%와 Descriptor색인의 64.42%이었다. 그리고 KWIC색인과 Descriptor색인의 再現率 평균차는 9.53%를 나타냈고, 각 탐색 건당 KWIC색인과 Descriptor색인의 再現率 차이는 탐색식 1번, 7번 및 10번의 최저 12.5 %에서 8번의 최고 22.2%까지의 범위를 나타냈다. 한편, KWIC색인의 再現率 범위는 최고 40%에서 63.6%에 걸쳐 있고 그 차이는 23.6%이다. 그리고 Descriptor색인은 최저 42.9%에서 81.8%까지 걸쳐 있고 그 차이는 38.9%였다.

結 論

이상과 같이 KWIC索引의 발생, 원리, 구성 및 특징과 Descriptor索引의 개념 및 특징을 고찰하면서 각기 지니고 있는 특징속에서 과연 그 兩者의 檢索效率性이 어떤가를 문제로 삼고 1) 표본문헌의 선정, KWIC 索引法과 Descriptor 索引法에 의한 색인 작성, 탐색 작업 및 적합성 평가의 과정을 거쳐 再現率측정 방법을 통한 兩 색인의 검색효율성을 측정한 바 2) 이 실험에서 KWIC索引과 Descriptor索引의 평균 再現率은 각각 54.89%, 64.42%로 나타났고, 兩 색인의 再現率 차이는 9.53%였음이 밝혀졌다.

따라서 본 연구는 실험 결과로 나타난 KWIC索引과 Descriptor索引의 검색효율의 차가 약 10%를 기록하였으므로 연구의 가정에 부합되는 결과를 얻었다고 결론지을 수 있다. 이번 실험에서 얻어진 바, KWIC索引은 컴퓨터가 입력 문헌의 표제에서 미리 작성되어 컴퓨터에 입력된 Stopword List를 근거로 키워드를 추출하면 색인이 자동적으로 작성되기 때문에 색인 작성에 전문 인력이 필요하지 않고 색인 작성의 소요시간이 적어진다는 것이 증명되었으며, 이와는 달리 Descriptor索引은 색인 작성자가 문헌의 주제를 분석하고 디소러스를 색인 작성의 도구로 하여 색인을 작성하므로 색인 작성을 위해 전문 인력이 필요하고 색인 작성의 소요 시간이 많다는 사실이 증명되었다. 그러므로 KWIC索引이 Descriptor索引에 비해 경제성과 신속성면에서 더욱 효율적인 것이다.

비록 본 실험의 결과, KWIC索引의 평균再現率이 Descriptor索引에 비해 약 10%가 낮았지만 再現率은 탐색시 색인작성의 망라성에 의해 크게 영향을 받으므로 탐색시 디소러스나 주제명표목 등을 이용하여 망라성이 풍부한 탐색식을 작성해서 검색한다면 再現率을 보다 더 증가시킬 수 있을 것임을 附言해 둔다.