「情報管理硏究」, vol. 23, no. 2, 1992. 6., pp. 1～13

標題와 抄錄의 索引性과 情報量 分析

Indexability and Information Quantity Analysis in Title and Abstract

金 在 洙*
(Kim, Jae Soo)

南 泳 準**
(Nam, Young Joon)

抄 錄

本 硏究에서는 自動索引의 索引語 抽出에 있어 주요한 索引源이 되는 標題와 抄錄등의 색인성을 비교 분석하였다. 결과는 표제만을 혹은 抄錄만을 索引源으로 선정할 경우에 적절한 索引語를 추출할 수가 없었으며, 標題와 抄錄을 동시에 索引源으로 선정할 경우가 좀 더 적절한 索引語를 확보할 수가 있었다.

키 워 드

自動索引, 標題 索引性, 抄錄 索引性, 情報量 分析, 索引語 抽出.

ABSTRACT

This study intends to measure the indexibility and the information quantity in title and abstract. The result of analysis was that when the source was title or abstract, result was not good. But when it was the title and abstract, the result was better.

KEYWORDS

Automatic indexing, Title indexibility, Abstract indexibility, Information quantity analysis, Indexing term extraction.

* 國防科學硏究所 責任技術員(Agency for Defence Development).
** 中央大學校 博士課程 (Graduate in Chung Ang University).

標題는 전통적으로 情報檢索에 있어서 접근점으로서의 중요한 역할을 하며, 抄錄은 그 情報가 이용자에게 필요한 정보인지를 파악하여 이용자의 reading overload를 줄여주는 또하나의 주요한 역할을 하고 있다. 또한 이들은 自動索引에 있어 候補索引語를 제공하는 情報源(information source)으로서의 중요한 역할도 수행하고 있다.

自動索引方法은 그 기법이 어떻든간에 반드시 분석대상이 되는 문장이 컴퓨터에 입력되어야 하기 때문에, 일반적으로 標題가 1차적인 분석대상이 되며, 시스템에 따라 抄錄이나 本文이 2차 분석대상이 된다.

그러나 현재 索引語自動抽出 방법은 대부분이 標題와 抄錄, 本文이 候補索引語를 추출하는 분석대상일 경우에는 각기의 특성을 고려하지 않고서 일괄적으로 하나의 분석대상물(동일한 입력문장)으로 간주하여 索引語를 추출한다.

情報의 크기는 발신자의 위치나 형태에 따라서 차이가 있다. 즉, 동일한 情報라 할지라도 그 정보가 저명한 學術誌에 게재된 경우와 일반 時事誌에 게재된 경우에 그 情報의 무게는 다를 수 밖에 없다.

標題는 저자가 자신의 사고를 최대한 압축하여 이용자에게 가장 간략하게 자신의 논문의 내용을 전달될 수 있도록 한 것이며, 그러므로 표제를 구성하고 있는 單語들은 本文에 나타난 다른 單語들보다는 그 文獻의 表意性이 훨씬 높을 것이다. 그러나 標題는 길이의 제한때문에 표의성을 가지고 있는 적절한 수의 키워드를 추출하기에는 부적절한 情報源이다. 즉, 표제는 하나의 文獻을 대표하는 索引語를 추출하기 위한 情報源으로서 정도성은 높으나 재현성이 떨어질 것이다. 이러한 재현성의 補整을 위하여 自動索引에서는 抄錄과 本文이 또하나의 情報源으로서 활용되고 있다.

本 硏究는 標題의 정도성과 재현성을 측정하여 자동색인의 情報源으로서의 가치를 평가하고, 抄錄과 本文의 가치도 측정하고자 한다. 이 결과는 앞으로 보다 효율적인 索引語를 추출할 수 있는 自動索引技法을 개발하는데 필요한 기초자료로 활용될 수 있을 것이다.

Ⅰ. 索 引 性

1. 標題의 索引性

標題는 著者가 본문의 내용을 최대한 압축하여 다른 情報와의 類別性을 보장하도록 한 것이며, 또한 이용자로 하여금 본문내용을 어느 정도 파악할 수 있도록 하는 1차 식별자 역할(aboutness)을 수행하도록 한다. 標題의 구성에는 일정한 규칙 또는 단어수, 표제형식에 제한을 두고 있지 않다. 그러므로 標題는 著者의 개인적인 성향에 따라 다양한 형태를 취하게 된다.

그러나 이러한 불특정적인 형식에도 불구하고 標題는 東西를 막론하고 情報源으로서의 그 효율성을 인정받고 있다. 일찍이 Luhn은 文獻의 내용을 標題에 나타난 單語를 이용하면 이용자는 이를 근거로 文獻의 내용을 연상하여 該當情報에 충분히 접근할 수 있다는 사고에 근거를 두고 KWIC 索引을 개발하였다.[1] 여기에서 Luhn은 하나의 명사(구)는 반드시 하나이상의 의미를 지니고 있으며, 표제에 출현한 명사(구)는 해당 논문의 내용을 가장 잘 표현한다고 주장하였다.

이러한 索引方法이 개발된 이래로 標題를 索引對象으로 한 類似標題利用索引方式이 다양하게 개발되었으며, 현재에도 폭 넓게 사용되고 있다. 국내에도 自然言語處理方式을 지향하는 몇몇 기관[2]에서 이러한 방식은 아니지만 표제를 自然言語處理主要對象으로 하여 索引語를 추출하고 있다. 이와 같이 標題를 索引의 역할을 하는 주요 분석대상으로 하는 이유는 다음과 같다.

i) 自動索引시스템構成의 편이성

標題分析시스템은 索引語 抽出에 있어 빈칸이나 특수기호를 구분자로 하여 용어를 추출하고 이를 효과적으로 배열하는 방식이어서 분석시스템이 단순하지만 抄錄과 本文과 같이 문장형태로 구성된 그 밖의 情報源을 분석대상으로 한다면 自動索引시스템構成이 그에 비례하여 복잡해질 것이다.

---

1) H. P. Luhn, "A statistical approach to mechanized encording and searching of library information ", *IBM Research Development*, vol. 1, no. 4, 1957, pp. 309~317.

2) 國會圖書館의 學位論文記事索引도 이와 같은 自然語方式을 사용하며, 佛敎放送의 데이터베이스도 標題分析에 있어서 自然語方式을 활용하고 있다. 서울대학교 圖書館에서도 索引語抽出에 있어서 標題를 활용하고 있다.

ii) 效率性 問題

抄錄이나 全文을 索引語抽出의 情報源으로 확장하여 처리한다면 시간과 경비를 고려한 효율성에 대한 의문이 제기된다. 이에 대해 Kraft는 표제에 나타난 키워드 가운데에 60% 이상이 主題語로서 자격이 있다는 연구결과를 제시하였으며,[3] 우리와 유사한 文法體系를 갖고 있는 日本에서는 齊藤孝가 표제에 출현된 단어가 정도성이 다른 어떤 곳에서 출현한 단어보다도 정도성이 높다는 것을 입증하였다.[4] 이태영도 일련의 연구를 통하여 표제의 表意性이 약 60%선이라고 결론을 내리고 있다.[5] 그러므로 標題를 가지고도 충분히 情報와 情報間에 有別性과 表意性[6]을 유지할 수 있기 때문에 표제는 索引語抽出을 위한 주요한 情報源이 될 수 있다.

그러면 標題分析方法이 이와 같은 특성을 지니고 있음에도 불구하고 보편적으로 사용되지 못하고 있는 이유는 다음과 같은 표제의 한계성에 기인한다.

i) 標題에 출현한 單語의 수가 일정하지 않다.

표제의 길이는 學問分野와 言語마다 차이가 있으며, 또한 著者의 개인적인 성향에 따라서도 많은 차이가 있다. 표제길이의 차이는 候補索引語의 갯수의 차이를 의미하며, 이로 인해 情報의 균질적인 저장과 다양한 접근점의 제공이 어렵다. 이는 표제만으로는 일정한 索引語를 확보할 수 없다는 것을 의미한다.

예를 들면,

① 소식사 연구

② 中國의 불전간행 및 目錄編纂에 관한 연구

③ 마케팅調査의 결합분석법을 활용한 참고봉사의 평가

위의 예는 모두 博士學位論文으로서 그 내용에 있어서 정보의 양은 동일하다고 간주한다. 일반적인 방법[7]으로 분석을 하면 1번 예는 2개의 단어로 절단이 되며, 2번 예는 6개의 단어로 절단이 이루어진다. 마지막으로 3번 예는 5개의 단어로 절단이 된다. 즉, 동일한 情報量을 가지고도 각 文獻이 갖게 되는 標題의 길이에 의해 후보색인어의 수는 많은 차이를 보여주게 된다. 즉, 표제가 일정한 정보량을 제공하는 情報源의 기능을 하기에는 부적격하다는 것을

---

3) D. H. Kraft, "A comparison of keyword-in-context indexing of titles with a subject heading classifecation", *AD*, vol. 15, 1964(Jan), pp. 48~52.
4) 齊藤孝, "索引作業のための自然語處理の研究", *LIS*, 1965, No. 51967, 51~72.
5) 이태영, "韓國語論文記事의 標題에 대한 分析的 考察", 「圖書館」, 1985(2), p. 7.
6) 표의성이란 문헌이용자에게 해당정보의 내용을 알 수 있도록 하는 성질을 의미한다.
7) 빈칸이나 특수기호에서 절단하여 주요어를 색인어로 선정하는 방법

의미한다.

ii) 표제의 맞춤법에 따라서도 단어의 수가 달라진다.

英美系의 표제에서는 발생하지 않으나 한글표제에서는 띄어쓰기에 따른 단어의 혼동이 있다.

예를 들면,

1 意味 情報를 이용하는 중심어 주도의 韓國語파싱

2 意味情報를 이용하는 중심어주도의 韓國語파싱

위의 예에서 한글맞춤법개정안에 따르면 위의 예문은 모두 맞춤법에 어긋나지 않는 문장이지만, 띄어쓰기 방법에 따라 1번 표제는 7개의 단어로 구성이 되는 반면에 2번 표제는 4개의 단어로 구성이 된다. 또한 2번 예에서 복합명사 '意味情報'는 수작업 색인시 단일표목으로 선정이 되지만, 빈칸을 단어와 단어간의 구분으로 간주하는 자연어 처리방식에 있어서는 1번 예는 '의미'와 '정보'라는 두개의 표목이 선정된다. 이 경우에 전자와 후자간의 정도성은 많은 차이가 있게 된다. 후자의 표목 '意味'와 '情報'는 모두 기능어에 속하기 때문에 키워드가 되지 못하지만 전자는 언어학적인 관점에서 주요한 키워드가 되기 때문에 索引語로 선정이 될 수 있다. 즉, 標題索引에서 단어절단은 특수한 기호나 빈칸을 기준으로 이루어지기 때문에 이와 같이 동일한 단어를 두고 상이한 결과가 나올 수 있다.

iii) 標題 自體의 單語의 수만으로는 색인어 선정의 정보원으로는 불충분하다.

일반적으로 표제를 구성하는 단어의 수가 국가나 國際標準案으로 제정된 규정은 없다. 그러나 情報接近點으로서의 索引語의 개수는 대부분이 10~15개 내외로 간주한다. 이에 비해 표제를 구성하고 있는 단어의 수는 英美系의 경우는 15개 내외, 日本의 경우는 8개 단어 내외로 간주하고 있다. 이 가운데서 해당정보의 내용을 표현할 수 있는 수준의 키워드는 學問 영역에 따라 차이가 있겠으나 대략 1개에서 5개 사이의 單語만이 索引語로 선정될 수 있을 것이다. 이에 비해, 데이터베이스의 구축시에 적절한 색인어(키워드)의 수는 5~10개가 바람직하기 때문에[8] 표제만을 대상으로 키워드를 선정한다는 것은 무리가 있다. 즉, 표제로부터의 정보구별은 정보의 급증과 표제의 산술적인 증가,[9] 한정된 主題語의 사용으로 인하여 적절한 수의 情報를 적출한다는 것이

8) 日本科學技術情報센터, 學術論文의 構成 및 그의 요소 SIST-1986, 동센터, 1986, p. 4.
9) 이태영, *ibid.*

어렵게 되었다. 이는 표제를 분석대상으로하는 자동색인의 단점을 의미하며 自動索引의 분석범위를 초록으로 확장해야하는 중요한 이유가 된다.

2. 抄錄의 索引性

抄錄은 源文의 내용을 요약해 놓은 문장의 집합으로 일반적으로 50 내지 500단어의 길이로 기술하여, 情報檢索에 있어서 중요한 판단역할을 하는 기능을 갖고 있다. 한글과 유사한 文法體系를 갖고 있는 日本에서는 國家標準案으로 日本規格集에서 이러한 抄錄作成時의 원칙을 다음과 같이 규정하고 있다.[10]

1 표제의 내용을 반복해서 사용하지 않는다.

1 원칙적으로 본문에 실려있는 전문용어를 그대로 사용한다.

1 그 길이는 일본어의 경우 200~400문자, 구문의 경우는 100~200문자를 표준으로 한다. 단지 짧은 記事文獻일 경우는 日本語의 경우 150~200문자, 歐文의 경우는 70~100단어를 표준으로 한다.

日本規格集에서 특징적인 것으로는 學術資料의 표제에 사용되었던 표현과 語彙는 가능한한 반복해서 사용하지 않을 것을 권고하고 있는데, 이는 이용자에게 보다 많은 정보를 제공하고자 하는데 그 이유가 있다. 즉, 표제는 표제나름대로 이용자에게는 필요할지도 모르는 정보에 대한 접근점의 역할을 수행하도록 하며, 抄錄은 그 情報를 직접 입수할 것인가를 판단하는 또 다른 접근점과 기준이 되도록 한다.[11] 이는 표제의 情報源만으로는 이용자에게 보다 다양한 정보접근점을 제공할 수 없다는 것을 의미한다.

3. 인용부의 索引性

인용부는 著者가 저자자신의 논문에 대한 학문적 배경이나 자신의 硏究方法을 유도 및 지지하는 자료를 선정하여 인용한 것이다. 특히, 인용부는 著者가 해당 피인용문을 충분히 숙지한 후 자신이 필요한 주요 부분만을 인용하는 것이기 때문에 이는 論文의 核心部分이라 할 수 있다. Kwok는 인용부에 나타난 引用文의 표제는 그 文獻의 주제를 표현할 수 있다는 것과 그 引用文標題에 출현한 단어들이 그 문헌의 내용을 대표할 수 있다는 것을 실험하였다. 그는

---

10) 日本科學技術情報센터, SIST 01-1980 抄錄作成, 동센터, 1980, pp. 2~3.

11) Raya Fidel, "Writing Abstracts for Free-Text Searching", *Journal of Documentation*, vol. 42, no. 5, 1986, pp. 15~18.

표제에 출현한 單語와 引用文의 표제와는 밀접한 연관성이 있다는 것과 抄錄에서 표현하지 못하는 내용을 引用文의 표제를 분석하여 알 수 있다고 주장하였다.[12] 그러나 아직까지 引用符는 文獻의 내용을 표현할 수 있는 표목의 역할보다는 오히려 동시인용을 판단하여 핵심저널의 선정과 일반저널의 수명을 판단하는 計量書誌學的 판단가치가 되고 있을 뿐이며 引用文은 아직까지 키워드를 추출할만한 情報源으로서의 각광을 받지 못하고 있다.

4. 本文의 索引性

索引語를 추출할 수 있는 가장 훌륭한 情報源으로서는 단연 본문을 들 수 있다. 또한 가장 적절한 색인어 추출방식으로는 해당분야 專門家가 해당 문헌의 標題나 抄錄과 같은 요약문외에 본문을 숙독한 후에 적절한 索引語를 적절한 수만큼 부여하는 것이다. 그러나 이러한 방식은 東西古今을 막론하고 그 실행에 많은 어려움이 있다. 지금까지의 自動索引語抽出硏究가 주로 짧은 記事文이나 標題와 抄錄을 대상으로 한 이유가 입력에 따른 막대한 시간과 예산을 들이는 것과 각 學問分野의 전문색인가와 새로운 분야의 학문에 능동적으로 대처할 수 있는 전문색인가를 확보할 경우 비용대 效率性(cost-effectiveness) 문제에 대한 확실성이 없기 때문이다. 특히, 自動索引의 경우에는 그 분석대상을 표제와 초록과 본문간의 분석결과를 비교할 때 效率性 側面은 抄錄을 근거로 한 분석이 原文全體를 대상으로한 것보다 유리하다는 주장[13]과 표제만 활용한 것이 초록을 대상으로 한 것보다 檢索効率이 우수하다는 주장[14]이 있다. 이는 本文이 情報源으로서 가치가 있지만 抄錄의 索引性보다는 못하고 抄錄 역시 標題에 비해 효율이 떨어진다는 것을 의미한다.

Ⅱ. 情 報 量

Fidel은 최근의 硏究에서 抄錄의 길이는 거의 일정하나, 本文의 길이와 抄錄

12) K. L. Kwok, "The use of title and cited titles as document representation for automatic classification", *IPM*, vol. 11, 1975. pp. 201~202.
13) Karen S. Jones, "Automatic Indexing, *Journal of Doc.*, vol. 30, no. 4, 1974, pp. 393~432.
14) K. L. Kwok, *ibid*, p. 206.

의 길이와는 어떠한 연관성을 갖고 있지 않다고 하였다.[15] 이는 抄錄이 갖고 있는 정보의 양이 본문의 길이에는 큰 영향을 받지 않는다는 것을 의미한다.

1. 標題와 抄錄間의 관계

표제가 가지고 있는 情報量과 抄錄이 가지고 있는 情報量에 대해서 어느 정도의 크기를 갖고 있는지와 어느 정도의 차이를 보이고 있는지에 대해서 아직까지 國內外에서는 구체적인 실험이나 硏究結果가 제시된 적이 없었다. 단지, 정영미[16]는 抄錄을 구성하는 단어를 대상으로 표제를 분석한 결과 主題語의 수는 평균 5개이며, 標題의 主題語와 抄錄의 주제어와의 일치율은 논문에 따라 차이가 있으나 대략 25~30% 정도였다고 하였다. 또한 표제어와 일치하지 않는 抄錄의 주제어 중에는 색인어로 부적합한 색인어가 상당수 있었으며, 索引語로 적합한 주제어는 대체로 표제어와 어의적인 관계가 있음을 밝히고 있다. 즉, 표제에서 나타난 주제어 중 1~2개만이 抄錄에서 主題語로 동시에 사용되었다.

2. 標題와 抄錄의 情報量

本 硏究에서 情報量의 측정을 위하여 다음과 같은 2가지 실험을 하였다. 첫째, 한글로 된 論文을 대상으로 標題와 抄錄의 효율을 측정하였으며, 둘째, 英文데이터베이스를 대상으로 각 標題와 抄錄의 檢索効率을 조사하였다.

3. 한글문헌을 대상으로 한 實驗

본 분석은 齒醫學科 博士學位論文 15편과 電算學博士學位論文 5편을 대상으로 하였다. 대상으로 삼은 논문은 저자들이 어떠한 사전주문이나 통제도 받지 않고 자신이 생각한 키워드를 추출하였으며, 저자자신이 직접 작성한 초록문을 갖고 있었다. 또한 초록의 특징으로는 모두가 통보적 초록이었으며, 한글 및 漢字, 外來語가 標題, 抄錄, 索引語에 모두 출현하고 있다. 본 실험의 의도

---

15) Raya Fidel, "The Possible effect of abstracing guidelines on retrieval performance of free-text searching", *IPM*, vol. 22, no. 4, 1986. pp. 309~326.

16) 정영미, "우리말 情報資料를 처리하는 지능형 情報檢索시스템의 設計", 「韓國情報管理學會」, 제8권 2호, 1991, p. 16.

는 실제로 표제와 초록에 출현한 단어가 저자가 추출한 索引語와 어느 정도 일치하는가를 분석하는 것이다.

본 실험에서 제시하는 모든 수치는 다음과 같은 알고리즘으로 계수된 수치이다.

i) 單語의 區分은 빈칸과 마침표, 쉼표를 구분자로 인식하였으며 品詞의 종류는 고려하지 않았다.

예) 「치아광질이탈현상유인요인에 관한 연구」라는 標題의 경우에는 3단어로 간주하였으며, 「부분적 하악 과두 절제후 늑연골 이식에 관한 연구」라는 표제는 8개의 단어로 계수하였다.

ii) 英文이나 外來語의 경우는 單語間의 띄어쓰기가 있어도 하나의 단어구는 하나의 단어로 간주하였다.

예) Fibrin sealant(著者註:齒科에서 사용되는 약제명)은 2개의 단어로 구성이 되어 있으나 한글의 複合名詞와 같은 형식이기 때문에 하나의 단어로 계수하였다.

iii) 複合名詞는 하나의 단어로 계수하였으며, 띄어쓰기가 되어 있으면 예외로 하였다.

예) '치아광질이탈현상유인요인'은 2자씩으로 6번 중첩된 형태의 複合名詞이며, 맞춤법에서는 6개의 단어로 구분이 가능하지만 저자가 하나의 단어로 기술하였으므로 하나의 단어로 계수하였다.

(1) 標題의 情報量

標題는 論文이 지니고 있는 내용을 가장 압축적으로 나타내고 있으며, 다른 어떤 필드보다도 효율적인 접근점이 될 수 있다. 또한 표제에 출현한 단어는 索引語가 될 확률이 가장 높은 구문적 특성을 갖고 있다.

本 實驗에서의 標題의 평균 단어수는 7.55개이다. 이태영의 연구에 의하면 1992년 현재, 學術論文의 標題의 단어수가 약 9개로 예측하였으나 이러한 차이점을 보이는 것은 앞에서 전제한대로 複合名詞의 한 단어 處理方式에 기인한 것으로 본다. 또한, 표제에 후보색인어가 포함되어 있는 비율은 56.3%였다.

(2) 抄錄의 情報量

抄錄이 가지고 있는 평균 단어수는 148.9개이다. 순수하게 抄錄에서 索引語가 포함되어 있는 갯수는 15.2개이다. 이 숫자는 索引語가 반복해서 출현한

것도 각기 계수하였다. 중복 출현한 것을 하나로 계수하면 索引語의 평균 빈도는 4.8개이다. 抄錄에 키워드가 포함되어 있는 비율은 약 10.2%이다.

(3) 情報量 分析

本文이 지니고 있는 情報量을 100으로 정한다면,[17] 抄錄과 標題가 갖고 있는 정보의 양은 어느 정도인가.

본 조사에 의하면 하나의 문헌이 갖는 색인어의 수는 5.75개이며, 편차는 6이였다. 이에 대해 표제에서 索引語가 출현한 갯수는 3.12개였으며, 이를 백분율로 환산하면 54.3%이다. 抄錄에서 索引語가 출현한 갯수는 24개였으며 중복된 索引語를 제외하면 4.8개였다. 이를 백분율로 환산하면 83.5%이다. 또한 표제와 초록을 索引語 抽出 情報源으로 동시에 실행을 하면 87%에 이른다. 즉, 本文이 索引語를 추출하는 情報源으로서의 情報量을 100으로 할때 표제는 약 54라는 情報量을 갖게되며, 抄錄은 약 84라는 정보량을 갖게 된다. 이를 표로 나타내면 다음과 같다.

本文 標題 抄錄

	本 文	標 題	抄 錄
길 이	약 18,000	7.55	148.9
索引語數	5.75개	3.12개	4.8개(중복허용15개)
情 報 量	100	54	84

이와 같은 결과는 情報處理에 있어서 KWIC索引으로 접근점을 제공하면 예상 재현율이 54%인 것을 의미하며, 抄錄을 索引語를 추출하는 情報源으로 간주하면 예상 재현율이 84%일 것이다. 이와 같이 標題와 抄錄의 情報量이 높은 이유는 실험대상으로 선정한 抄錄이 통보적 抄錄에 가까웠으며, 그 형태는 일반적인 抄錄보다도 要約文(TERSE CONCLUSION)에 가까웠기 때문으로 생각된다. 왜냐하면 앞에서 언급한 일반적인 초록의 크기는 평균단어의 크기가 50단어에서 100단어 미만인 것에 비해 위의 실험에서 사용된 평균 抄錄의 크기는 약 150개의 단어로 구성이 되어 있기 때문이다.

17) 본문 가운데는 반드시 키워드로 선정할 단어가 반드시 포함되어 있으며, 索引語로 선정될 본문이 키워드를 전부 갖고 있는 경우를 情報量 100으로 간주한다.

4. 英文 데이터베이스를 대상으로 한 實驗

본 분석은 圖書館情報學分野의 文獻情報를 다루고 있는 LISA파일을 대상으로 하였다. 이 실험의 의도는 標題와 抄錄이 檢索對象으로 어느 정도의 表意性을 갖고 있는가를 측정한 것이다.

(1) 實 驗 環 境

앞서의 實驗은 標題와 抄錄이 어느 정도의 情報量을 갖고 있는지를 각 情報源 부분과 索引語와의 매칭상태를 비교하여 분석을 하였으며, 이번의 실험은 표제와 초록과 색인어간의 연관성을 비교하고자 한다. 實驗方法은 硏究者가 임의로 선정한 25개의 索引語[18]를 檢索語로 하여 직접 데이터베이스에서 결과를 추출하였다.

(2) 結 果 分 析

본 實驗에 出力한 計數情報는 다음과 같은 방식으로 얻었다.

A : 갑이라는 索引語를 갖고 있는 문헌의 총수(예 : AUTO INDEX/DE)
B : 갑이라는 索引語가 標題에 나타난 文獻의 총수(예 : AUTO INDEX/TI)
C : 갑이라는 索引語가 抄錄에 나타난 文獻의 총수(예 : AUTO INDEX/AB)
D : 갑이라는 索引語가 標題와 抄錄에 동시에 나타난 文獻의 총수
(예 : AUTO INDEX/TI and /AB)

A의 평균값은 1076.6개이며, B의 평균값은 272.24개, C의 평균값은 623.32개, D의 평균은 124.12개이다. 이상의 情報는 각 필드의 檢索効率을 측정하는데 중요한 역할을 한다. 재현율을 정확하게 측정하기 위해서는 그 檢索語를 索引語로 갖고 있는 정보군의 수를 정확하게 파악해야 하며 A는 이에 대한 정보를 제공한다.

i) 標題의 情報量

이용자에게 적합한 情報의 접근점이 표제에 하나 이상의 索引語로 들어있을 비율은 25%이다. 즉, 自動索引에서 標題만을 索引語抽出對象으로 선정할 경우는 索引語가 標題에 포함될 확률이 25%이란 것을 의미하며, 나머지 75%는 표제로서 접근점을 제공하지 못하며 이는 문헌군에서 정보의 死藏率이 75%임을 의미한다.

ii) 抄錄의 情報量

이용자에게 적합한 정보의 접근점이 抄錄에 하나 이상의 索引語로 입력되어

---

18) 이에 대한 情報는 附錄에 기재하였다.

있을 확률은 57.9%이다. 초록만을 대상으로 索引語를 선정한다면, 정보의 死藏率은 42.1%가 된다.

iii) 혼합한 情報量

索引語 抽出對象을 표제와 초록 모두를 선정하였을 경우는 색인어가 이에 포함되어 있을 확률은 71.7%이며, 정보의 死藏率은 28.3%이다. 이와 같이 情報量이 커지는 것은 標題와 抄錄間의 정보의 독립성이 國內 文獻보다 크기 때문이다.

Ⅲ. 結　　論

이상과 같은 실험에서 얻은 결과를 요약하면 다음과 같다. 自動索引은 입력 대상 부분에 따라 索引語數가 달라진다. 國內에서 구축되고 있는 많은 데이터베이스 가운데 키워드 추출의 대상으로는 표제만을 선택하고 있는데 이 방식으로 키워드를 선정한 실험에는 다음과 같은 결과를 얻을 수 있었다.

① 國內文獻의 경우는 標題만을 선정대상으로 할 경우에는 표제에 색인어가 포함되는 수가 53%에 이른다. 이 수치는 적지 않은 수치이지만 만약 표제가 2단어로 구성이 되어 있다면 실제 키워드의 수는 1개가 된다. 즉, 표제가 나타내는 索引性은 뛰어나지만 효과적인 접근점 확보는 전적으로 표제의 길이에 달려 있기 때문에 일정한 索引語를 선정하기 위해서는 표제만으로 색인어를 추출하는 것은 미흡하다.

② 抄錄까지를 自動 索引語 抽出 選定 對象으로 한다면 國內文獻의 경우는 84%의 索引語가 추출될 수 있으며, 이는 抄錄만으로는 索引語를 완벽히 추출할 수 있는 대상이 될 수 없음을 알 수 있다.

③ 데이터베이스를 이용한 外國文獻을 대상으로 한 情報量 測定은 표제는 주요 문헌을 나타낼 수 있는 정보량이 25%였으며, 抄錄까지 확장하였을 경우가 72%의 정보량을 갖는다. 이는 標題와 抄錄으로는 索引語를 추출할 수 있는 대상물로는 부족하다는 것을 의미한다.

④ 自動索引技法 活用時에 기존의 표제와 초록만으로 索引語를 추출할 경우는 적절한 수의 색인어를 선정할 수 없다.

그러므로 앞으로는 自動索引技法을 활용하여 효과적인 색인어를 추출하기 위해서는 본문에 근접한 완전한 情報量을 유지할 수 있는 다른 대용물의 개발이 모색되어야 할 것이다.

〈附　　録 1〉

1 automatic indexing
2 library management
3 reference service
4 library automation
5 information network
6 information retrieval
7 information technology
8 automatic abstracting
9 automatic classification
10 library network
11 library history
12 system analysis
13 library administration
14 collection development
15 index language
16 subject heading
17 citation analysis
18 university library
19 subject analysis
20 information processing
21 resource sharing
22 school library
23 library standard
24 natural language
25 thesaurus construction