「情報管理研究」, vol. 23, no. 4, 1992. 12., pp. 1～20

原文代表情報의 比較評價에 관한 연구

An Experimental Study on the Evaluation of Surrogates

兪 안 나
(Yoo, An Na)

抄 錄

本 硏究에서는 표제, 초록, 목차를 대상으로 原文獻에 대한 반영도를 비교·평가하는 실험을 수행하여, 어떤 정보가 보다 反映度가 높은지 알아보고자 하였다. 이를 위해 우선 標題, 抄錄, 目次의 특성과 역할을 살펴보고, 代表價에 의한 평가와 검색실험을 통한 평가를 실시하였다.

키 워 드

代用物, 反映度, 情報壓縮

ABSTRACT

In this study, titles, abstracts, and tables of content were evaluated and compared. The evaluation of the representativeness of surrogates and the evaluation through retrieval experiments were carried out in this study.

Through the two experiments it was proved that titles are the best surrogate to condense information but they can't contain sufficient information because of the degree of condensation. Abstracts and tables of content are more effective than titles, while tables of content are the most effective.

KEYWORDS

Surrogate, Representativeness, Information condensation.

Ⅰ. 序　　論

情報를 탐색할 때 처음에는 原文獻의 내용을 축약적으로 나타내고 있는 것을 이용하여 정보요구에 대한 원문헌의 適合性을 예견하게 되는 경우가 많은데, 이때 이용되는 것을 '原文代表情報'(surrogates)라 할 수 있다. 原文獻의 內容을 축약적으로 나타내고 있는 것을 일반적으로 '代替物' 또는 '文獻代用物'이란 용어로 표현하고 있으나, 原文獻을 대표한다는 의미를 강조하기 위하여 본 연구에서는 原文代表情報란 용어로 사용하기로 하였다. 原文代表情報는 원문헌의 존재를 밝히고, 적합여부를 예견하게 하며, 원문헌을 대신하는 기능을 한다고 볼 수 있다. 原文代表情報가 제기능을 발휘하려면 가능한 한 原文獻의 내용에 대한 反映度가 높아야 할 것이다. 원문헌을 보고 내리는 판단과 같은 판단을 내릴 수 있어야만이 原文代表情報로 인한 誤謬가 발생하지 않을 것이며, 情報의 損失을 방지할 수 있다.

따라서 原文代表情報의 평가는 원문헌과 관련해서 이루어져야 하며, 그 핵심은 原文獻에 대한 反映度를 평가하는 것이다. 잘 구성된 原文代表情報라면 원문헌에서 중요하게 다루어진 내용을 나타낼 것이며, 이러한 반영도를 살펴봄으로써 原文代表情報에 관한 평가를 할 수 있을 것이다.

지금까지의 연구에서 평가된 原文代表情報는 주로 標題와 抄錄이었다. 이 두 原文代表情報는 여러 연구에서 다루어질 만큼 전형적인 原文代表情報이며, 확고한 原文代表情報의 위치를 가지고 있다고 볼 수 있다. 그러나 최근에 특히 온라인 목록 분야에서 主題接近에 관한 연구를 살펴보면, 目次에 대한 관심이 높아지고 있다는 것을 알 수 있다.[1)]

그러므로 본 연구에서는 지금까지 크게 주목을 받지 못한 목차를 포함하여 檢索의 접근점이 되는 原文代表情報들을 평가하여 보다 효율적인 原文代表情報가 어떤 것인가를 알아보고자 한다.

---

1) Virgil Diodato, "Tables of Contents and Book Indexes : How Well Do They Match Reader's Descriptions of Book?", *Library Resources and Technical Services*, vol. 30, no. 4, 1986, p. 403.

Ⅱ. 原文代表 情報의 평가에 관한 理論的 背景

1. 原文代表 情報

(1) 標　　題

標題란 '다른 著作物과 구별시켜 주는 이름'으로 정의된다.[2] 標題는 原文獻에서 다루고 있는 주제내용을 가장 압축된 형태로 간결하게 나타내기 때문에 이용자의 눈에 가장 잘 들어오는 요소이다. 이와 같은 이유로 標題는 문헌내용을 일별하게 하는 데 도움을 줌으로써 情報探索者가 文獻內容을 판단하는데 맨 먼저 영향을 미치는 요소가 된다.[3]

標題는 壓縮性이 뛰어나며, 거의 모든 문헌에 부여되어 있기 때문에 실제 分類作業이나 索引作業에 이용되고 있으며, 따라서 이에 관해 많은 연구가 이루어져 왔다. 특히 KWIC 索引이 구현되어 標題索引을 이용하는 것에 대한 가능성을 인정받은 이후로 標題에 대한 관심이 증대되었다.

그러나 標題의 특성으로 인한 短點도 있다. 標題는 내용의 일부만을 나타내는 것, 暗示하는 것, 象徵에 지나지 않는 것[4] 등이 있으므로 標題만으로는 정확한 판단을 내릴 수 없다. 특히, 標題는 資料를 선택하는 데 필수적인 특정성에 있어서 많은 경우 非通報的이며, 簡潔해야 하기 때문에 지나치게 일반적이되므로 문헌 선정에 있어서는 대단히 불량한 안내자가 될 가능성이 있다는 의견도 있다.[5]

(2) 抄　　錄

抄錄은 2次情報 중 索引과 함께 그 중요성을 더해가고 있는 原文代表情報로서, 이는 抄錄雜誌의 증가와 自動抄錄에 관해 계속되는 연구를 통해 쉽게 알 수 있다. 抄錄이란 용어는 拔萃, 要約 등과 유사한 개념으로 성격이나 구성방법이 조금 다르기는 하지만 이들 모두를 통칭하여 쓰이기도 하는데, 그 정

---

2) *Harrod's Librarians' Glossary of Terms Used in Librarianship, Documentation and Book Crafts*, 6th ed., Compiled by Ray J. Prytherch, Aldershot Hants : Gower Publishing, 1987, p. 783. ; *ALA Glossary of Library and Information Science*, Edited by Heartsill Young, Chicago : American Library Association, 1983, p. 229.

3) Richard A. V. Diener, "Informational Dynamics of Journal Article Titles", *Journal of ASIS*, vol. 35, no. 4, 1984, p. 223.

4) 이재철, 「주제명 목록의 연구」, 서울 : 연세대학교 도서관학과, 1959, p. 80.

5) R. K. Maloney, "Title versus Title/Abstract Text Searching SDI System", *Journal of ASIS*, vol. 25, no. 6, 1974, pp. 370～373.

확한 의미를 정의를 통해 살펴보면, 抄錄은 '知識記錄物을 요약하여 나타낸 原文代表情報로, 관련된 데이터나 경우에 따라서는 批評的 論評을 포함하는, 문헌에 대한 說明的 記述'[6]이라고 할 수 있다.

이러한 抄錄의 기능과 용도를 살펴보면, 抄錄은 원문헌의 내용을 요약하여 원문헌을 대신하는 기능과 적합문헌 판정에 도움을 주는 情報檢索의 기능을 아울러 가지고 있다.[7] 즉, 抄錄은 索引言語 다음으로 문헌의 적합여부를 판별할 수 있도록 구비된 장치이며,[8] 이러한 기능 외에도 情報配布와 情報檢索 시스템에서 情報蓄積 및 探索을 위한 키워드 추출에 이용된다.[9]

또한 抄錄을 이용함으로써 이용자나 색인자는 모두 時間을 節約할 수 있다. 즉, 이용자는 原文獻을 읽기에 앞서 抄錄을 살펴봄으로써 원문헌을 읽고 선택하는 경우보다 시간을 절약할 수 있다. 索引作業時 抄錄을 이용하게 되면 原文獻을 이용하는 것보다 훨씬 빨리 작업을 할 수 있는데, 이는 읽어야 할 단어의 수가 적고 색인될 요소가 더 잘 식별되기 때문이다. 따라서 색인작성에 드는 時間과 費用을 줄일 수 있고, 質的인 면도 거의 손실이 없어 抄錄은 索引의 효율을 높이는 역할을 한다고 볼 수 있다.[10]

抄錄의 機能과 活用度를 살펴볼 때 대단히 우수한 原文代表情報이지만, 그 정의에 합당한 抄錄을 작성하는 일이 결코 쉬운 것은 아니다. 얼핏 생각하기에는 著者가 抄錄을 직접 작성하는 것이 가장 이상적일 것으로 판단되나, 抄錄作成에 숙련되지 못한 사람이 작성해야 할 경우가 대부분이기 때문에 主題와 관련된 枝葉的인 생각을 포함시키는 경우가 많으며,[11] 자신의 연구에 대해 客觀的으로 판단하기 어렵기 때문에 균형을 잃게 되거나, 原文獻의 서론적인 내용이 되기 쉽고, 문장의 統一性이 결여될 수 있는 단점이 있다.[12] 좋은 抄錄을 작성하기 위해서는 主題分野에 관한 광범위한 專門的 知識이 있어야 하며,

---

6) Donald B. Cleveland and Ana D. Cleveland, *Introduction to Indexing and Abstracting*, 2nd Edition, Englewood, Colorado : Libraries Unlimited, 1990, p. 291.
7) 정영미, 「정보검색론」, 서울 : 정음사, 1987, p. 173.
8) 李泰榮, "韓國語 抄錄文의 文章과 內容에 관한 硏究", 「情報管理硏究」, vol. 21, no. 1, 1990, p. 1.
9) C. Guinchart and M. Menou, 「정보관리론」, 사공철·김태수 공역, 서울 : 구미무역(주) 출판부, 1987, p. 176.
10) H. Borko and C. L. Bernier, *Abstracting Concepts and Methods*, New York : Academic Press, 1975, pp. 6~8.
11) Masse Bloomfield, "Simulated Machine Indexing Part 2 : Use of Words from Title and Abstracts for Matching Thesauri Headings", *Special Libraries*, vol. 57, no. 2, 1966, pp. 232~235.
12) 高仁哲, "情報傳達媒體로서의 抄錄에 관한 小考", 「國會圖書館報」, vol. 164, no. 3, 1983, p. 72.

抄錄作成의 諸原則을 숙지하고 있어야 한다. 따라서 抄錄을 작성하는 데는 많은 비용이 드는 단점이 있으며, 모든 抄錄이 반드시 우수한 原文代表情報라고 할 수 없다.

(3) 目　　次

目次란 '圖書 등의 내용을 그 구성에 따라 순서대로 표시한 次例'[13]로 정의된다. 目次는 지금까지 크게 주목받지 못해 온 것이 사실이나 分類作業이나 索引作業에서 실제로 이용되고 있다. 분류에 있어서 標題만으로 정확한 主題를 파악할 수 없거나, 標題가 原文獻의 내용을 명확하게 나타내고 있다고 판단되는 경우라도 반드시 目次를 통하여 主題를 판단하도록 권하고 있는데, 이는 目次가 原文獻의 내용을 서술한 순서를 비교적 자세하게 나열하고 있기 때문에 主題를 파악할 수 있을 뿐만 아니라, 그 主題가 다루어진 관점까지도 추측할 수 있어 分類에 있어서 중요한 위치를 차지하고 있기 때문이다.[14]

또한 自動抄錄에 관한 연구에서 標題, 目次 등의 특수한 부분에 나타나는 단어를 단서어로 사용하여 가중치를 부여하는 標題語技法은 目次가 구분된 해당 기술 부분에서 가장 핵심적인 내용으로 구성되며 나누어진 각 부분에 대해 標題와 같은 역할을 한다는 것을 前提로 한 것이다.[15]

目次에 대한 관심은 보다 나은 主題探索에 대한 요구에서 비롯되었다. 온라인 目錄에 관한 評價 프로젝트에서 이용자 설문조사 결과 目次, 要約, 卷末索引을 탐색할 것을 희망하고 있으며, 이는 주제접근 행태를 고려해 볼 때 그리 놀라운 사실이 아니다.[16] 또한 記入語를 살펴본 결과 圖書에 부여된 주제명 표목에서보다는 目次, 卷末索引에서 나온 용어가 많다는 것을 알 수 있었다. 따라서 이용자의 주제탐색을 돕기 위해서 目次나 卷末索引을 포함시킬 것을 고려하게 되었다.[17], [18]

---

13) 「圖書館用語辭典」, 圖書館問題研究會, 圖書館用語委員會 編著, 東京 : 角川書店, 1982, p. 621.
14) 김명옥, 「자료분류법」, 서울 : 구미무역(주) 출판부, 1986, p. 190.
15) H. P. Edmunson, "New Methods in Automatic Extracting", *Journal of ACM*, vol. 16, no. 2, 1969, pp. 265～285.
16) Karen Markey, "Subject Experiences and Needs of Online Catalog Users : Implications for Library Classification", *Library Resources and Technical Services*, vol. 29, no. 1, 1985, p. 40.
17) Karen Markey, "Users and the Online Catalog : Subject Access Problems", *The Impact of Online Catalogs*, Edited by Joshep R. Matthews, New York : Neal－Schuman Publishers, 1986, pp. 37～38.
18) Carol A. Mandel, "Enriching the Library Catalog Record for Subject Access", *Library Resources and Technical Services*, vol. 29, no. 1, 1985, pp. 5～15.

코크레인(P. A. Cochrane)은 目次에서 주제를 기술하는 용어를 경제적으로 도출해낼 수 있다는 것을 보여 주었으며,[19] 이들 연구에서의 제안으로 현재는 최근 출판된 單行本에 대한 目次 데이터 베이스인 'Current Book Contents'가 개발되어 판매 중이다.[20] 또한 主題接近에 있어 目次의 중요성을 인식하여 目次를 2次情報에 포함시키는 정책을 수립해야 한다는 주장도 있다.[21]

2. 代表價에 의한 평가

原文代表情報가 잘 구성되었다는 것은 原文獻에서 중요하게 다루어진 개념이 原文代表情報에 잘 나타나 있다는 것을 의미한다. 따라서 원문헌에서 중요하게 다루어진 개념이 얼마나 잘 나타나 있는가를 통해 原文代表情報에 관한 정확한 평가를 할 수 있을 것이다. 또한 이는 原文代表情報의 정의가 원문헌의 기능을 대신한다는 것을 상기해 보면 더욱 분명해진다.

原文代表情報를 구성하는 경우, 우선 原文獻의 主題分析, 즉 원문헌에서 중요하게 다루어진 것을 파악해야 할 필요가 있듯이, 逆으로 原文代表情報를 평가하는 경우에는 原文代表情報에 나타난 요소들이 원문헌에서 얼마나 중요하게 다루어진 것인가를 알 필요가 있다. 예를 들어, 自動抄錄을 구성하는 경우, 핵심문장을 파악하기 위해서 문장의 代表價를 구하였는데, 문장의 代表價는 문장을 구성하는 單語의 代表價의 合이다.[22] 여기서 '代表價'(representativeness)라는 것은 단어나 문장이 그것이 속한 문장이나 원문헌의 의미를 얼마만큼 잘 전달하고 있느냐 하는 것이며,[23] 이와 유사한 의미의 重要度(significance)는 단어나 문장이 원문헌의 내용을 나타냄에 있어서의 價値이다.[24]

---

19) P. A. Cochrane, "Books Are for Use : Final Report of the Subject Access Project", *Redesign of Catalogs and Indexes for Improved Online Subject Access*, Phoenix : Oryx Press, 1985, pp. 394~457.
20) Diodato, 앞의 글.
21) F. E. Dehart and Karen Matthews, "Subject Analytics and Table of Contents in Essay Collections : Implications for Searching", *Technical Service Quarterly*, vol. 6, no. 3;4, 1989, pp. 57~69.
22) H. P. Luhn, "The Automatic Creation of Literature Abstracts(Auto-Abstracts)", *IBM Journal of Research and Development*, vol. 2, no. 2, 1958, pp. 159~165 ; P. E. Baxendale, "Machine-Made Index for Technical Literature-An Experiment", *IBM Journal of Research and Development*, vol. 2, no. 4, 1958, pp. 354~361 ; Edmunson, 앞의 글.
23) Borko and Bernier, 앞의 책, p. 232.
24) Borko and Bernier, 앞의 책, p. 233.

代表價를 산출하여 原文代表情報를 구성한 것과 유사하게 原文代表情報를 평가할 수 있다. 즉, 原文代表情報에 나타난 요소들이 원문헌에서 얼마나 중요하게 다루어진 것인가를 파악하는 데도 代表價를 활용하는 것이다. 이를 활용한다면 原文代表情報의 구성성분이 원문헌에서 갖는 중요도를 산출함으로써 原文代表情報의 원문헌에 대한 반영도를 알 수 있을 것이다.

Ⅲ. 原文代表情報의 평가 실험

原文代表情報를 평가하기 위한 실험은 代表價에 의한 평가와 檢索實驗을 통한 평가로 구성된다. 본 실험에서 평가할 標題, 抄錄, 目次의 범위를 기술하면 다음과 같다.

標題는 학술잡지의 記事名과 학위논문의 論文名을 포함하며, 초록은 國文抄錄을 대상으로 하며, 국문초록이 없는 경우 論文槪要, 要約을 모두 포함한다. 目次는 문헌에 따라 그 구성 단계가 다양하므로 章題와 節題만을 대상으로 한다.

1. 代表價에 의한 실험

(1) 實驗의 개요

原文代表情報의 구성성분이 원문헌에서 갖는 중요도를 반영하여 原文代表情報를 평가하고자 原文代表情報 구성성분의 원문헌에서의 출현빈도에 근거한 原文代表情報의 代表價를 산출한다. 본 실험에서는 原文代表情報의 구성성분 단위를 명사구로 하고, 중요도를 原文代表情報에 나타난 명사구가 원문헌에서 갖는 가중치로 산출한다.

1) 構成成分 單位로서의 명사구

명사구를 단위로 하는 이유는 명사구가 정보전달의 기본적 단위로서 構文的·心理的 內容指示者[25]로 인식되고 있으며, 또한 실제로 檢索의 접근점들이 명사구의 형태를 취하고 있고, 명사구를 이용함으로써 더 나은 檢索効率을 얻을

25) Robert Kenneth Waldstein, "The Role of Noun Phrases as Content Indicators", Ph. D. Dissertation, Syracuse University, 1981, pp. 95~103.

수 있기 때문이다.[26] 본 실험에서 설정한 명사구의 정의는 다음과 같다.

〈名詞句의 정의〉
〈名詞句〉::={명사}* ¦ 〈명사〉{〈조사〉〈명사〉}*
〈助　詞〉::='의' ¦ '와'

2) 명사구의 가중치와 原文代表情報의 代表價

代表價는 原文代表情報에 나타난 명사구들이 원문헌에서 갖는 가중치를 합하여 얻을 수 있다. 본 실험에서 사용한 가중치 산출방식은 出現頻度에 근거한 것이다.

출현빈도에 근거한 加重値 算出方式이 많이 제시되어 있으나, 대부분의 산출방식은 특정 문헌집합 내에서 用語의 文獻頻度를 고려하여 가중치를 산출하는 것이다. 이러한 가중치 산출방식은 본 실험에서와 같이 個別文獻에 대해 가중치를 산출하는 경우와는 전제조건이 다르다. 한편, 個別文獻에서 加重値를 산출한 연구는 존스(L. P. Jones)의 연구에서 찾아볼 수 있다.[27] 명사구(P)를 구성하는 단어의 출현빈도의 합을 W, 명사구의 출현빈도를 F, 명사구 구성 단어수를 N이라고 하면, 명사구 P의 가중치 W(P)는 다음과 같은 공식으로 산출한다.

$$W(P) = W \times F \times N^2$$

여기서 N^2은 N값을 변화시켜 실험한 결과 경험적으로 얻은 것이다. 이 加重値 算出公式은 출현빈도가 높은 單一名詞보다는 名詞句에 보다 높은 순위를 부여하고자 한 것으로, 이는 단일명사보다는 명사구가 보다 더 나은 內容指示語이기 때문이다. 순위를 부여하기 위해서 가중치를 산출한 것이므로 가중치를 원문헌에서의 重要度로 볼 수 있다.

따라서 原文代表情報에 나타난 명사구가 원문헌에서 갖는 가중치를 합산하여 原文代表情報의 代表價를 산출한다. 原文代表情報의 代表價가 높을수록 즉, 가중치의 합이 클수록 그 原文代表情報는 원문헌을 잘 반영하고 있는 것으로 볼 수 있다.

---

26) L. P. Jones, et al., "INDEX : The Statistical Basis for an Automatic Conceptual Phrase-Indexing System", *Journal of ASIS*, vol. 41, no. 2, 1990, pp. 87~96.
27) 위의 글.

(2) 實驗環境

본 실험에서는 標題, 抄錄, 目次를 평가하기 위해서 이 세가지 原文代表情報를 모두 구비하고 있는 「정보관리학회지」의 기사를 대상으로 했다. 「정보관리학회지」 기사 중 10편을 무작위로 추출하여 실험대상으로 하였으며, 실험대상 문헌의 리스트는 〈附錄 1〉에서 열거하였다. 이 세가지 原文代表情報는 모두 著者에 의해 작성된 것이다.

또한 출현빈도 산출을 위한 프로그램은 조합형 한글을 지원하는 HSV 하에서 TURBO-C 2.0 버전으로 작성되었다.

(3) 實驗節次

原文代表情報의 代表價 산출을 위한 실험은 다음과 같은 단계로 이루어진다.

1) 原文獻 파일 생성

名詞句가 원문헌에서 출현한 빈도를 구하기 위해 우선 원문헌 전체를 입력하여 텍스트 파일을 구성한다. 이때 원문헌에 대해 다음과 같은 통제를 가한다.

첫째, 각 原文代表情報 즉, 標題, 抄錄, 目次는 출현빈도에 영향을 미치므로 제외한다.

둘째, 圖表와 脚註, 參考文獻은 제외한다.

셋째, 漢字는 한글로 변환하여 입력한다.

넷째, 한 문헌 내에서 표기를 달리한 외래어는 출현빈도가 높은 쪽으로 통일하여 입력한다.

2) 名詞句 探索 파일 생성

각 原文代表情報에서 원문헌에 대한 내용 지시 기능이 없는 불용어를 제외한 후 標題, 抄錄, 目次의 명사구 探索 파일을 생성하는데, 이때 명사구를 구성하는 單一名詞들도 출현빈도 산출을 위해 探索單位가 된다. 명사구 탐색파일 생성시 불용어에 대한 통제 외에 다음과 같은 통제를 가한다. 즉, 原文代表情報에서 중복되는 어휘가 생략된 경우 생략된 것을 복원하여 名詞句를 형성한다.

예) 학문활동의 세분화와 지역화

→ 학문활동의 세분화

학문활동의 지역화

3) 出現頻度 산출

名詞句 探索 파일로 원문헌을 탐색하여 명사구와 명사구를 구성하는 單一名詞들이 원문헌에서 출현한 출현빈도를 구한다. 이때 완전히 일치하는 경우에만 출현빈도에 포함시키므로 同義語나 關聯語는 출현빈도에 포함되지 않는다.

4) 加重値 산출

출현빈도 산출결과로부터 가중치 산출공식에 의해 각 명사구의 加重値와 合, 平均을 구한다. 존스의 공식 중 명칭을 약간 수정하여 명사구(P)를 구성하는 단일명사의 출현빈도의 합을 Wf, 명사구의 출현빈도를 Pf, 명사구 구성 단어수를 N이라고 하면, 명사구 P의 가중치 W(P)는 다음과 같이 구해진다.

$$W(P) = Wf \times Pf \times N^2$$

(4) 實 驗 結 果

출현빈도 산출결과, 출현빈도가 0인 명사구에 대해서는 원문헌에서 출현한 명사구의 형태로 나누어 준 뒤, 나눈 형태로 출현빈도를 다시 구한다. 명사구를 구성하는 단일명사들의 합과 명사구 출현빈도, 단어수로 가중치를 산출하는데 加重値 算出 結果物은 〈圖 1〉과 같다.

〈圖 1〉의 가중치 산출 결과를 살펴보면, '非公式 커뮤니케이션'이란 名詞句는 '비공식', '커뮤니케이션'의 두 개의 單一名詞로 구성되는데, 각 단일명사의 원문헌에서의 출현빈도는 171회, 161회이므로, 단일명사의 출현빈도의 합 332가 구해지며, 명사구의 원문헌에서의 출현빈도는 91회, 명사구를 구성하는 단일명사의 수가 2이다. 따라서, 加重値 算出公式 $W(P)=Wf\times Pf\times N^2$에 의해 가중치 120,848이 구해진다.

실험문헌에 대한 原文代表情報의 代表價 산출 결과는 〈表 1〉과 같다. 먼저 〈表 1〉에서 각 原文代表情報의 명사구의 수를 살펴보면 標題는 1~6개로 평균

〈圖 1〉 加重値 산출 結果의 예

Source	Weight 2
硏究(207) 활동(34) ; 241 15 2	14,460
비공식(171) 커뮤니케이션(161) ; 332 91 2	120,848

Sum 2 = 135,308, Average 2 = 67,654.0

3.7개, 抄錄은 8~30개로 평균 17.7개, 目次는 7~31개로 평균 17개로 이루어져 있다. 文獻當 代表價는 標題가 60,057, 抄錄이 226,019, 目次가 266,160이다.

이러한 결과는 原文代表情報를 구성하는 명사구 수에 영향을 받은 것이며, 原文代表情報의 길이에 영향을 받지 않도록 명사구 수로 代表價를 나누어 보면 〈表 2〉와 같다. 〈表 2〉에서 명사구의 수로 각 代表價를 나눈 결과를 문헌별로 살펴보면 標題가 가장 높은 값을 갖는 경우가 4편, 抄錄이 가장 높은 값을 갖는 경우가 3편, 目次가 가장 높은 값을 갖는 경우가 3편이었다. 代表價를 명사구의 수로 나누어 준 평균값은 標題가 가장 높았으며, 抄錄보다는 目次가 우세한 것으로 나왔다.

〈表 1〉 原文代表情報의 代表價 산출 결과

文獻番號	標題		抄錄		目次	
	名詞句數	代表價	名詞句數	代表價	名詞句數	代表價
1	6	25,927	11	33,374	7	36,179
2	5	1,446	22	16,008	12	13,894
3	3	46,636	19	476,945	13	164,549
4	2	135,308	18	585,433	21	791,512
5	4	14,289	30	44,151	14	21,336
6	3	5,573	19	50,890	8	19,695
7	4	10,588	8	41,145	31	152,103
8	6	237,307	16	460,872	25	1,071,085
9	1	5,100	15	78,486	29	171,818
10	3	118,392	19	472,881	10	219,424
合 計		600,566		2,260,185		2,661,595
平 均		60,057		226,019		266,160

〈表 2〉 原文代表情報 代表價의 평균값

文獻番號	標題	抄錄	目次
1	4,321	3,034	5,168
2	289	727	1,157
3	15,545	25,102	12,657
4	67,654	32,524	37,691
5	3,572	1,471	1,524
6	1,857	2,678	2,461
7	2,647	5,143	4,906
8	39,551	28,804	42,843
9	5,100	5,232	5,924
10	39,464	24,888	21,942
合 計	180,000	129,603	136,273

2. 檢索效率 비교 실험

· (1) 實驗의 개요

原文代表情報를 검색 측면에서 평가하기 위해 각각의 原文代表情報別로 색인어 파일을 구축, 검색실험을 실시하여 검색효율을 비교하였다.

檢索技法으로는 불리안 검색기법을 채택하였는데, 그 이유는 이 기법에 대한 단점에 대해 많이 보고되고 있기는 하지만 현행 情報檢索 시스템에서 보편적으로 사용되는 기법이기 때문이다. 따라서 檢索實驗에서는 불리안 검색기법의 특성이 검색결과에 영향을 미칠 수 있다.

(2) 實驗對象 文獻의 선정

검색실험을 위한 실험대상 문헌은 「정보관리학회지」의 기사 중 학술논문 49편과 국립중앙도서관에 입수된 정보학 분야(전산학분야 제외)의 학위논문 중 標題, 國文抄錄, 目次를 갖추고 있는 53편을 합하여 총 102편으로 구성하였다. 학위논문에서 국문초록이 없는 경우 論文槪要, 要約 등을 모두 포함하였다.

(3) 實驗 데이터

실험대상 문헌 102편을 대상으로 각 문헌의 標題, 抄錄, 目次에서 색인어를 추출하여 實驗 데이터를 구성하였는데, 이때 각 原文代表情報別로 색인어 파일을 마련하였다. 즉, 표제 색인어 파일, 초록 색인어 파일, 목차 색인어 파일의 3개의 색인어 파일이 구축되었다.

索引語는 가능한 한 原文代表情報에 나타난 용어를 그대로 추출하였으며, 다음과 같이 약간의 통제를 가하였다.

첫째, 색인어가 복합개념인 경우 복합개념의 색인어와 분리한 형태의 색인어를 추출하였다.

예) 이용자연구 → 이용자, 이용자 연구

비공식 커뮤니케이션 → 커뮤니케이션, 비공식 커뮤니케이션

둘째, 表記方式이 다양한 外來語의 경우는 출현빈도가 높은 것으로 통일하였다.

예) DB, 데이타 베이스, 데이터 베이스 → 데이터 베이스

셋째, 漢文은 한글로 변환하였으며, 英語는 그대로 입력하였다.

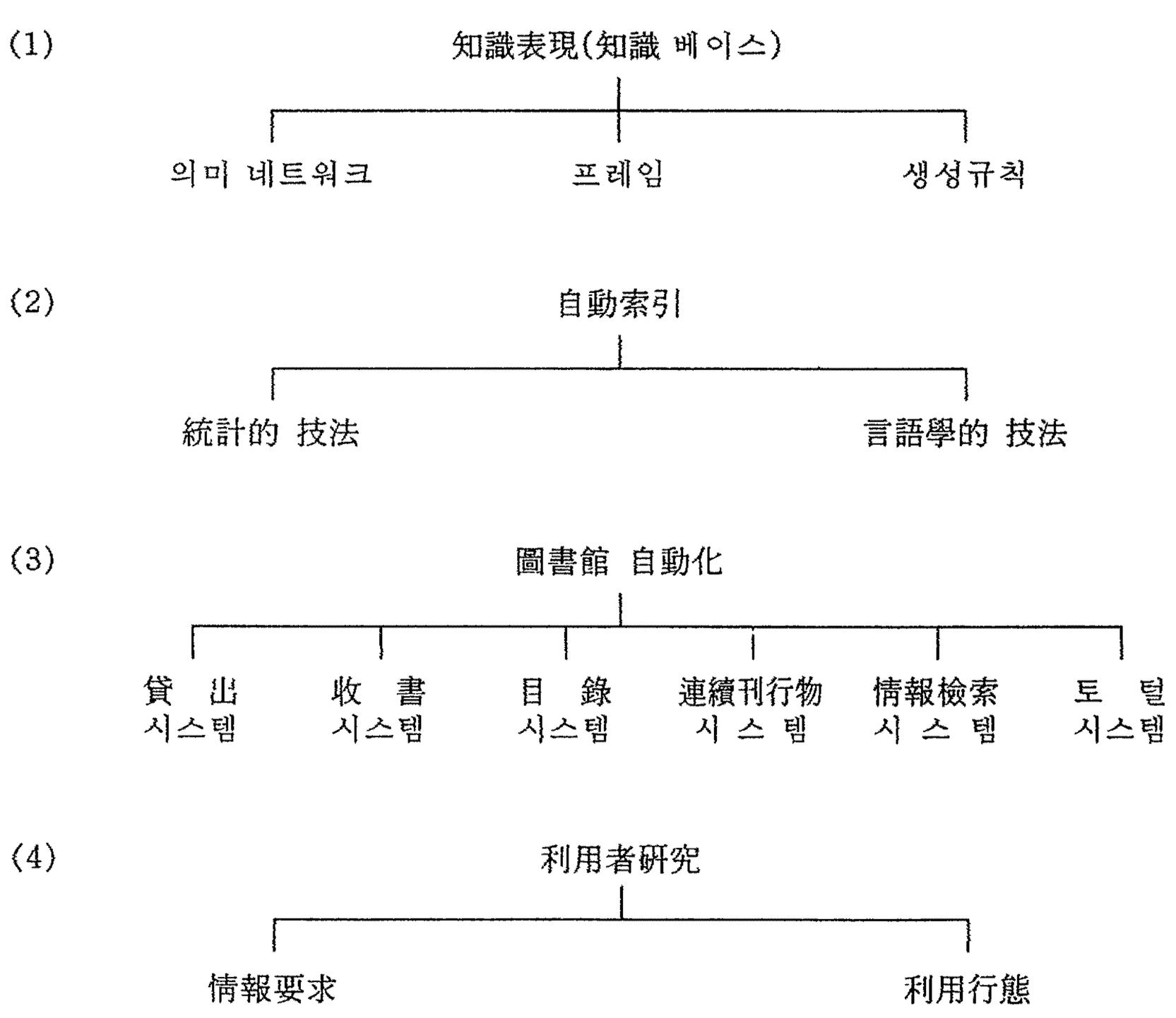
이러한 방식으로 각 原文代表情報別로 색인어 파일을 구성한 결과 색인어수는 標題 索引語 파일이 평균 3.7 단어, 抄錄 索引語 파일이 9.7 단어, 目次 索引語 파일이 7.8 단어였다.

(4) 探索文 형성

탐색문은 原文代表情報의 일반성과 특정성을 각각 살펴보기 위하여 광범위한 탐색문과 범위를 좁혀 특정성을 높인 하위 탐색문으로 형성하였다.

실험대상 문헌에서 다루어진 주제 중 다루어진 문헌의 수가 비교적 많은 '知識表現', '自動索引', '圖書館 自動化', '利用者 硏究'의 주제를 택하여 광범위한 탐색문을 형성하였다. 하위 탐색문은 광범위한 탐색문의 하위 주제를 〈圖 2〉와 같이 나누어 형성하였다. 그런데 '자동색인', '도서관 자동화'에서와 같이 하위 주제의 용어만으로 下位 探索文을 형성하는 경우에는 광범위한 탐색문에 대해 범위를 좁힌 下位 探索文이 될 수 없으므로 上位 探索語를 'AND'로 조합하여 하위 탐색문을 형성하였다.

〈圖 2〉 探索文 계층



(5) 實驗 결과

각 原文代表情報의 검색효율은 재현율과 정확률로 평가하였다.

$$\text{再現率} = \frac{\text{검색된 적합문헌 수}}{\text{적합문헌 총수}} \qquad \text{正確率} = \frac{\text{검색된 적합문헌 수}}{\text{검색된 문헌 총수}}$$

검색실험 결과 原文代表情報別로 검색된 문헌수는 〈表 3〉과 같다.

〈表 4〉는 광범위한 探索文에 대한 檢索效率인데, 광범위한 탐색의 결과를 살펴보면 再現率은 표제 색인어 파일의 경우가 0.42로 초록과 목차 색인어 파일의 0.73보다 매우 낮았으며, 正確率은 표제 색인어 파일이 0.93, 목차 색인어 파일이 0.92로 초록 색인어 파일의 0.84보다 높았다. 正確率은 각 原文代表情報別로 큰 차이가 없었으나, 再現率의 경우 표제 색인어 파일이 현저하게 떨어졌다.

〈表 5〉는 下位 探索文에 대한 檢索效率인데, 특정성이 높은 탐색문의 경우 재현율은 표제 색인어 파일이 0.27로 광범위한 탐색에서와 같이 많이 떨어지며,

〈表 3〉 原文代表情報別 검색문헌수

探索文番號	總適合文獻數	標題		抄錄		目次	
		檢索된文獻數	檢索된適合文獻數	檢索된文獻數	檢索된適合文獻數	檢索된文獻數	檢索된適合文獻數
1	3	1	1	1	1	2	2
2	8	3	3	7	7	7	7
3	5	1	1	3	3	3	3
4	5	0	0	3	3	3	3
5	1	0	0	0	0	1	1
6	12	10	9	12	10	11	10
7	2	1	1	1	1	1	1
8	9	1	1	2	2	4	4
9	14	4	4	13	12	12	11
10	5	1	1	3	3	4	4
11	2	1	1	1	1	2	2
12	2	0	0	1	1	2	2
13	2	1	1	1	1	2	2
14	4	0	0	2	2	2	2
15	4	1	1	4	4	0	0
16	17	8	6	19	9	12	9
17	10	4	4	9	7	4	4
18	9	8	8	6	5	6	4

〈表 4〉 廣範圍한 探索文에 대한 검색효율

探索文 番 號	標 題		抄 錄		目 次	
	再現率	正確率	再現率	正確率	再現率	正確率
1	0.33	1.00	0.33	1.00	0.66	1.00
2	0.38	1.00	0.88	1.00	0.88	1.00
6	0.75	0.90	0.83	0.83	0.83	0.92
9	0.29	1.00	0.86	0.92	0.75	0.92
16	0.35	0.75	0.53	0.44	0.53	0.75
平 均	0.42	0.93	0.73	0.84	0.73	0.92

〈表 5〉 下位 探索文에 대한 검색효율

探索文 番 號	標 題		抄 錄		目 次	
	再現率	正確率	再現率	正確率	再現率	正確率
3	0.20	1.00	0.60	1.00	0.60	1.00
4	0.00	0.00	0.60	1.00	0.60	1.00
5	0.00	0.00	0.00	0.00	1.00	1.00
7	0.50	1.00	0.50	1.00	0.50	1.00
8	0.11	1.00	0.22	1.00	0.44	1.00
10	0.20	1.00	0.60	1.00	0.80	1.00
11	0.50	1.00	0.50	1.00	1.00	1.00
12	0.00	0.00	0.50	1.00	1.00	1.00
13	0.50	1.00	0.50	1.00	1.00	1.00
14	0.00	0.00	0.50	1.00	0.50	1.00
15	0.25	1.00	1.00	1.00	0.00	0.00
17	0.40	1.00	0.70	0.78	0.40	1.00
18	0.89	1.00	0.55	0.83	0.44	0.67
平 均	0.27	0.69	0.52	0.89	0.64	0.89

초록 색인어 파일보다 목차 색인어 파일이 더 높은 효율을 나타냈다. 正確率은 표제 색인어 파일이 약간 낮고, 초록 색인어 파일과 목차 색인어 파일은 유사하다.

실험결과를 종합해 보면 〈表 6〉의 평균 검색효율이 보여주는 것처럼 초록 색인어 파일과 목차 색인어 파일의 검색효율이 再現率 0.57, 0.66, 正確率 0.88, 0.90으로 거의 유사한데 목차가 약간 우세하며, 표제 색인어 파일은 재현율 0.31로 재현율에 있어서 많이 떨어졌다.

〈表 6〉 原文代表情報의 평균 검색효율

原文代表情報	標題		抄錄		目次	
檢索效率	再現率	正確率	再現率	正確率	再現率	正確率
平均	0.31	0.76	0.57	0.88	0.66	0.90

3. 實驗結果 분석 및 활용방안

본 연구에서는 原文代表情報를 평가하기 위해서 原文代表情報의 代表價 算出實驗과 檢索實驗을 행하였으며, 각 실험의 결과를 살펴보면 다음과 같다.

우선 原文代表情報의 代表價 산출 실험결과 名詞句의 數를 고려하지 않은 경우 代表價는 標題, 抄錄, 目次의 순으로 증가하였으며, 특히 標題는 抄錄과 目次에 비해 그 값이 많이 낮았다. 그러나 名詞句의 數를 고려한 경우에는 標題의 代表價가 가장 높았고, 그 다음은 目次, 抄錄의 순이었다. 따라서 標題가 情報의 壓縮性이 가장 뛰어난 原文代表情報임을 알 수 있었으며, 또한 실험대상 문헌의 標題가 원문헌의 내용을 잘 나타내고 있다는 것을 의미한다. 目次와 抄錄간의 차이는 근소했는데, 이는 抄錄이 指示的인 性格을 띠는 경우 目次와 거의 유사한 내용으로 이루어졌기 때문인 것으로 보인다.

따라서, 代表價 실험의 결과를 종합해 보면 標題가 정보의 압축성에 있어서는 뛰어나지만, 代表價 合이 너무 적어서 정보의 손실이 많을 것으로 생각되며, 抄錄과 目次의 경우는 유사한데 목차가 약간 더 우세한 것으로 나왔다.

두번째, 原文代表情報를 검색실험을 통해 평가하기 위해서 原文代表情報別로 색인어 파일을 구성하여 각 색인어 파일로 탐색한 경우의 검색효율을 비교하였다. 正確率은 세 가지 原文代表情報의 색인어 파일이 유사했다. 再現率의 경우는 표제 색인어 파일이 초록 색인어 파일이나 목차 색인어 파일보다 많이 떨어졌으며, 特定性이 높은 하위탐색의 경우 그 차이가 더 심했다. 따라서 標題는 자료를 선택하는 데 있어 필수적인 특정성이 결여된다는 점이 입증되었다. 초록 색인어 파일과 목차 색인어 파일은 그 길이를 나타내는 索引語 數에 있어서 抄錄이 약간 많으나, 檢索效率에 있어서는 目次 索引語 파일을 탐색한 경우가 더 나았다.

따라서 原文代表情報 중 표제는 초록이나 목차보다 代表價와 檢索效率 모두가 많이 떨어지며, 초록이나 목차의 代表價와 檢索效率은 거의 유사하나 목차

가 조금 더 나은 것을 알 수 있어 目次가 보다 효율적인 原文代表情報라고 볼 수 있다.

이러한 原文代表情報에 대한 평가 결과를 통해 目次를 索引 對象物에 포함시키거나 목차를 포함한 데이터 베이스를 구축하는 방안을 생각해 볼 수 있다. 즉, 온라인 閱覽 目錄이나 書誌情報檢索 시스템에서 목차에서 추출한 용어를 포함시켜 주제탐색을 도울 수 있으며. 索引을 하지 않더라도 目次를 포함한 書誌 데이터 베이스를 구축하여 이용자가 원하는 경우에 目次를 通覽할 수 있도록 하는 기능을 추가시킬 수 있다.

Ⅳ. 結　　論

본 연구에서는 原文代表情報인 標題, 抄錄, 目次를 대상으로 비교평가를 실시하였다. 이를 위해 우선 표제, 초록, 목차의 특성과 역할을 살펴보고, 原文代表情報에 대한 여러 평가방법을 살펴보았다. 보다 효율적인 原文代表情報가 어떤 것인가를 파악하기 위해 代表價에 의한 평가와 검색실험을 통한 평가를 실시하였다.

代表價에 의한 평가와 검색실험을 통한 평가를 통해 標題는 情報의 壓縮性이 뛰어나지만 그 길이가 짧아서 그 결과 再現率이 다른 原文代表情報에 비해 많이 떨어진다는 것을 알 수 있었으며, 抄錄과 目次는 두 실험에서 거의 유사하거나, 목차가 약간 우세한 결과를 보여 목차가 초록과 유사하거나 좀 더 효율적인 原文代表情報라는 것을 알 수 있다.

따라서 온라인 열람 목록이나 書誌情報檢索 시스템에 목차에서 추출한 용어를 색인어로 부가하거나, 이용자가 원하는 경우 목차를 通覽하는 기능을 추가한다면 이용자의 主題探索에 도움을 줄 수 있을 것이다.

본 연구의 실험을 위하여 실험대상 문헌을 살펴본 결과 標題나 目次는 대부분의 문헌이 갖추고 있는 반면, 抄錄은 그렇지 못했다. 抄錄을 마련하는 데는 主題分野에 대한 지식뿐 아니라 抄錄作成에 대한 지식도 숙지하고 있어야 한다. 따라서 抄錄이 미리 구비되지 못한 경우 초록작성에는 많은 비용이 들 뿐 아니라 실제로 급속도로 증가하는 文獻에 대해 抄錄을 작성하는 것이 불가능할 것이며, 이는 自動抄錄에 대한 연구에서 알 수 있다. 抄錄은 거의가 문장형

태로 이루어지기 때문에 이해하기는 쉬우나 문장을 구성하기 위해 기본적으로 필요한 機能語 즉, 助詞, 語尾 등이 포함된다. 따라서 機能語도 포함해서 처리해야 하므로 自動索引을 하거나 抄錄을 데이터 베이스에 포함시키는 경우 入力에 소요되는 시간이나 기억장소, 처리시간 등에 드는 비용이 標題나 目次보다는 더 많이 들 것이다.

경우에 따라서는 抄錄을 제공하는 檢索 시스템에서도 抄錄 대신 目次를 제공할 때도 있으며, 대부분의 초록이 指示的 抄錄인 점을 감안하면 目次의 內容과 크게 다르지 않다는 것을 알 수 있다. 따라서, 구비된 정도와 처리에 드는 비용을 포함하여 각 原文代表情報의 효율을 고려해 본다면 현재 原文代表情報로 사용되지 않고 있는 目次에 대해 관심을 기울일만하다는 것을 알 수 있다.

〈參 考 文 獻〉

- 高仁哲, "情報傳達媒體로서의 抄錄에 관한 小考", 「國會圖書館報」, vol. 164, no. 3, 1983, pp. 69～79.
- 김명옥, 「자료분류법」, 서울:구미무역(주) 출판부, 1986.
- 이재철, 「주제명 목록의 연구」, 서울 : 연세대학교 도서관학과, 1959.
- 李泰榮, "韓國語 抄錄文의 文章과 內容에 관한 硏究", 「情報管理硏究」, vol. 21, no. 1, 1990, pp. 1～33.
- 정영미, 「정보검색론」, 서울 : 정음사, 1987.
- Baxendale, P. E., "Machine-Made Index for Technical Literature-An Experiment", *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354～361.
- Bloomfield, Masse, "Simulated Machine Indexing, Part 2 : Use of Words from Title and Abstracts for Matching Thesauri Headings", *Special Libraries*, vol. 57, no. 2, 1966, pp. 232～235.
- Borko, H. and Bernier, C. L., *Abstracting Concepts and Methods*, New York : Academic Press, 1975.
- Cleveland, Donald B. and Cleveland, Ana D., *Introduction to Indexing and Abstracting*, 2nd Edition, Englewood, Colorado : Libraries Unlimited, 1990.

- Cochrane, P. A., "Books Are for Use : Final Report of the Subject Access Project", In *Redesign of Catalogs and Indexes for Improved Online Subject Access*, Phoenix : Oryx Press, 1985, pp. 394～457.
- Dehart, F. E. and Matthews, Karen, "Subject Analytics and Table of Contents in Essay Collections : Implications for Searching", *Technical Service Quarterly*, vol. 6, no. 3/4, 1989, pp. 57～69.
- Diener, Richard A. V., "Informational Dynamics of Journal Article Titles", *Journal of ASIS*, vol. 35, no. 4, 1984, pp. 222～227.
- Diodato, Virgil, "Tables of Contents and Book Indexes : How Well Do They Match Reader's Descriptions of Book?", *Library Resources and Technical Services*, vol. 30, no. 4, 1986, pp. 402～412.
- Edmunson, H. P., "New Methods in Automatic Extracting", *Journal of ACM*, vol. 16, no. 2, 1969, pp. 265～285.
- Guinchart, C. and Menou, M., *General Introduction to the Techniques of Information and Documentation Work*, Paris : UNESCO, 1983, 한국어 역본 :「정보관리론」, 사공 철·김태수 공역, 서울 : 구미무역(주)출판부, 1987.
- Jones, L. P., et al., "INDEX : The Statistical Basis for an Automatic Conceptual Phrase－Indexing System", *Journal of ASIS*, vol. 41, no. 2, 1990, pp. 87～96.
- Luhn, H. P., "The Automatic Creation of Literature Abstracts(Auto－Abstracts)", *IBM Journal of Research and Development*, vol. 2, no. 2, 1958, pp. 159～165.
- Maloney, R. K., "Title versus Title/Abstract Text Searching SDI System", *Journal of ASIS*, vol. 25, no. 6, 1974, pp. 370～373.
- Mandel, Carol A., "Enriching the Library Catalog Record for Subject Access", *Library Resources and Technical Services*, vol. 29, no. 1, 1985, pp. 5～15.
- Markey, Karen, "Users and the Online Catalog:Subject Access Problems", In *The Impact of Online Catalogs*, Edited by Joshep R. Matthews, New York : Neal－Schuman Publishers, 1986, pp. 37～38.
- _____, "Subject Experiences and Needs of Online Catalog Users : Implications for Library Classification", *Library Resources and Technical Services*, vol. 29, no. 1, 1985, pp. 34～51.
- Waldstein, Robert Kenneth. "The Role of Noun Phrases as Content Indicators", Ph. D. Dissertation, Syracuse University, 1981.

〈부록 1〉　　代表價 算出實驗을 위한 實驗對象 文獻 리스트

- 「정보학의 개념과 교육의 기본방향에 관한 연구」, 「情報管理學會誌」, vol. 5, no. 2, pp. 127～144.
- 「구문 및 의미분석을 통한 한국어 자동색인」, 「情報管理學會誌」, vol. 8, no. 2, pp. 96～107.
- 「온라인 데이터 베이스 탐색자의 탐색행태에 관한 연구」, 「情報管理學會誌」, vol. 8, no. 2, pp. 32～73.
- 「연구활동과 과학지식 생산성에 있어서 과학연구 전산망의 역할」, 「情報管理學會誌」, vol. 7, no. 1, pp. 96～120.
- 「언어학적 분석기법에 의한 신문기사 자동색인 시스템 설계에 관한 연구」, 「情報管理學會誌」, vol. 9, no. 1, pp. 78～97.
- 「하이퍼 텍스트의 개념과 응용에 관한 고찰」, 「情報管理學會誌」, vol. 6, no. 2, pp. 3～19.
- 「대형 컴퓨터를 이용한 도서관 업무의 토털 시스팀 설계」, 「情報管理學會誌」, vol. 4, no. 2, pp. 2～30.
- 「연구활동에 있어서의 비공식 커뮤니케이션」, 「情報管理學會誌」, vol. 1, no. 1, pp. 127～145.
- 「우리나라 학술잡지 발달과정 연구」, 「情報管理學會誌」, vol. 6, no. 1, pp. 3～13.
- 「문헌정보학 영역 지식기반 시스템에서의 지식표현」, 「情報管理學會誌」, vol. 7, no. 2, pp. 35～57.