주제분석기법으로서의 자동색인

이 영 자*

〈목 차〉

Ⅰ. 서 언

입수한 자료중에서 필요한 것만을 선택하여 그 내용을 분석하고 그 결과를 초록, 분류, 색인으로 작성하여 어떤 형태의 매체에 축적하여 두고 이용자의 정보요구에 대응하는 것이 정보관리업무의 핵심을 이루고 있다고 할수 있다. '현대의 정보검색 시스템에 있어서 노력의 80%(경비 75%)는 초록, 색인과 같은 이차정보작성에 소모되고 있으며[1] 이차정보작성의 성공여부는 주제분석의 일관성과 정확성에 달려있다고 볼때, 이의 중요성은 아무리 강조해도 지나치지 않을 것이다.

그런데도 불구하고 주제분석을 통정하는 명확한 원리원칙이 없을뿐만 아니라, 주제분석분야보다 편목에 대한 방법들의 진보가 훨씬 앞서고있어, 공인할 수 있는 편목업무를 위한 기준의 개발을 가져 올 수 있었으나 고도로

* 경북대학교 사회과학대학 도서과학과 부교수

1) Lancaster, F.W. "Mechanized Document Control: a Review of Some Recent Research," Aslib Proceeding, Vol,16,No.4,1964,pp.132-152. 情報檢索論. 서울 : 아세아출판사, 1977, p.18에서 재인용.

체계화된 주제분석의 기준에 대한 개발은 아직 요원한 상태에 놓여있다.[2]

이 점을 고려하여 Liston Howder는 1974년에서 1976년사이의 주제분석에 관련된 문헌 206편[3]을 수집, 검토하였고 Travis는 1977년에서 1981년사이의 205편[4]의 문헌을 수집, 검토하여 주제분석에 관한 원칙들의 수립을 위한 기초작업으로써 연구동향을 파악하려고 하였으며 그 한가지 두드러진 경향은 자동화가 주제분석에 미친 영향이라고 지적하고 있으며 특히 자동색인방법들이 주제분석의 새로운 기법으로 등장한 점이라고 지적하고 있다.

따라서 본 연구는 주제분석의 새로운 기법으로서의 자동색인방법의 중요성과 아직 이 방면에 관한 포괄적 연구가 이루어지지 않고 있는 실정을 인식하여 그 방법들을 각각 통계적 분석방법, 구분적 분석방법, 의의적 분석방법으로 나누어 개관하였다. 이는 어디까지나 선행연구들을 검토,종합하여 얻어진 개관으로서 각각의 방법에 대한 보다 구체적이고 상세한 연구와 그 적용사례는 앞으로의 연구과제로 남겨두었다.

Ⅱ. 주제분석과 자동색인

1. 주제분석의 정의 및 변천양상

1) 주제분석의 정의

'주제분석(主題分析)은 초록이나 색인을 작성하기 위하여 문헌의 요점을 가려내는 작업으로 그 결과를 자연어의 문장으로 표기하면 초록이 되고, 소정의 기호로 표기하여 소정의 순서에 따라 배열하면 색인이 되는 것'[5]이라고 일반적으로 정의될 수 있다. 즉 새로 입수한 기록자료에 포함된 정보

2) Liston, JR., David M., and Howder,Murray L., "Subject Analysis," ARIST. Vol.12, 1977, p.82.
3) 上揭書 pp.105-118.
4) Travis, Irene L., "Subject Analysis," ARIST. Vol.17,1982,pp.144-157.
5) 菊池敏典 "主題分析", 情報管理, Vol.9, No.9,1966,p.460, 司空 哲 情報檢索論. p.18에서 재인용.

에 대한 이용자의 요구가 발생할 때, 그 정보를 용이하게 확인하기 위한 방법으로 정보를 조직해야 하고, 그 조직의 결과가 초록이나 색인 등이 되며, 초록이나 색인의 내용은 기록자료의 개념분석결과라 할 수 있다.

힉키(Doralyn J. Hickey)는 주제분석(subiect analysis)의 정의를 '메시지 원(message source)의 의도를 정확하게 추론할 가능성이 가장 큰 메시지의 속성들을 확인해내는 과정'[6]이라고 설명하였다. 그리고 그는 그 속성들을 확인해내기 위하여 메시지를 이루고 있는 단어, 단어군(單語群), 단어빈도수를 분석하는 요소분석(elemental analysis)과 이 요소들의 관계를 더 관심대상으로하는 구조분석(structural analysis)로 구분할 수 있는데 이 중에서 요소분석이 주제분석의 중심을 이루고 있다[7]고 피력하고 있다. 특히 불용어 리스트(stop-words list), 어간사전(語幹辭典)작성, 단어빈도 측정기법, 문법적 문장분석(parse), 구문적 분석 등을 컴퓨터로 처리하는 것이 이 범주에 속하는 영역이라[8]고 다시 부언하고 있어 이는 기록정보의 개념분석의 과정을 기계에 의하여 수행하는 측면을 부각시키는 정의라 할 수 있겠다.

수작업에 의한 주제분석의 정의나 기계에 의한 주제분석의 정의의 기본 개념은 동일하다고 볼 수 있다. 즉 둘다 원기록자료에 대하여 이용자의 정보 요구발생시의 접근점(access point)을 제공해주기 위하여 행하여지는 과정이라는 점에서 공통점을 갖고있다. 따라서 주제분석의 문제점은 여기에서 유래된다고 볼 수 있는데 즉 이용자가 어느 관점에서 특정탐색을 이행할 것인가를 예측하기가 매우 힘들기 때문이다.

즉 주제분석가는 저자가 내용변수에서 선택한 가치 즉 메시지를 발견해야 하는데, 각기 다른 인간들이 같은 자료에 직면하더라도 같은 가치(意味)를 선택하기를 기대할 수 없기 때문에, 주제분석의 일관성 유지가 어렵다는 말이 될 수 있다.

6) Hicks, Carol E, et al., "Content Analysis," in Belzer, Jack et al., Encyclopedia of Computer Science and Technology, New York: Marcel Dekker, Inc., 1976, March, p.101.in David M. Liston, JR., "Subject Analysis," p.88 에서 재인용
7) 上揭書, 同面.
8) 上揭書, 同面.

켄트(Kent)는 주제분석에 대한 이러한 문제점들을 주제분석의 일반론으로 집약했다. 즉 '첫째 아무리 상세한 주제분석이라 할지라도 개개인의 정보요구에 대한 완전한 탐색을 충족시켜줄만큼 망라적이지 못하다. 둘째 어떠한 주제분석도 특정 기록문헌에 대하여 요구될 수 있는 모든 가능한 관점이나 이용을 다 예측할 수는 없다'[9]고 한 것이 바로 그 일반론인 것이다.

2) 주제분석의 변천양상

기록문헌과 이용자를 연결시키기 위한 방법, 즉 자료의 분석 및 조직방법의 역사는 인류의 문자사용의 시기로 거슬러 올라갈 수 있으나, 1876년 Cutter의 '사전체목록규칙'의 출현 이전까지는 거의 전부가 문헌(原籍)의분석, 즉 저자명, 서명, 출판사항, 대조사항 등을 중심으로하는 자료의 외형적 요소의 분석에 머물러 있었다.

1876년에 간행된 Cutter의 사전체목록규칙은[10] 도서관의 전통적인 저자명 및 서명목록에 분류번호순의 분류목록대신 알파벳순의 주제명목록의 사용을 제안한, 소위 저자명, 서명 및 주제명을 혼합한 사전체목록시스템으로서, 특히 161조에서 188조까지의 주제명목록규칙은 이른바 주제색인의 이론과 발전을 위한 토대를 마련해 주었다고 볼 수 있다.[11]

1876년 이후의 수년간의 주제분석은 거의 대부분 주제목록(subject cataloging)과 분류(classification)이라는 본질적으로 분리된 두 개의 분야로 나누어지게 되었다.[12] 주제목록은 자료의 내용을 기술(記述)하고 있는 적절한 용어의 선정에 관한 연구로서, 논리적으로는 記述目錄에 대한 노력과 상당한 관련을 갖고있고, 분류는 근본적으로 자료를 의미있는 방법으로 유별(類別)하려는데 관심을 둔 것이기 때문이다. 따라서 1945년,

9) Kent, Allen, Information Analysis and Retrieval. New York: Becker and Hayes, Inc., 1971, p.94.
10) Cutter, C.A. Rules for a Dictionary Catalog. 4th ed.Washington: Government Printing Office, 1904.
11) 윤구호, "주제색인의 이론과 실제", 圖書館學論集. 제 10 집, 1983, p.104.
12) Hickey, Doralyn J., "Subject Analysis: An Interpretive Survey," Library Trends, Vol.25, No.1, (1976. July), p.273.

제 2 차세계대전이 끝날때까지 주제분석을 통정하는 원리 원칙은 주로 주제명표목표(Sears Subject Headings과 LA Lists)와 분류표들(DDC, LC, UDC, CC등)에의존하였는데 이들은 시대에 뒤떨어지고, 일관성이 결여되고 인종, 종족, 종교, 성(性)에 대한 편견을 많이 포함하고 있는 등 주제분석을 위한 훌륭한 기준이 될 수 없었다. [13]

1945년 이후의 두드러진 정보폭발현상은 주제분석 방법상에 변천을 강요하였고, 주제명표목이나 유(類)의 선정이 보다 신속한 시간에 이루어져야하고 반면에 주제분석의 심화 및 정확성에 대한 요구는 증대되었다. 또한 폭발하는 정보의 이용은 더이상 저자명이나 서명으로 접근하는 일이 점점 어렵게 됨으로해서, 보다 용이한 접근을 가능하게하는 방법으로 主題索引이 고안되기에 이르렀다.

주제색인은 전통적인 자료조직방법에 근거하여 색인작성시에 색인엔트리를 확정하여 탐색시에는 그 색인엔트리로서만 접근할 수 있는 전조합색인시스템(前組合索引시스템)의 발전과*1이용자의 多元的接近을 도모하기 위하여 색인시에는 단위개념으로 색인엔트리를 정하고, 탐색시에 조합할 수 있게하는 후조합색인시스템(後組合索引시스템)의 발전으로*2그 양상을 들어내게 되었다.

또한 특정 정보에 대한 정확한 주제분석에 필요한 주제배경을 갖춘 전문적 주제분석가의 부족에 대한 대안으로 1959년의 Luhn에 의한 키-워드 색인*이 탐색의 접근점으로서는 부정확하고 일관성이 결여되어 있어, 이를 극복하기 위한 방법으로 어휘통제(vocabulary control)에 대한 연구가 이루어졌고 이의 결과로써 디소러스(thesaurus)가 개발되기에 이르렀다.

한편 1950년대에 들어와서 컴퓨터기술이 발전한 것과 병행하여 주제색인의 여러가지 난해성을 컴퓨터로 해결시켜보려는 연구와 시도와 실험들이 이루어지게 되었다. 정보수명의 단축으로 인한, 신속한 주제분석의 필요성, 정보폭발로 인한 색인대상의 방대한 량, 주제전문가의 부족, 주제분석의 일관

13) Hickey, Doralyn J. 上揭論文, p.283
*1 Kaiser, Ranganathan, Coates, Lynch, Austin 으로 이어짐
*2 Mortimer Tauber의 uniterm system이 그 효시가 됨
* KWIC 색인

성 결여, 다원적인 탐색점 제공의 필요성 등의 여러 복합적 요인은 주제분석의 자동화에 관심을 갖도록 한 필연적인 동기가 되었다고 할 수 있다.

즉 최초의 주제분석은 자동색인과 자동초록, 그리고 자동문헌분석을 중심으로 발전되고 있다고 볼 수 있다.

2. 주제분석의 요소 및 자동색인

Liston과 Howder는 1974년에서 1976년 사이에 발표된 주제분석에 관련된 문헌들을 분석하여 〈그림 1〉[14]와 같은 주제분석 요인들을 확인해내었다.

〈그림 1〉에서 알 수 있듯이 주제분석과정을 크게 ① 문헌평가와 선정 ② 주제분석 ③ 색인작성 방법들 ④ 어휘관리 ⑤ 자동언어처리로 구분하고 있다. 주제분석은 다시 개념파악, 개념평가 및 선정, 개념표현, 개념간의 상관관계로 세분되고, 색인작성방법들은 분류, 주제명표목 및 색인작성, 단위개념색인작성, 인용색인작성, 전문(全文)색인작성(탐색), 데이타 플레깅/태깅(data flagging/tagging)으로 세분되어 있다. 이 가운데 단일항목으로는 단위 개념색인작성에 관한 연구가 88건으로 가장 많이 연구되었고, 다음은 어휘관리가 68건, 분류가 41건 순으로 나타나 있다.

한편 주제분석에 관련된 여러 측면을 주제분석과정과 연결지어 분석하고 있는데 여기에는 시스템내(intra-system)의 측면으로는 일관성, 분석과 표현의 심도, 색인언어 및 어휘통제의 문제들이 관련되고 있고, 시스템간(inter-system)의 측면에는 표준 및 표준화, 양립성, 전환성 협력문제가 관련되어있다. 분석과정평가 측면에는 조사, 검토, 문헌에 의거한 평가, 평가방법, 평가비교, 비용대 효과 문제들이 연루되어있고, 기계화 측면에는 자동화, 컴퓨터 조력 문제들이 관련되고 있다. 주제분석과정의 연구 및 개발 측면에는 가설 및 오피니언, 이론, 원리, 실험, 발전적 시스템 연구가 포함되어있고, 마지막으로 교육 및 훈련의 측면이 지적되고 있다.

14) Liston, JR., David M. 前揭書, p.83.

ASPECTS OF THE PROCESSES / THE PROCESSES			QUALITY AND/OR CONTROL OF THE PROCESSES														EVALUATION OF THE PROCESSES				MECHANIZATION		RESEARCH AND DEVELOPMENT OF THE PROCESSES						Total Number of Documents on the Subject
			INTRA-SYSTEM ASPECTS										INTER-SYSTEM ASPECTS																
						INDEXING LANGUAGE/ VOCABULARY CONTROL																							
							THESAURI																						
			Consistency	Depth of Analysis	Depth of Representation	Natural or Uncontrolled	Intra-Language Aspects	Inter-Language Aspects	Creation	Updating and Maintenance	Syntactic Control	Standards/Standardization	Compatibility	Convertability	Cooperation	Standards/Standardization	Surveys/Reviews/Texts	Evaluations	Comparisons	Cost/Benefit Studies	Computerization—Automation	Computer Assistance	Hypothesis/Opinion	Experiments/Studies	Theory/Principles	Developmental Systems	Education/Training	
			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
Document Assessment and Selection		1																1							2			2
SUBJECT ANALYSIS	Concept Recognition	2															5		2		7	1		1	5	1		21
	Concept Evaluation and Selection	3															5		1		6	1			3	1		16
	Concept Representation	4									1						1		1		1				2	2		8
	Concept Interrelation	5									3						5	1	2		5	1		1	5	2		19
INDEXING METHODOLOGIES	Classification	6	2														16	4	1		9			5	15	3	2	41
	Subject Heading Indexing	7	2					1	5		1						2	2	1	1	1	1		1	1	1		16
	Unit Concept Indexing	8	2					3	11		6						11	7	4		32	11		14	21	2	2	88
	Citation Indexing	9																			2				2			4
	Full Text Indexing (Searching)	10															1	2	2									3
	Data Flagging/Tagging	11																1						1				1
Vocabulary Management		12					1	7	21	1	2	1	5	5	1	1	7	8	3		8	4		1	11	2		68
Automatic Language Processing		13									2						1	2	2		9	2		9	17		1	29
Total Number of Documents on the Subject			2				1	7	25	1	11	1	5	5	1	1	36	24	11	1	52	18		29	62	8	4	

그림 1. 주제분석의 요소들

각 주제당 문헌수에 있어서는 자동화에 관한 연구가 52건으로, 연구·개발 측면의 이론 및 원리 연구의 62건 다음으로 많이 연구되고 있음을 알 수 있으며, 특히 단위개념색인작성의 자동화 문제가 32건으로 주제분석 요소가운데 가장 중요한 비중을 차지하고 있는 것으로 나타나 있다.

이러한 현상은 앞에서도 언급한 바와같이, 주제분석가들이 검색 목적으로 중요하다고 생각되는 것으로서의 주제를 각기 다르게 선택할 수 있는 가능성과, 같은 주제를 선택했을 경우에라도 다른 용어를 사용해서 표현할 수 있는 가능성으로 인한 일관성 결여가 나중에 탐색실시를 위하여 상당한 불확실한 결과를 초래할 것이 문제점이 되고 있어, 기계만이 절대적인 일관성을 보증할 수 있다(물론 이 일관성 있는 분석만이 반드시 요구충족의 최선을 제공한다고 볼 수는 없지만)는 데에 연유하고 있다고 볼 수 있다.

그러나 인간의 지적활동을 기계가 분석하는데에는 아직 상당한 한계성을 가지고 있는바 앞으로도 계속하여 이 복잡하고 난해한 주제분석의 향상을 위한 갖가지 연구와 실험이 시도되리라고 믿는다.

Ⅲ. 주제분석기법으로서의 자동색인

'자동색인의 목적은 검색시스템에서 사용하기에 적합하도록 문헌의 압축적인 표현을 성취하려는 것이다. 보수적인 관점에서 본다면 문헌이란 이은데가 없는 전체이며 그 전체를 나타내기 위하여 부분들을 추출하는 것은 그 본내용을 어느 정도 파괴할 것이다. 그러나 최근의 검색시스템은 본문에서 추출한 비교적 소수의 조각들이, 그것이 단어들이건 구들이건, 검색에 훌륭한 결과를 성취시킬 수 있다는 가정위에 운영되고 있고 실제적인 경험과 광범위한 실험이 이 견해를 지지해주고 있다. 그러므로 자동색인은 본문에서 내용어들이나 구절(content bearing words or phrases)을 선정하기 위하여 통계적인 혹은 언어학적인 기준을 사용하여 문헌의 주제를 분석, 표현하려고 한다.'[15]

15) Dillon,Martin, "FASIT:a Fully Automatic Syntactically Based Indexing System," Journal of American Society for Information Science,34(2):1983,p.99.

따라서 주제분석의 기법으로서의 자동색인에 대하여, 통계적으로 분석하는 접근방법에 있어서는 문헌의 요소를 분석하여 확인해 내는 기준설명과, 색인어를 산출해내는 각종 통계적 수학 공식을 검토하며, 구문적, 의의적분석방법에 있어서는 각각 그 의미와 이를 위한 선결조건으로서의 각종 사전에 대하여 살펴보려고 한다.

1) 통계적 혹은 빈도분석방법

색인작성의 통계적 방법이 근거하고 있는 가설은 어느 문헌에 어느 단어가 많이 사용되면 될수록 그 단어는 문헌의 주제의 지시자(indicator) 이기 쉽다[16]는 것이다. 이 가설에 근거하여 컴퓨터 프로그램은 특정문헌의 모든 단어들을 리스트하여 출현빈도에 따라 모아주고 빈도순 이내에서 알파벳 배열을 한다. 기능어(function words : 관사, 접속사, 전치사, 그리고 대명사)는 보통 제외되며 같은 어간을 가진 단어는 같은 단어로, 혹은 다른 단어로 헤아려 질 수 있다.

단어를 헤아리는 것이 모든 자동색인작성기법의 기초이며, 가장 단순한 수준의 방법으로는 특정한 최소빈도보다 더 많이 사용된 모든 단어들을 열거하여 색인용어로 선택하도록 프로그램된 것이다. 즉 특정문헌의 전 내용으로부터 주제 내용을 기술하고있는 일조(一組)의 키-워드를 확인해내는 업무는 자동화된 색인작성 및 정보검색시스템에서 필요한 단계로서 ① 이렇게 확인된 키-워드가 바로 엔트리 포인트로 사용되는 경우가 있고 ② 이들을 하나의 베이스로 사용하여 그 키-워드들을 전거파일에 있는 요소들로 변환시켜서 색인어로 사용하거나, 혹은 용어절단을 행하거나 혹은 부울리언 논리를 적용하거나 혹은 상관계수기법을 활용하여 색인작성을 하게된다.[17]

16) Borko, Harold and Bernier, Charles L., Indexing Concepts Concepts and Methods, New York : Academic Press. Inc., 1978. p.115.

17) Harter, p. Stephen, "Statistical Approaches to Automatic Indexing," Drexel Library Quarterly., Vol.14, No.2, (April 1978), pp.57.

어휘빈도분석에 의한 자동색인 작성은 두단계 문제로 간주될 수 있으며, 첫째 특정 학문분야의 문헌을 나타내는 특성으로서의 전문적 어휘(technical vocabulary)를 확인해내는 단계와, 둘째 그 학문분야를 이루고있는 개개 문헌의 어휘의 요소(vocabulary element =주제명표목, 색인용어,혹은 디스크립터)를 확인해내는, 즉 개개 문헌의 키-워드를 확인해내는 단계이다.[18]

Harter는 이 두 단계에서 문헌군의 요소와, 개개 문헌의 키이워드를 자동적으로 선정하는 몇가지 기준을 제시하고 있다.

1) 전문적인 성격을 띤 문헌군의 요소들을 자동적으로 선정하는 기준[19]

첫째 기준 : '문헌의 어휘에서의 디스크립터는 이용집단의 이용자가 요구할 가능성이 큰 용어라야 한다.'

여기에 관련된 문제는 특정용어가 이용자에 의하여 선정될 확률의 평가인데 이 확률평가를 위하여 과거의 탐색역사가 유지되어 있어야하며, 또한 문헌통계로부터 용어사용정도와 이용자 탐색의 확률간에 존재하고있는 직접관계를 추정하여 얻을 수도 있다. 여기에서 두번째의 기준이 얻어진다.

둘째 기준 : '전문적인 어휘의 요소들은 그 문헌군내의 문헌들에서 빈번히 다루어지고 있는 주제들을 명명(名命)하는 것이라야 한다.'

왜냐하면 문헌에서 빈번히 다루어지고 있는 주제들은 실제적인 연구흥미를 나타내기 때문이다.

셋째 기준 : '전문적인 어휘의 요소들은 그 문헌군내의 문헌들에서 드물게 사용된 용어라야 한다.'

즉 드물게 발생하는 용어들이 그 희소성(rarity)때문에 본질적으로 보다 더 가치있는 것이라고 생각되며, 주제가 드물게 취급되면 그 특정문헌이 그 주제를 부가적으로 중요하게 취급하는 것이 되며, 따라서 그 주제들을 나타내는 용어들이 색인작성시에 우선권이 주어져야 한다는 것이다.

---

18) Harter, 上揭論文, p.58.

19) 上揭論文 pp.60-64.

넷째 기준: '전문적인 어휘의 요소들은 문헌내에서 너무 빈번히 취급되지도 않고, 너무 드물게 취급되지도않는 주제들을 표현하고 있어야한다'.

룬(H.P.Luhn)은 [20] 언어데이터처리에 관한 논문에서 요소단어는 빈도수가 너무 큰 것도, 너무 작은 것도 아닌 중간 정도의 것이라는 것을 데이타·베이스내의 단어들의 총 빈도수 계산에 의하여 밝혀냈다. 이러한 룬의 제안은 문제점이 있는데 하나는, 단순히 높은 빈도의 혹은 낮은 빈도의 용어들을 제거함으로써 재현율 및 정도율의 저하를 가져올 수 있다는 점과, 다른 하나는 한계빈도를 결정하기가 어렵다는 점이다.

파오 [21] (Miranda Lee Pao)는 단어빈도 형태에대한 지프의 법칙(Zif's Law)을 검토하고 특정 텍스트(text)에 있어서 내용어(centent bearing words)에 대한 빈도분포의 낮은 기준치(a lower threshold)를 확인해내는 방법을 제시하였다.

이 개념은 맨처음 고프만(Goffman)이 [22] 제시했고 Goffman은 다음과 같은 과정에 의하여 색인어들이 될 수 있는 용어의 범위설정을 정할 수 있는 방법을 제시했다.

Goffman은 [22] 지프의 첫번째 법칙($r \times f = c$. r은 빈도, f는 텍스트내에서 그 단어가 사용된 수, c는 주어진 텍스트를 위한 상수(常數))이 작용하는 조건에는 단지 고빈도 단어의 발생만이 고려된다는 것을 관찰했다. 즉 특정 텍스트내에서의 고빈도 단어들에 있어 어떤 두개의 단어도 동일한 빈도를 가지지 않는다. 다시말해서 단 한개의 단어만이 가장 빈도가 높고, 두번째 빈도가 높은 단어도 한개이며, 이렇게 계속된다. 그래서 개정된 두번째의 법칙($I_1 / In = (4n_{2-1})/3 : In \rightarrow$ 그 텍스트에서

20) Luhn, H.P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No.4, 1957, in Harter, 上揭論文, p.58. 에서 재인용.

21) Pao, Miranda Lee, "Automatic Text Analysis Based on Transition Phenomena of Word occurrence," Journal of American Society for Information Science, Vol.29, No.3, 1978, p.121.

22) Miranda Lee Pao와 Goffman 과의 personal communication이 있었음. Miranda Lee Pao, p.121.

n 번째 발생하는 상이한 단어의 수, I,→그 텍스트에서 1번밖에 나타나지않는 상이한 단어의 수)은 발생의 저빈도 단어들만을 설명하고 있다.

많은 저빈도의 상이한 단어들은 동일한 빈도를 갖고있음이 나타났다. 이 두 개의 완전히 다른 규칙들은 어떤 특정 텍스트내의 단어분포에 두가지 극단을 예측하고 설명하여준다. 그래서 그것들은 고빈도로부터 저빈도현상이 발생하는 단어형태의 전환에 있어 중요한 영역을 추측케해준다. 더 나아가 그는 전환영역(transition region)은 특정 텍스트를 위한 내용을 가장 많이 포함하는 단어들이 속해있을 것이라고 가정했다. 그 전환점을 얻기 위하여 저빈도 단어들의 공식($I_1/In = n(n+1)/2$)이 고빈도 단어들의 특징을 띄기시작할 것이며, n빈도를 가지는 단어들의 수가 하나의 통일된 지점으로 접근하게 될 것이다. 결국 $I_1/1 = n(n+1)/2$ —이를 다시 정리해보면 $n^2+n-2\,I_1/1=0$.

이를 다시 풀어보면 $n=(-1\pm\sqrt{1+8\,I_1/1})/2$, 여기에서 단지 양수의 값에만 관심이 있으므로 $n=(-1+\sqrt{1+8\,I_1/1})/2$가 되고 여기에서 n값만 계산하면 쉽게 전환영역 n의 단어들을 확인해 낼 수 있을 것이다.

즉 559개의 상이한 단어들을 포함하고있는 텍스트에서 한번밖에 발생하지 않는 단어가 256개일때 n값은 22로 나타나, 논문에서 22번가량 발생하는 단어들 주변에 고빈도와 저빈도 성격의 전환영역이 이루어지고 이 영역내에서 기능어를 제외시킨 단어들을 색인어로 선정할 수 있다.

다섯째 기준 : '전문적 어휘의 요소는 문헌집성의 여러잡다한 구성원들간에 「구별」을 지을 수 있어야 한다.'

이 기준은 샐톤[23]등의 문헌분리가 (discrimination value)에 의한 법칙에 근거하는 색인어 선정에서 적용되고 있다 하겠다. 즉 문헌분리가가 높은 색인어(좋은 색인어)는 문헌들을 분리시켜 문헌집단의 밀집도를 낮추므로 이웃 문헌들로부터 쉽게 구별하게 해주며, 문헌분리가가 낮은 색인

23) Salton, G. and Yang, C.S., "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, Vol.29, No.4, (December 1973). pp.351-372.

어(나쁜 색인어)는 문헌들을 함께 밀집시키므로 구별을 어렵게 한다고 했다.

문헌분리가를 측정하는 공식은

$DVk = \overline{SIMk} - \overline{SIM}$ [24] 가 되며 이때 값이 양수이면 좋은 색인어, 음수이면 나쁜 색인어로 판정된다.

이상에서 자동적으로 한주제분야의 전문적인 어휘요소를 선정하는 기준에 대하여 살펴보았으며 이 기준들의 하나 하나에 대한 구체적인 검토와 실험이 이루어져야 할 것이며 이 기준들간의 관계에 대한 연구도 이루어져야할 것이다.

2) 개별 문헌의 키-워드를 선정하는 기준[25]

첫째 기준: '하나의 문헌에 대한 키-워드로 선정되는 단어들은 그 문헌에서 취급되고 있는 주제를 명명(名命)하고 있어야 한다.'

즉 한 문헌에서 개념을 표현하는 단어의 빈도수가 그 문헌의 주제의 취급정도(degree of treatment)를 반영한다는 가정이 나올 수 있다. 그러나 이러한 절대개념(absolute concept)으로서의 취급정도는 색인작성 목적으로 반드시 매우 유용하다고 말할 수 없고, 따라서 '취급의 상대적인 정도'라는 개념이 나타날 수 있으며, 이때 '상대적'(relative)인 것의 비교의 표준(standard of comparison)이라는 문제에 당면하게 된다. 여기에는 ① 특정 문헌에서 취급되고 있는 하나의 주제를 같은 문헌에서 취급된 다른 여러 주제개념들과 비교함으로써 그 정도를 알아낸다. (단어빈도 측정) ② 특정문헌이 특정주제를 다루고 있는 정도를 다른 문헌들이 동일 주제를 취급하고 있는 정도와 비교하여 본다. (문헌빈도 측정)

둘째 기준: '한 문헌의 색인레코드로 선정되는 키-워드는 그 문헌에서 가장 중요하게 취급된 주제를 명명하는 것이라야 한다.'

24) 이태영, "계량군집색인의 모형제시", 정보관리연구, Vol.16, No.3,(1983), p.73.

25) Harter, Stephen P. 上揭論文, pp.64-67.

이 개념은 마론(M.E.Maron)[26]의 "R-aboutness"라는 견해와 관련되는데, 특정용어에 관한 특정문헌의 R-aboutness란 이용자가 그의 정보요구를 충족시켜줄 것이라고 생각하여 그 용어아래에서 요구정보를 탐색할 확률을 말한다. R-aboutness가 주제의 취급정도와 관련이 있고 '취급정보'가 그 문헌에서 발생하는 개념을 표현하는 용어의 빈도수에 비례한다는 두개의 가정이 성립한다면 이 측정법에 의한 문헌의 색인 레코드는 색인어휘(index vocabulary)에 상관없이 문헌내용으로부터 직접 추출한 단어들이 될 것이며, 그래서 문헌내의 발생빈도(within document frequency of occurrence)에 따라 순위를 정하고 불용어를 제거한 후의 가장 빈번히 발생한 단어부터 맨꼭대기에 배치하고 둘째기준에 따라 절단점을 정하여 그 이상의 빈도수를 가진 단어들이 색인용어로 선정될 수 있다.

셋째 기준 : '특정 키-워드에 의하여 색인된 문헌은 가장 빈번하게 키-워드가 명명한 주제를 다루고 있는 문헌일 것이다.

이것은 둘째 기준과는 아주 다른 '취급정도'의 개념을 내포하고 있다.

즉, 셋째기준은 같은 주제를 취급하고 있는 '여러 다른 문헌들의 비교에 관심이 있는데 이 개념의 공리는 특정 문헌이 유용하다는 것은 그것이 다른 문헌에 비하여 동일 주제를 더 많이 취급하고 있기 때문이라는데 있다 하겠다.

자동색인에 있어서 기준 3은 한 문헌집성의 통계적 특성을 고려해야 하고 문헌내(within documant)의 특성도 고려해야 함을 시사해주고 있다.

둘째, 셋째 기준간의 모순점이 어떻게 해결되어야 하는가는 아직 연구되어야 할 영역으로 남아있다 하겠다.

넷째 기준 : '한 문헌의 키-워드로 선정된 단어들은 어떤 의미에서 그 문헌의 검색의 가능치를 극대화해야 한다.(또는 가능 손실치를 극소화해야 한다)

---

26) Maron, M.E., "On Indexing, Retrieval, and the Meaning of About," Journal of the American Society for Information Science. Vol.28, (January,1977) pp.38-43. Stephen P. Harter. 前揭論文, p.66에서 재인용.

이것은 색인작성과 검색행위의 관계를 요구하며, 색인작성과 검색행위는 의사결정과정으로 간주되나 이 분야에 대한 의사결정의 이론적 방법은 아직 매우 이론적으로 치우쳐 오고 있고 자동색인 작성에 적용할 수 있는 잘 정의된 이론이 개발되어 있지 않고 있다[27]

이상에서 Harter가 제시한 자동색인을 위한 전문분야 어휘요소 확인 기준과 개별 문헌의 키-워드 확인 기준을 살펴 보았다.

Harter의 기준을 구체적으로 실현하거나 Harter의 기준에서 언급되지 않은 것으로, 자동색인의 통계적 접근방법에 있어서 단순한 빈도측정의 불완전성을 보완하여 보다 효과적으로 색인어를 선정하는 기법으로 단어에 가중치를 부여하는 방법, 부호-잡음측정(signal-noise calculation)방법, 그리고 상관계수(association value)를 이용하는 방법등이 있다.

1) 가중치(weighting)를 부여하는 방법

이것은 단어의 중요도에 따라 가중치(加重値)를 부여하여 색인어를 선정하는 것으로 다음과 같은 방법[28]이 있다.

첫째, 문헌에 나타나는 단어의 위치, 즉 제목명, 각 문장의 처음, 끝 등의 위치에 따라 가중치를 부여하는 방법으로, 본문에 나오는 단어보다 표제속에 나타나는 단어에 더 높은 가중치를 부여하는 등의 방법이다.

둘째, 특정 단어의 빈도수에따라 가중치를 부여하는데, 이 또한 발생빈도가 높은 단어가 그 문헌의 주제를 잘 나타내리라는 가정에서 출발하고 있다.

셋째, 문헌빈도가 낮은 단어에 높은 가중치를 부여하는 방식[29]으로 이는

27) Harter, Stephen p. 前揭論文, p.69.
28) 송미연, "자동색인방법과 자동색인시스템성능", 정보관리연구, Vol.17, No.3 (1984.9) p.8.
29) Jones, K.Spark, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," Journal of Documentation, Vol.28, No.1, 1972, pp.11-21.

색인어의 특정성을 문헌의 수와 관계가 있으리라는 개념에서 출발한다 하겠다. 구체적인 가중치부여방법으로는 샐톤의 단어분리가중기법, 스파크·존스의 역문헌 빈도 가중기법, 하트의 2-포와손 가중기법, 수정된 하트의 2-포와손 가중기법, 룬과 샐톤 등의 TS가중기법 등이 있다. [30]

2) 시그널-잡음계산(Signal-noise calculation)에 의한 기법

정보이론(information theory)에서는 전달 가능한 메시지의 수에 따라 한 메시지가 전달하는 평균정보량이 계산된다. 이러한 정보이론의 통계적 특성은 색인언어에도 그대로 적용된다. [31] 하나의 학문분야는 10개의 하위 주제로 나누어질수 있고, 20개, 30개, 100로도 나누어질수 있는 것이며 주제를 세분할 수록 색인어의 수는 증가되고 그만큼 색인언어의 특정성은 커진다. 정보이론적으로 특정한 색인언어가 갖는 색인어휘의 수가 많을수록 문헌의 주제를 자세하게 색인할 수 있고 이 색인언어가 전달하는 문헌내용에 관한 정보는 보다 구체적이라고 할 수 있다.

이에 대한 구체적인 방법으로 샤논(Shannon)의 정보이론에서 유추한 것으로, 쿠퍼가 제안한 '시그널·잡음 계산'은 다음과 같이 이루어진다. [32]

즉 n개의 문헌집합에대한 K용어의 잡음

$$N^k = \sum_{i=1}^{n} \frac{f_k}{F_k} \cdot \log \frac{F_k}{f_i^k}$$ 로 나타나며

시그널S는 $S^k = \log F^k - N^k$가 된다. 그래서 용어 K가 각 문헌에 한번씩만 나타나면 모든 $f_i^k = 1$이 되어

$N^k = \sum_{i=1}^{n} \frac{1}{n} \cdot \log \frac{n}{1} = \log^n$ 으로 $F^k = n$인 경우 시그널 S^k는 0이 된다.

30) Hyun-Hee, Kim, "An Investigation of Automatic Term Weighting Techniques," 정보관리학회지, Vol.1, No.1, 1984, p.44.

31) 정영미, "색인이론과 실제", 연세논총, 제17집 (1980), p.32.

32) Cooper, W.S., "Is Interindexing Consistency a Hobglollim?" American Documentation, Vol.20, No.3,(July,1969) pp.268-278. 송미연, 上揭論文, pp.6-7에서 재인용.

반대로 집중분포를 갖는 용어는 빈도 F^k를 갖고 한 문헌에만 나타난다면 잡음이 "0"이 되고 시그널은 최대가 된다. 이처럼 시그널과 잡음의 관계를 측정하여 색인어를 선정하기도 하는데 이의 기준은 시스템의 성능, 즉 검색효율 정도에 따라 달라지게 된다. [33]

3) 상관계수(Association value)를 이용하는 방법

이는 단어간의 관련성 정도를 계산 하는 방법[34][35]으로 같은 문헌내에 두개의 단어가 밀접하게 나타나는 정도에 따라 색인어를 선정해준다. 알고리듬에 의하여 한 문헌에서 두 용어가 동시 발생하는 빈도를 계산하여 상관계수를 구하는데, 용어 A와 B가 문헌에서 여러번 함께 나오면 두 용어는 높은 값을 갖게된다.

용어 j와 k간의 상관계수 f(j,k)는

$$f(j,K) = \frac{\sum_{i=1}^{n} (fi^j . fi^k)}{\sum_{i=1}^{n} (fi^j)^2 + \sum_{i=1}^{n} (fi^k)^2 - \sum_{i=1}^{n} fi^j, fi^k}$$

(fi^j : j용어의 발생빈도
fi^k : k용어의 발생빈도) 의

공식에 의하여 구하고 그 결과 나온 "0"에서부터 "1"까지의 상관계수는 두용어쌍간의 용어상관행열을 만들어 색인어를 선정하는 방법이다.

33) Ibid. p.7.
34) Doyle, L.B. "Indexing and Abstracting by Association," Amesican Documentation, Vol.13, No.4,(October,1962) pp.378-390, 송미연 上揭論文, p.8 재인용
35) Stiles, H.E. "The Association Factor in Information Retrieval, Journal of ACM, Vol 8. No.2, (April,1961) pp.271-279. 上揭論文, 재인용.

2. 자동색인을 위한 구문적 및 의의적 분석방법

1) 구문적 분석방법

자동색인의 통계적 분석기법이, 빈번히 사용된 단어들(특히 중요한 내용어)을 확인해내는 것이라면, 구문적 분석은 문장내에서의 단어의 역할, 즉 그 단어의 문법적인 유별(類別 : 예를들면 '동일한 단어, work가 명사로 사용되었느냐 혹은 동사로 사용되었느냐, 와 같은-), 문장내에서의 단어들의 관계(예를들면, 'Dog bites man.'과 'Man bites dog,'을 확인해내는 것과 같은-)를 확인해 내는 것[36]이라고 간단히 말해 볼수 있다.

자동색인의 구문적 분석방법은 보통 모든 자동색인의 기초가 되는 통계학적 기준과 결합하여 언어학적인 기법을 사용함을 뜻하며, 통계적 시스템에 의한 색인용어 선택이 여러가지 문제점, 특히 정보검색의 정도률을 저하시키는 문제를 내포하고 있어, 이것을 지향하는 방안으로서 연구되어왔다고 볼 수 있다.

문헌의 본문(text)에 대한 언어학적 분석의 깊이의 정도에 따라 자동색인시스템의 차원도 달라질 수 있으며 그 분석의 정도를 보통 완전한 문장분석(full parsing), 중간수준의 문장분석(middle level parsing), 그리고 부분적 문장분석(partial parsing)으로 [37] 구분하고 있다.

가장 정교하고 완전한 수준의 문장분석을 실시한 실례로서는 LSP(Linguistic String Project) [38]를 들 수 있다. 이 수준의 시스템들은 본문에대한 완전한 구문분석을 이끌어내는데, 즉 문장속의 개별 단어들을 구(句)로 결합시키고, 구들을 절(節)로 결합시키는 것 등이 포함된다.

36) Borko & Bernier 前揭書, p.117.
37) Joner, Karen Spark, "Autometic Indexing," Journal of Documentation, Vol. 30, No.4, (December, 1974) pp. 400-401.
38) Sagee, N. Natural Language Information Processing: A Computer Grammer of English and Its Application, Reading, MA : Addison Wesley, 1981. Martin Dillon p.100

문장의 완전분석을 위해서는 광범위한 의미론적 지식이 시스템에 통합되어야 하는데, 보통 사전의 기입사항(dictionary entries)의 여러 속성의 형태로 통합된다. 예를들면 human (boy,girl,children……)을 언급하는 명사들을 LSP에서는 HUMAN이라는 속성에 관련시켰는데 그들이 행위자(agent)나, 피동자로서의 인간을 요구하는 문맥속에 출현할 수 있다는 것을 지적한다.

또한 'believe' 'study' 혹은 'design'과 같은 동사들은 올바른 문장분석을 위해서는 인간 주어를 가져야하고(즉 The violin believed.는 틀림), 반면에 동사 'edit', 'learn', 혹은 'summarize,'는 인간, 명사를 목적으로 가져서는 안된다는(즉, He edited the boy. 는 틀림)것 등을 지적해준다. 이러한 완전한 문장분석에 근거한 자동색인은 완전한 문장분석이 어떤 방법에 의해서든지 미리 사전으로 작성되어 있어야 한다.

LSP 다음으로 정교한 수준의 문장분석, 즉 중간수준의 실제적 시스템으로서, PHRASE를[39]를 들 수 있다. PHRASE는 LSP와 마찬가지로 본문의 완전한 구문분석은 하고 있지만 의미론적 요소가 통합되어 있지않으며, 명칭이 나타내고 있는데로 PHRASE는 내용을 표현하는데 가장 적합하다고 생각되는 본문의 요소들을 선정하기 위하여 본문을 구성구절(Component phrase로 압축 시키도록 고안되어있다. 이 시스템은 용어들의 허용된 구문적 관계(The acceptable syntactic contexts of terms)를 표로 만든 사전을 사용하는데 이 사전은 허용되지 않는 구문적 해석을 제거하기위한 여과기 역활을 한다고 볼 수 있다. 예를들면, 'give' 뒤에는 두개의 명사(구)가 따라와야한다는 것을 명시함으로써 「The man gave the girl biscuits.」라는 문장과 「The man ate the dog biscuits」라는 문장은 구별될 수 있는 것이다

실제적인 검색시스템에서 LSP나 PHRASE와 같은 시스템을 사용하는데 있어서의 중요한 장애점은 이들 시스템들의 일반화가 어렵다는 점이다.

39) Dillon, Martin, 上揭論文, pp.100-101.

LSP는 본문내의 모든 어휘가 사전속에 포함되어야하며 각 엔트리는 적절하게 구문적, 그리고 의미론적 제약점들을 표시해주는 코드가 부여되어야 하는 것이다. LSP보다 덜 엄격한 요구사항을 가지고있는 PHRASE 역시 그 방대한 단어통활사전(Word governor dictionary)를 작성해야되기 때문에 역시 일반화가 어렵다고 하겠다.

세번째 수준의, 즉 부분적 문장분석을 실시한 시스템의 실례로는 FASIT (A Fully Automatic Syntactically Based Indexing System)를[40] 들 수 있다.

FASIT는 내용어들이나 구들이 일정한 구문적 범주, 혹은 그 범주들이 결합한 범주에 속하고있다는 생각에 그 근거를 두고있다. 즉 본문내의 단어들을 각각 그해당 범주들의 유형에 근거하여 개념들을 선정한다. 그리고 이 개념들의 변형들을 전거형태(authoritative)로 조정하여, 표준형으로 만들어 하나의 어간에 중첩된 모든 표준형태들을 같은 것으로 모아지게 된다. 즉 'catalog libra'는 'catalog'으로 표현되는 그룹에서 그 멤버·쉽을 갖게된다. FASIT가 사용한 구문적 범주들은 「Standard American English in Green」[41]을 분석하여 채택한 것이고 그것은 영어의 전통적인 팔품사(八品詞)에 근거하고 있는 것이다.

이상에서 살펴본 구문적 분석방법에 의한 자동색인시스템들의 실례는 모두 수행될 내용분석의 유형이나 분석자체가 자동적으로 이행되어야하기 때문에 미리 요구된 단계의 특정화나 지시를 준비해 두는 것으로 시작하고 있고, 이러한 지시들이 다양한 유형의 사전형태로 되어있음을 알 수 있다.

최초의 정보분석이나 분류가 정보검색의 효율을 결정할만큼 중요하다는것을 생각해 볼 때 사전작성과 사용에 관련된 문제들을 상세히 연구되고 검토되어야 할 것이나, 이는 본 논문의 범위를 넘고있어, 자동색인을 위한 구문적 및 의의적 분석방법과 관련하여, 의의적분석방법의 설명다음에 개략적으로 검토하려고 한다.

40) Dillon, Martin, 上揭論文, pp.101-108.
41) Green B., and Rubin, G.M.,Automated Grammatical Tagging of English, Providence : Brown University Department of Linguistics, 1971, 上揭論文, p.101 .

2) 의의적(意義的) 분석방법

자동색인을 위한 의의적 분석(Semantic analysis)는 단어들을 단위개념으로 연결시킬 수 있도록 예증관계(paradigmatic relation)혹은 유관계(class relation)을 확립하도록 도와주는 것이며, 다음과 같은 방법[43]들이 연구되어오고 있다.

첫째 접미어나 접두어를 절단시킴으로써 키-워드를 표준에 일치시키는(normalize), 소위 용어절단법(term truncation)이 있다. 이것은 같은 어근(語根 : Stem)을 갖는 용어들을 모아줌으로써 매치되는 용어의 수를 늘려주는 것이며 이 방법에 의하여 정보검색의 재현율이 증대될 수 있다. 둘째 본문에서 추출한 단어나 구를 색인어로 정하기위하여 이미 만들어놓은 디소러스를 사용하기 위하여, 디소러스를 작성하는 방법인데 이 때는 본문의 용어에서 색인어로 연결해주는 폭넓은 유사어통제가 필요하다. 셋째, 관련어들을 묶어주는 다양한 용어 분류방법이 있는데 용어간의 유사성을 측정할 수 있는 매칭함수를 이용하여 그래프로 용어군을 만드는 방법이 있고, 이는 또다시 스트링(string)유형, 스타(star)유형, 그리고 클리크(clique)유형으로 구별할 수 있다. 그레프이론 방법 이외에 한 용어의 속성이 여러개로 나타날 때 이 용어가 다른 용어군에도 중복되어 나타나게 할 수 있는 분류방법으로 글럼프방법이 있다.

3) 구문적, 의의적 분석을 위한 사전작성(辭典作成)

언어의 복잡성과 구문적, 의미론적 구조를 지배하는 불규칙성으로 인하여 사전작성에는 많은 어려움이 발생하며 Salton[44]이 지적하고 있는 사전작성시의 문제점들을 열거해보면 다음과 같다.

43) Borko & Bernier , 上揭書, p.117.
44) Salton, Gerald, Automatic Information Organization and Retrieval, New York: Mc Graw-Hill, 1968. p.22-23.

(1) 직접으로는 특정 정보내용에 관계는 없지만 구문적 기능을 하는 단어들을 제거해야 한다. (ex. 'can')

(2) 동의어 혹은 관련의미로 사용되는 단어들이 문제가 된다.

(3) 문맥에 따라 상이(相異)한 의미로 사용되는 많은 단어들은 문제가 된다.

(4) 완전히 상이한 구조가 동등한 일반적 개념을 표현하는데 사용됨으로해서 발생하는 여러 유형의 구문적인 동의어 (syntactic equivalences)들은 문제가 된다.

(5) 간접적 참조(indirect reference)의 사용, 즉 대명사, 군집명사 및 관사들의 문제가 있다.

(6) 단어간에 존재하는 관계에 유의해야 한다.

(7) 시간에 따라 변화하는 여러 단어들의 의미 혹은 새로운 단어들이 문제가 된다.

이러한 제반 문제들은 사전작성시 완전히 해결 내지 제거될 수는 없으나 적절한 사전(dictionary mapping)알고 리듬을 이용하여 여러가지 불규칙한 결과들을 상당히 감소시킬 수 있다.[45]

문헌의 자동색인을 위하여 사용되는 사전들의 유형에는 부정적사전(negative dictionary), 즉 제외사전(exception dictionary), 동의어 사전(synonym dictionary), 즉 디소러스(thesaurus), 단어의 어간 디소러스(stem thesaurus)와 접미어 리스트(suffix dictionary), 구 사전(phrase dictionary), 그리고 개념의 계층사전(The cencept hierachy)등이 있다.

(1) 부정적 혹은 제외사전

이것은 정보확인 목적으로 사용해서는 안될 용어나, 범주등을 보유하고 있는 사전이다. 통계적 분석에 의한 자동색인의 경우, 불용어(stop words) 사전을 미리 만들어두고 이와 대조하여 이 사전에 있는 단어가 아닌 것의 단어들의 빈도수를 헤아려 어느 수준의 빈도수 이상을 색인어로 선정하는

45) Salton, Gerald, 上揭論文, p.24.

경우에 사용되고 있다.

FASIT의 구문적 분석에 근거한 자동색인의 경우에는 단어들의 제거사전(exception dictionary)과 단어 어미들의 접미어사전은 구문적 범주(〈그림 2〉참조 [46])를 나타내는 조기성(助記性) 테그(tag)를 본문에서 발견되는 각 단어, 숫자 그리고 구둣부호에 할당하는데 사용된다 (〈그림 3〉 [47] 참조)

Syntactic Category	Examples
AP (adverb or preposition)	by, around
APP (adverb, preposition, or particle)	in, on, over
GN (general noun)	analysis
JJ (adjective)	administrative
MD (modal auxiliary)	can, may
NN (singular noun)	library
NNS (plural noun)	libraries
PP (preposition)	of, to, from
PPS (singular nominative pronoun)	I, he, she
PQL (prequalifier)	all, half
QL (qualifier)	all, more
SC (subordinating conjunctions)	for, then
VB (uninflected verb)	choose
VBD (past tense verb)	chose
VBN (past participle verb)	chosen
VBZ (third person singular verb)	chooses

그림 2. 구문적 범주의 예

46) Dillon, Martin, 上揭論文, p.101.

47) 上揭論文, p.102.

A Sample of Text from a Query:

I would like all information on library catalogs produced by automated methods ...

Tagging and Disambiguation (Steps 1-2)

Text	Tag	Dictionary	Disambiguated
I	PPS	Exception	
would	MD	Exception	
like	VB-SC-JJ	Exception	VB
all	PQL-QL	Exception	
information	GN	Exception	
on	APP	Exception	
library	NN	Exception	
catalogs	NNS-VBZ	Suffix	
produced	VBD-VBN	Suffix	
by	AP	Exception	
automated	VBD-VBN	Suffix	VBN
methods	NNS	Exception	

Concept Selection (Step 3)

Concept	Form
library catalogs	NN NNS-VBZ
automated methods	VBN NNS

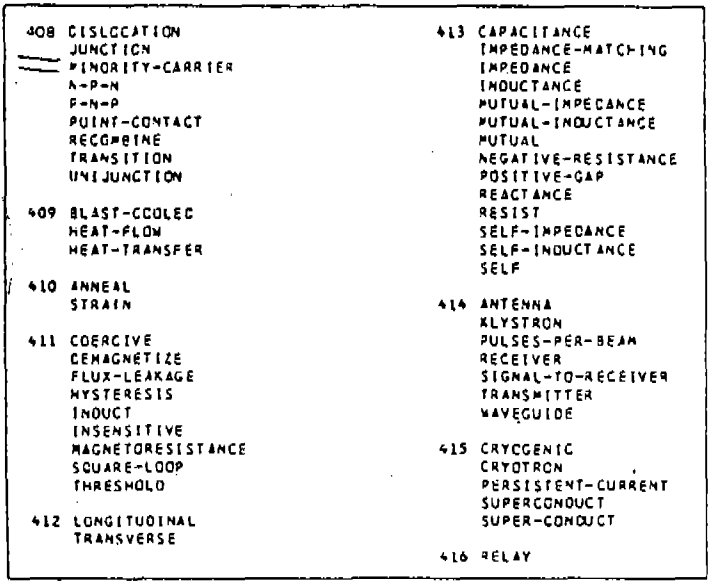
그림 3. 개념선정

(2) 동의어 사전[48]

동의어 사전, 즉 디소러스는 단어, 단어의 어간들을 어떤 주제범주로 묶는 것인데 즉 개념군(concept class)을 형성시킨다.

대표적인 예로는 〈그림 4〉[49]을 들 수 있겠는데 개념군이 세자리 숫자로 나타나 있고 각각의 엔트리는 각 개념의 번호아래 예시되어 있다.

48) Salton , Gerald, 上揭書 , pp.25-29.
49) Salton, Gerald, 上揭書 , p.26.



```
408 DISLOCATION
    JUNCTION
    MINORITY-CARRIER
    N-P-N
    P-N-P
    POINT-CONTACT
    RECOMBINE
    TRANSITION
    UNIJUNCTION

409 BLAST-COOLED
    HEAT-FLOW
    HEAT-TRANSFER

410 ANNEAL
    STRAIN

411 COERCIVE
    DEMAGNETIZE
    FLUX-LEAKAGE
    HYSTERESIS
    INDUCT
    INSENSITIVE
    MAGNETORESISTANCE
    SQUARE-LOOP
    THRESHOLD

412 LONGITUDINAL
    TRANSVERSE

413 CAPACITANCE
    IMPEDANCE-MATCHING
    IMPEDANCE
    INDUCTANCE
    MUTUAL-IMPEDANCE
    MUTUAL-INDUCTANCE
    MUTUAL
    NEGATIVE-RESISTANCE
    POSITIVE-GAP
    REACTANCE
    RESIST
    SELF-IMPEDANCE
    SELF-INDUCTANCE
    SELF

414 ANTENNA
    KLYSTRON
    PULSES-PER-BEAM
    RECEIVER
    SIGNAL-TO-RECEIVER
    TRANSMITTER
    WAVEGUIDE

415 CRYOGENIC
    CRYOTRON
    PERSISTENT-CURRENT
    SUPERCONDUCT
    SUPER-CONDUCT

416 RELAY
```

그림 4. Thesaurus excerpt in concept number order.

〈그림 5〉[50]는 알파벳 순서로 배열된 단어들의 개념번호가 중간란에 나타나 있는데 정보의 식별자가 될 수 없는 보통 단어에는 32,000 이라는 개념번호가 붙어 있고 마지막 란에는 구문적 목적으로 사용되어지는 단어에 구문분석 목적으로 구문코드가 붙어있다.

동의어사전, 즉 디소러스를 작성할 때에 반드시 유의해야 할 사항으로는

① 매우 희귀한 개념들을 문헌과 탐색질문간을 잘 매치시켜줄 수 없으므로 디소러스에서 제외되어야 한다.

② 아주 보편적인 고빈도의 용어들도 위와 마찬가지이므로 제외시킨다.

③ 중요하지 않은 단어들은 제외사전에 포함되기 전에 신중히 검토되어야 한다.

50) 上揭書 , 同面.

	CONCEPT NUMBERS	SYNTAX CODES
BLOCK	663	070043040
BLUEPRINT	58	070043
BOMARC	324	070
BOMBARD	424 0343	043
BOMBER	346	070
BOND	105	070043
BOOKKEEPING	34	070
BOOLEAN	20	001
BORROW	28	043
BOTH	32178	008080012
BOUND	523 0105	070043134135
BOUNDARY	524	070
BRAIN	404 0235	070
BRANCH	48 0042	070042
BRANCHPOINT	23	070
BREAK	380	043040070
BREAKDOWN	689	070
BREAKPOINT	23	070
BRIDGE	105 0458 0048	070043
BRIEF	32232	001043071
BRITISH	437	001071
BROAD-BAND	312	001071
BROKE	380	134104
BROKEN	380	135105
BUFFER	24	070043

그림 5. Thesaurus excerpt in alphabetical order.

④ 애매한 용어들은 관련된 문헌군에 출현될 때에 가지는 의미로 명시되어야 한다.

⑤ 각 개념계층은 매치시키는 특성이 카테고리내에서 같은 단어가 될 수 있도록 거의 같은 빈도의 용어들을 포함해야 한다.

3) 단어의 어간 디소러스와 접미어 리스트 [51]

어간 디소러스는 대표적인 문헌군에 나타난 단어를 사용하여 단어 어간에

51) 上揭書, pp.30-33.

일련번호를 부여한 간단한 단어 어간 목록으로 구성되어 있다. 어간 디소러스의 일련번호는 정규 디소러스에 포함된 개념번호와 일치하고 있다.

〈그림 6〉[52] 의 어간 디소러스는 단어 어간이 알파벳 순서가 아닌, 문헌군내의 발생빈도 순서로 열거되어 있다.

FREQUENCY	STEM	SUFFIX	SEQUENCE NUMBER	FREQUENCY	STEM	SUFFIX	SEQUENCE NUMBER
11	MOCULE	S	2099	12	DECIS	ION	2114
11	PLACE	S	2100	12	DEPOSIT	ED	2115
11	RESPONSE		2101	12	DUE		2116
11	RF		2102	12	ECONOM	ICAL	2117
11	SOURCE		2103	12	ESAKI		2118
11	THICK		2104	12	EXAMIN	ED	2119
11	TRUNC	ATION	2105	12	FUNCTION	AL	2120
11	WAVE		2106	12	GRAPH		2121
11	WHERES	T	2107	12	HAV	ING	2122
11	WIR	ING	2108	12	IMPROVE	MENT	2123
12	ALPHABET	ICAL	2109	12	IMPROV	ED	2124
12	BASE		2110	12	INDIVIDU	AL	2125
12	CAP	ABLE	2111	12	LEAST		2126
12	CENT		2112	12	MAGNETIZ	ATION	2127
12	CONCEPT		2113	12	MAIN		2128

그림 6. Word stem frequency list (null thesaurus).

만일 완전어(full word)가 아닌 단어 어간을 열거하고 싶다면 먼저 접미어 절단 시스템을 만들어야 하고 이것을 위하여 접미어사전에 만들어 지는데 〈그림 7〉[53] 이 대표적인 예다.

〈그림 7〉에서 각 접미어는 일련번호와 하나 혹은 그 이상의 구문적 코우드를 가지고 있는데 오른편의 일련번호는 나중에 어간과 접미어를 완전한 단어로 재조합하는 것이 필요불가결할 때 이용되어진다.

52) 上揭書, p.30.
53) 上揭書, p.32.

Alphabetic suffix list		*Syntactic suffix codes*			
FICATION	058	058	NOUS		
FICATIONS	059	059	NOUP		
FIED	060	060	V00C0	P00 0	ADJ
FIER	061	061	NOUS		
FIERS	062	062	NOUP		
FIES	063	063	V00S0		
FOLD	064	064	ADJ	NOVC	
FUL	065	065	ADJ	NOVC	
FULLY	066	066	AV1		
FY	067	067	V00P0	I00 0	
FYING	068	068	R00 0	G00S0	NOUS ADJ

그림 7. Typical suffix dictionary entries.

4) 구 사전(句辭典)[54]

문헌의 주제분석시에 주제를 개개 단어의 조합으로 이루어지는 구 개념으로 표현하려고 할 때 구 사전을 작성함으로써 주제식별에 사용할 수 있다.

스마트(Smart)시스템[55]에서는 구절 탐지에 두가지 전략의 사전을 이용하는데 통계적 구 사전과 구문적 구 사전이다.

통계적 구 사전(statistical phrase dictionary)은 구절의 구성요소의 통계적 동시발생 특성을 헤아리는 구절탐지 알고리듬에 근거하는 것으로 〈그림 8〉는[56] 그 한 부분을 보여주고 있다.

구문적 구 사전(syntactic phrase dictionary)은 탐지되어야 할 특정 구절의 구성요소의 명세서일 뿐만 아니라 구문 의존관계에 관한 정보도 포함하는 것으로 〈그림 9〉[57]와 같다. 여기에서 구의 첫번째 구성요소는

54) 上揭書, pp.33-38.
55) 上揭書, p.34.
56) 上揭書, p.35.
57) 上揭書, p.36.

PHRASE CONCEPT	COMPONENT CONCEPTS					
543	544	609	-0	-0	-0	-0
282	280	281	-0	-0	-0	-0
282	306	281	-0	-0	-0	-0
280	69	648	-0	-0	-0	-0
280	69	215	-0	-0	-0	-0
694	1285	1284	-0	-0	-0	-0
291	265	290	-0	-0	-0	-0
291	265	496	-0	-0	-0	-0
422	646	185	-0	-0	-0	-0
640	309	290	-0	-0	-0	-0
294	21	293	-0	-0	-0	-0
393	21	635	-0	-0	-0	-0
393	635	106	-0	-0	-0	-0
294	21	245	-0	-0	-0	-0
695	44	150	-0	-0	-0	-0
78	572	565	-0	-0	-0	-0
411	370	328	-0	-0	-0	-0
411	370	389	-0	-0	-0	-0
411	370	476	-0	-0	-0	-0
666	46	601	-0	-0	-0	-0
666	330	53	601	-0	-0	-0
666	347	46	-0	-0	-0	-0
666	347	290	-0	-0	-0	-0
666	347	601	-0	-0	-0	-0
666	357	290	-0	-0	-0	-0

그림 8. Excerpt from statistical phrase dictionary.

185 혹은 624개념으로 이루어져야하며, 두번째는 225의 개념을 표현해야 한다. 달러표시($)의 지시자는 구문적 정보를 전달해준다. 즉 유형 7, 15,16을 말하는데 구체적으로 명사구, 주어-동사관계, 동사-목적관계, 주어-목적 관계 등 4가지의 구문관계 아래 약 20개의 유형이 있다.

5) 개념의 계층사전[58]

주제분석시스템에서는 단어나 단어간의 계층적 배열을 정보식별과 검색목적으로 사용될 수 있으며 대표적인 개념의 계층사전의 예로서 〈그림 10〉[59]

58) 上揭書, pp.38-40.
59) 上揭書, p.39.

```
NAME  OUTPUT   FIRST    SECOND  TYPE 7  TYPE 15  TYPE 16
 OF    CON-    NODE     NODE    SERIAL  SERIAL   SERIAL
TREE   CEPT     CONCEPTS         143     398      399
MAGSWI=422(185,624)/(225)$7/143,15/398,16+
MANMCH=517(600)/(516)$7/144,15/400
MANROL=286(290)/(113)$7/145,5+,15/401,16+,19+
MATHOP=594(615)/(7,116,376)$7/147
MCHBKD=69(689)/(600)$1/148
MCHCOD=304(102,281)/(14,41,600,601)$1/149,15/404
MCHOPE=93(615)/(600)$7/150
MCHORI=41(513)/(600,601)$7/151,15/405
MCHTIM=691(617)/(52,600,601,605,1281)$7/152
MCHTIM=691(617)/(72,615)$1/153
MCHTRA=303(98)/(119,600)$1/154,4+,5+,6+,10+,15/406,16+,19+
MEMACC=593(672)/(121)$1/159,15/409
MEMCOR=557(669)/(121)$7/137,15/395
MEMEFF=284(64)/(121)$1/160,6+,15/410
MEMSPA=552(212)/(121)$1/162,13+,15/411
MHTOOL=471(337)/(600)$7/164
MINSTA=294(245)/(230)$7/165,15/412
MISTRA=668(46,341)/(346)$1/166,3+,15/413
MISTRA=668(341)/(344)$1/168,15/414
MISTRA=668(496)/(341,657)$7/169
MISTRA=668(657)/(346)$7/170
MLTACC=481(672)/(55)$7/171,15/415
MTHMOD=722(1273)/(116)$7/172
MTHSIM=722(353)/(116)$7/173
NATLNG=283(102)/(35,179)$7/174,15/416,16+
NETFNC=639(618)/(623)$7/175
NLNSYS=727(496,1273)/(297)$7/176
NOGCMP=91(601)/(604)$7/177,15/418
NOGDIG=288(604)/(603)$1/178,6+,15/419
NOGSIM=679(353)/(604)$7/180
```

그림 9. Excerpt from criterion tree dictionary.

를 들 수 있다. 이는 광의의 개념, 보다 일반적 개념들이 왼쪽에 나타나 있으며 계층구조의 근간이된다. 또한 상호참조를 나타내 주기도 하는데 원래의 개념과 그렇지않은 것의 일반적이고 비한정적인 유형을 표현해주고 있다.

계층적 주제배열을 형성할때는 어떤 일반적인 응용 가능한 알고리듬을 가지는 것이 도움이 되는데 대체로 광의의 개념은 계층꼭대기 근처로, 특정 개념은 바닥 근처에 와야한다는 것이 통념이 되고있다.

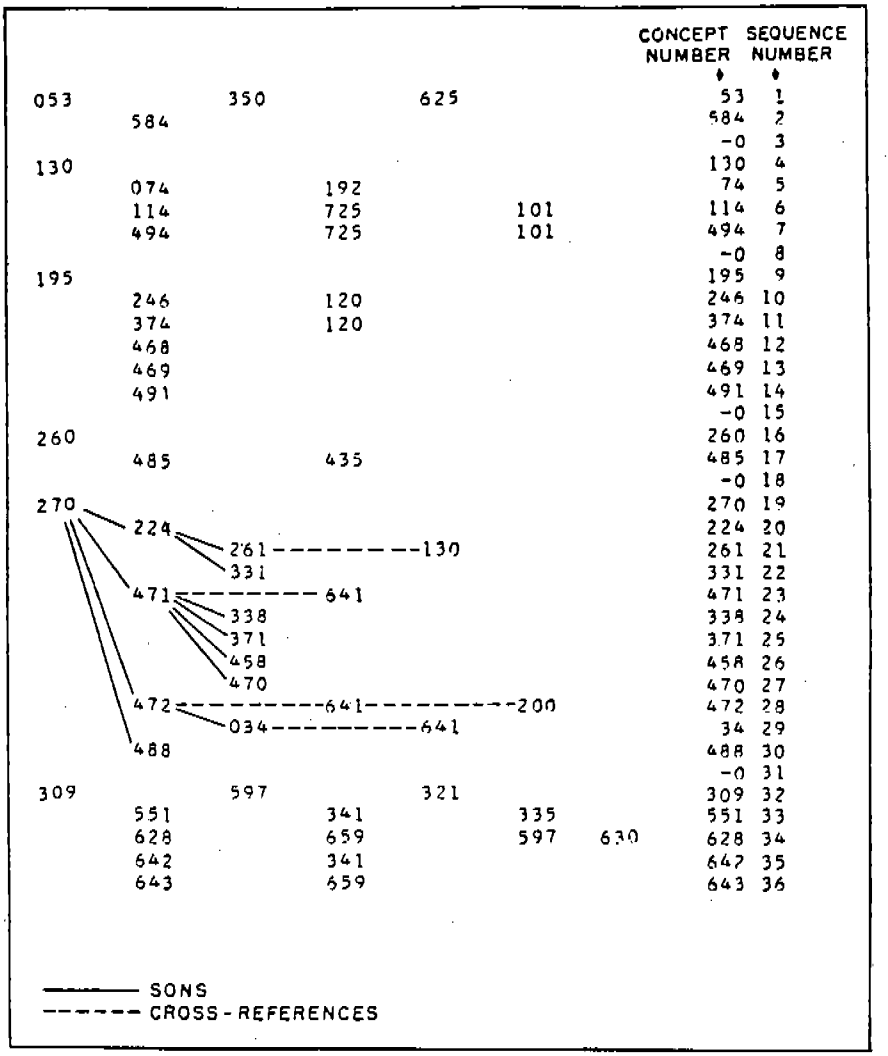


그림 10 Hierarchy excerpt.

이상에서 구문분석과 의어적 분석을 위한 사전을 그 유형 중심으로 개략적으로 살펴 보았는데 앞으로 구체적인 사전작성 과정, 사전의 성능,사전의 이용 등에 관한 연구가 이루어져야 할 것이다.

3. 주제분석기법으로서의 자동색인의 문제점

주제분석기법으로서의 자동색인은 30여년간의 연구, 실험을 거듭하여 왔으면서도 아직 일반화, 보편화되고 있지 않은 상태라는 점을 고려하여 보더라도, 상당히 복잡하고 어려운 문제들을 내포하고 있다고 할 수 있으며, 그 중에서 가장 중요한 것으로 생각되는 것만 몇가지 들어보면 다음과 같다.

첫째, 주제분석의 관련분야로 보고있지 않으면서 사실은 주제분석 특히 자동적인 주제분석기법에 깊은 영향을 미치고 있는 많은 영역이 있으며 (그림 11 참조) [60] 이 학문영역에 대한 교육실시 및 이 분야의 숙련을 달성하는 일이 정보학분야에서 어려운 과제라는 점이다.

둘째, 자동색인의 비용 효과에 대한 명료한 해답이 얻어지지 않고 있다고 할 수 있으며, 몇 시스템들이 생산적 환경에서 실제로 자동색인 시스템을 운용하고 있다는 사실이 비용/효과의 타당성을 입증하여 주는 것이 아니라는 점이다. [61]

셋째, 온-라인 상호작용시스템의 출현으로 색인작성과 탐색간의 경계선이 점차 희미해져가고 있음으로해서 용어통제와 색인언어 양립성 문제가 중요한 과제로 대두되고 있는 점이다. [62]

넷째, 지금까지 자동색인을 위한 실험의 거의 모든 알고리듬은 약 60%의 정확성을 나타내고 있어 [63] 구문분석, 어휘빈도분석 및 다양한 사전작성을 위한 보다 정교한 알고리듬이 필요한 반면에 정교하고 복잡한 알고리듬의

60) Liston, JR. David M. et al, 上揭書 , p.102.
61) Travis, Irene, 前揭論文 , p.143.
62) 上揭論文 , p.143.
63) 上揭論文 , p.141.

DISCIPLINES EXPECTED TO CONTRIBUTE TO THE DEVELOPMENT OF THE FUNCTIONS \ FUNCTIONS PERFORMED ON CONCEPTS IN SUBJECT ANALYSIS	• DETECTION • RECOGNITION	• EXPRESSION • REPRESENTATION • ENCODING/DECODING	• ORGANIZATION • STORAGE • RETRIEVAL	• COMMUNICATION • TRANSFER	• MANIPULATION • ANALYSIS • SYNTHESIS
• MATHEMATICAL/THEORETICAL LINGUISTICS • AUTOMATIC TRANSLATION • AUTOMATIC LANGUAGE PROCESSING	X	X	X		X
• STATISTICS/PROBABILITY • INFORMATION THEORY • PATTERN RECOGNITION	X	X	X	X	X
• COMPUTER SCIENCE AND TECHNOLOGY			X		X
• SYMBOLIC LOGIC • ARTIFICIAL INTELLIGENCE		X			X
• MECHANO/BIOLOGICAL RESEARCH	X	X	X	X	X
• PSYCHOLOGICAL RESEARCH	X	X			X

그림 11. DISCIPLINES EXPECTED TO CONTRIBUTE TO THE DEVELOPMENT OF SUBJECT ANALYSIS FUNCTIONS

일반화는 현 단계에서는 아직 어렵다는 점이다.

다섯째, 지난 20년 내지 30여년에 걸쳐 각종 검색기법들을 상호보완적으로 보던 견해들이 점차 서로 상반되는 것으로 간주하는 경향으로 나아가고 있는데, 즉 분류와 색인, 인용색인과 주제색인, 자연언어시스템과 통제

언어시스템 등의 관계가 그러하며, 그기다가 여러 형태의 자동색인기법들이 공존하고 있다. 이들간의 원리 원칙들을 통정하고 체계화한 주제분석의 기준이 개발되어 있지 않다는 점이다.

Ⅳ. 결 언

수작업 주제분석의 치명적인 결점이라 할 수 있는 일관성 결여와 주제분석의 두 가지 목적들(즉 ① 특정 주제영역에서의 검색을 위한 내용의 확인—주제명표목과 관련, ② 관련자료끼리 검색될 수 있도록 하는 내용의 확인—분류와 관련)사이의 모순을 기계분석방법이 해결할 수 있다는 가능성 때문에 자동색인에 대한 갖가지 접근방법이 개발, 연구, 실험되어오고있다.

본 논문에서는 자동색인을 위한 통계적 분석방법, 구문전, 의의적 분석방법을 검토하였으며, 앞으로 이러한 자동기법의 합당한 적용시기를 고려하는데는 비용/효과에대한 치밀한 계산이 필요하며, 점점 폭발적으로 증가하는 정보량으로 인한 온·라인 탐색과의 긴밀한 관계에서 색인방법이 연구되어야 한다는 것을 알게되었다.

앞으로 점차 기계가독형으로된 문헌내용이 널리 사용되게 될 가능성이 크기 때문에 자동색인시스템의 연구가 보다 용이하게되고 따라서 자동색인 시스템이 실제로 증식될 것이라고 전망한다.

〈참고문헌은 각주로 대신함〉

Automatic Indexing as a Subject Analysis Technique

Lee, Young-Ja *

〈 Abstract 〉

The human subject anaysis of a document has some critical problems. The method results in the inconsistency in analysis process and the contradiction of two objects of the subject analysis (one is the identification of the content for the retrieval of specific items and the other is to identify the content for the grouping of related materials).

Since the subject analysis by mechanized has been recognized to be the possible way to aggreviate the problems of manual analysis, various approaches of automatic indexing have been studied and experimented.

This study is to examine the automatic indexing as one of the promising subject analysis techniques by statistical, syntactical and semantic approaches.

In conclusion, the reasonable application time of the automatic indexing should be made a decision based on the through investigation on the cost verse effectiveness, and automatic indexing system should be developed in the close relationship with the on-line search which is a good retrieval

* Department of Library Science, College of Social Science, Kyungpook National University.

system for information explosion society.

From now on, since the machine-readable document-text will be envisaged to be more and more available due to the rapid development of computer technology, the more substantial research on the automatic indexing will be also possible, which can bring about the increasing of practical automatic indexing systems.