콤퓨터에 의한 정보검색

柳 京 熙*·李 正 一**

<차 례>

〈머 리 말〉

과학기술 계산용 계산기계로서 콤퓨터가 등장한 지는 30년도 채 못되지만 이 동안의 진보는 눈부신 바 있다. 그리고 현재와 같은 정보화 사회에 있어서는 필요 불가결의 것이 되고 있다.

이것은 콤퓨터가 단순히 숫자만을 취급하는 계산기계가 아니라 문자정보와 도형정보도 취급할 수 있는 정보처리기계이기 때문이다. 이와 같은 콤퓨터를 정보검색에 사용하게 되었다는 것은 실로 너무도 당연한 일인 것이다. 특히 정보검색의 본질이 데이터의 축적과 탐색임을 고려할 때, 콤퓨터의 기억장치의 발달은 이 분야의 진보에 크게 기여하고 있다.

본고에서는 이러한 콤퓨터를 정보검색에 어떻게 적용하면 좋을까 하는 것을 중심으로 해설코자 한다.

Ⅰ. 정보검색의 콤퓨터화[1)]

콤퓨터는 종래의 기계와는 달리 기억, 계산, 판단, 탐색, 인쇄 등 대단히 광범한 기능을 가지고 있으며, 콤퓨터화한다는 것은 지금까지의 기계화와는 전혀 다른 의미를 가지고 있다. 즉, 콤퓨터화라는 것은 업무의 일부, 예를 들면, 탐색이라든가, 인쇄라고 하는 것을 하나하나 기계화하는 것이 아니라, 하나의 업무전체의 방법을 변화시키는 것이다. 따라서, 정보검색을 콤퓨터로써 행하고자 할 경우에는 전체 연구개발 시스팀을 고려하고, 그 서브시스팀으로서 정보관리 시스팀을 염두에 두고, 그 한 가지 요소로서 색인이나 탐색을 생각하여야 하는 것이다. 이것을 보다 구체적으로 말하면 색인이나 탐색과 함께 도서의 구입, 대출관리, 연구관리, 특허관리 등을 포함한 하나의 시스팀을 콤퓨터화할려고 하는 어프로우치가 필요한 것이다.

다음에 정보 시스팀의 하나의 서브시스팀인 정보검색 시스팀의 구성을 보면 이 시스팀은 3개의 기능을 가지고 있음을 알 수 있다.

① 색인의 작성—不特定多數의 이용자를 대상으로 데이터로부터 색인을 만든다.

② SDI 서어비스—특정개인을 대상으로 신착자료 중에서 그 사람에게 필요하다고 생각되는 것만을 선택하여 배포한다.

③ Q-A(Question-Answering) 서어비스—특정 개인으로부터의 질문에 대하여 회답한다.

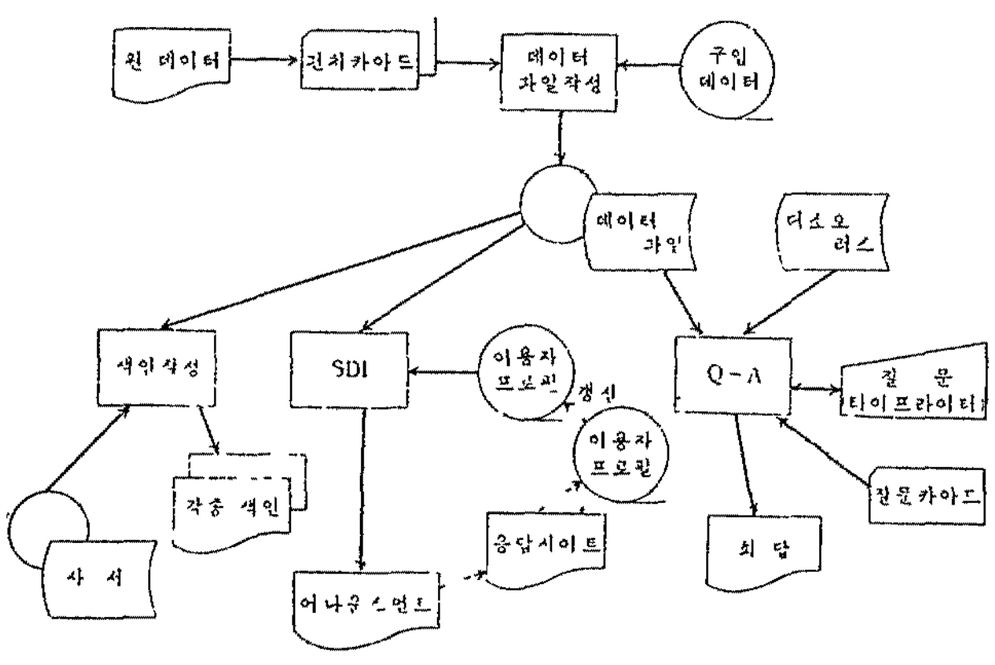
이러한 기능들은 각각 전혀 다른 서어비스이지만 취급하는 데이터는 동일하며 처리방법만이 다르다. 따라서, 이것을 유기적으로 결합함으로써, 정보검색 시스팀

*조사검색부 부장 **조사검색부원

의 효율을 향상시킴과 동시에, 이용자도 다양한 서어비스를 받게 되며, 인간을 포함한 시스팀 전체의 효율도 상승하게 된다.

정보검색 시스팀을 고려할 경우에는 이러한 기능 외에 데이터 파일 작성기능을 추가할 필요가 있다.

도-1은 이러한 관계를 잘 나타내고 있다.



도—1 콤퓨터화한 정보검색 시스팀[1]

Ⅱ. 데이터 파일의 작성

A. 데이터 파일[2]

정보검색 시스팀을 콤퓨터화하는 데는 먼저 처리해야 할 데이터 파일[2]을 만들지 않으면 안된다. 데이터 파일은 매뉴얼방식(Manual Method)인 경우의 매뉴얼카아드(Manual Card)에 해당한다. 따라서 먼저 무엇을 넣고, 무엇을 축적매체(蓄積媒體)로 할까 하는 것을 결정해야 한다. 넣어야 할 데이터로서는 매뉴얼카아드에서 보통 사용하고 있는 항목을 모두 그대로 사용할 수 있다. 또한 이러한 항목들은 코오드화할 필요가 없으며 자연어(自然語) 그대로가 좋다. 다만 쓰는 방법을 통일할 필요는 있다. 예를 들면, 연월일을 표시하는 데 양력과 음력을 혼용하는 것은 좋지 않다.

이러한 데이터는 입력매체(入力媒體)를 통하여 축적매체에 기록된다. 그리고 입력매체로서는 카아드나 종이테이프가 일반적으로 이용되고 있다. 이러한 매체에 어떤 항목을 어떻게 펀치할까 하는 것은 입력형식의 설계문제로서 매우 중요하다.

축적매체는 자기(磁氣)테이프 또는 자기디스크가 일반적이지만, 데이터가 소량일 경우에는 입력에 사용하는 카아드를 그대로 쓰는 것도 좋다. 축적매체상에 어떻게 데이터를 기록할까 하는 것은 입력형식과는 별도로 결정해야 하며, 이 형식의 설계 여하는 뒷처리나 데이터 파일의 수정에 영향을 미치기 때문에 신중히 결정하지 않으면 안된다.

이하, 이러한 설계에 대하여, 입력매체를 80란(欄) 카아드, 축적매체를 자기테이프로 하여 설명코자 한다.

B. 입력 데이터 카아드의 설계[3]

데이터로서 인풋(Input)되는 항목은 다음과 같다. 문헌번호, 표제, 부제(副題), 저자명, 저자 소속 기관명, 잡지명, 권, 호, 페이지, 발행년월, 발행소, 국명 분류, 키이워드(Keyword), 초록, 문헌형식, 인용문헌, 기타.

이러한 것을 80란(欄) 카아드를 사용하여 인풋하는 데는 문헌 1건에 수매 혹은 수십매의 카아드를 필요로 한다.

그러므로, 각 카아드는 하나의 문헌에서 만들어진 일조(一組)의 카아드에 공통기호인 인풋 ID, 각 카아드가 상기의 항목의 어디에 상당하는가를 표시하는 카아드 ID, 그리고 동일한 카아드 ID가 2매 이상일 때 순서를 정해 주는 카아드 NO의 3개의 필드(Field)가 필요하다(도-2).

1 2	3 4	5 74	75 80
카아드 ID	카아드 NO	내 용	인풋 ID

도-2. 카아드 형식

카아드 ID를 상기의 항목에 대하여 할당한 것이 표-1이다.

순서	ID	내 용
1	AD	서지적 정보
2	AU	저자명, 소속 기관명
3	TL	표 제
4	ST	부 제
5	HE	키이워드
6	AB	초 록
7	OR	원 문
8	CT	인용 문헌명
9	PU	발행소
10	OP	기타 필요한 사항
11	ME	메 모

표-1. 카아드 ID 대조표

내용란은 카아드 ID에 따라서 다르다. AD 카아드는 서지적(書誌的) 데이터를 인풋하는 데 사용하기 때문에 항목의 개수, 길이를 정해둔다(표-2).

이 카아드는 반드시 있어야 하며, 또한 이 카아드 중의 문헌번호는 기입하지 않으면 안된다. 그러나 다른 항목은 필요에 따라서 사용하면 좋다.

카아드 칼 럼	내 용	사용 문자
1~2	카아드 ID	A, N
3~4	카아드 NO	N
5~20	문헌 번호	A, N, K
21~22	분 류 (1)	A, N
23~24	분 류 (2)	A, N
25~44	잡지명	A, N, K
45~47	권	A, N
48~51	호	A, N
52~55	연 월	A, N
56~65	페이지	A, N
66~69	사용하지 않음	
70~72	국명, 국어	A, N, K
73~74	문헌 형식	A, N
75~80	인풋 ID	A, N

A : 알파벳, 특수 기호
N : 숫자
K : 가, 나, 다 문자

표-2. AD 카아드의 내용

AU 카아드는 저자명과 소속 기관명을 넣는 카아드이지만, 일반적으로 저자명이나 소속 기관명의 길이는 일정하지 않고, 또 개수도 공저자를 고려하면, 미리 정해 둘 수가 없다.

이러한 경우에 취할 수 있는 방법은 두 가지가 있다.

하나는 생각되는 최대 자수를 예견하여 카아드 형식을 설계하고, 매수도 최대인수분(最大人數分)을 취할 수 있도록 하는 것과 또 하나는 실제로 필요한 자수(字數)만을 쓰고, 2명 이상의 경우에는 차례차례 연이어 쓰도록 하는 방법이다.

도-3은 전자의 예로서 칼럼 5에서 34까지의 30자를 저자명으로, 칼럼 35~74까지의 40자를 소속 기관명으로 할당하고, 각각 좌단에서부터 쓰도록 한 것이다. 공저자가 있을 경우는 동일형식의 카아드를 2매째 이후에 계속한다.

1 2 3 4 5 34 35 74 75 80

AU	02	CHAN YOUNG PARK	KIST	700401
AU	01	SEUNG JUNE YUN	KORSTIC	700401

도-3. 저자카아드 예 1

1 2 3 4 5 74 75 80

AU	01	CHAN YOUNG PARK(KIST), SEUNG JUNE YUN (KORSTIC)	700401

도-4. 저자카아드 예 2

도-4는 후자의 예로서 칼럼 5에서부터 저자명을 쓰고 소속 기관명은 직후에 ( ) 안에 넣고, 공저자가 있을 경우는 , 로 구별하여 동일하게 쓴다. 전자는 카아드 매수가 증가하나, 콤퓨터에서의 처리가 쉽고, 또 프로그램도 간단하다. 후자는 쓰기 쉽고, 매수도 적지만 콤퓨터에서의 처리가 귀찮다.

TL 카아드는 표제를 위한 카아드로서 칼럼 5에서부터 표제를 그대로 기입한다. 표제가 70자 이상일 때는 2매째 이후에 계속한다. 이 경우 앞 카아드에서 다음 카아드로 넘어갈 때 단어가 도중에서 끊어지지 않도록 하는 것이 좋다.

부제를 쓰는 ST 카아드, 초록을 위한 AB 카아드 등도 TL 카아드와 마찬가지로 취급한다.

키이워드를 쓰는 HE 카아드는 표제와 부제만으로써는 그 문헌의 내용을 충분히 표시할 수 없을 경우에 사용하고, 키이워드인 단어 또는 구(句)를 , 로 구별하여 기입한다.

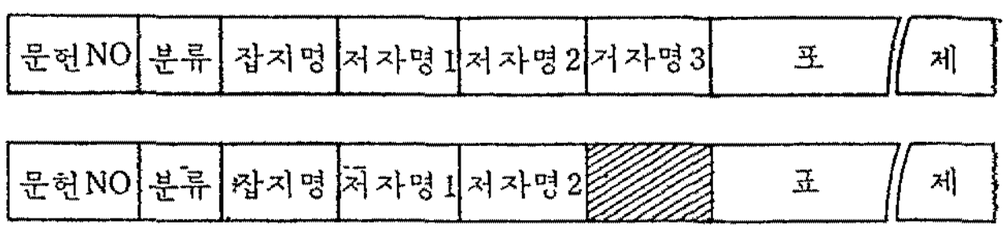
이상과 같이 설계하여 두면, 하나의 항목이 몇 매가 되도 상관없고, 불필요하다면 사용하지 않아도 좋다. 또한 새로운 항목을 추가하기도 쉽기 때문에 매우 융통성이 있는 시스팀을 만들 수 있다.

C. 데이터 테이프의 형식

인풋된 데이터를 축적하기 위한 자기테이프상의 데이터의 형식에는 다음과 같이 3 가지 형식이 있다.

① 고정장(固定長)

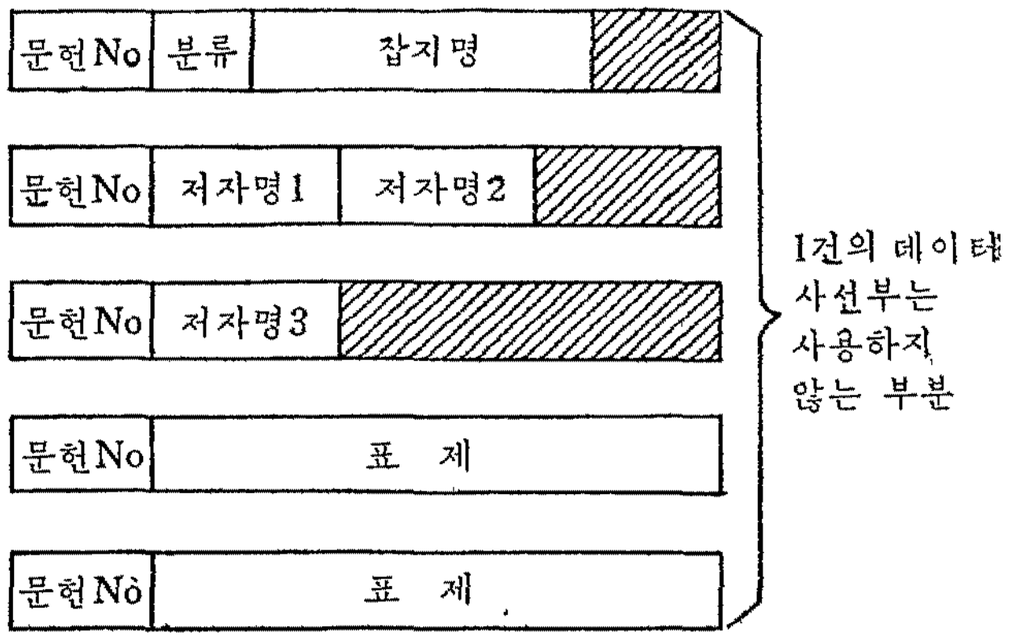
1건의 데이터의 전체의 길이 및 각 항목의 길이와 개수가 일정한 형식(도-5).



도—5. 고정장 형식

② 반고정장(半固定長)

1건의 데이터의 길이, 각 항목의 길이는 일정하지 않



도—6. **반고정장 형식**

지만 각 항목마다 일정한 길이를 주고, 나머지는 공백(Blank)으로 하며, 부족할 경우에는 동일 길이의 것을 필요한 수만큼 취한다(도-6).

③ 가변장(可變長)

1건의 데이터의 길이, 각 항목의 길이가 일정하지 않으며, 자기 테이프상에 기록할 때도 그대로 쓴다. 각 항목의 개시 위치는 그 장소를 알 수 있게 되어 있다(도-7).

문헌 NO	분류	잡지명	저자명 1	저자명 2	저자명 3	표 제

문헌 NO	분류	잡지명	저자명 1	저자명 2	표 제

도-7. 가변장 형식

정보 검색용 파일로서 사용되고 있는 것은 일반적으로 고정장과 반고정장이다. 고정장에 있어서는 사용하지 않는 항목도 지정한 자수분(字數分)만큼 확보하여 두지 않으면 안되지만, 콤퓨터는 처리하기 쉽다. NTT 通研의 REWDAC[5]은 이 형식이다.

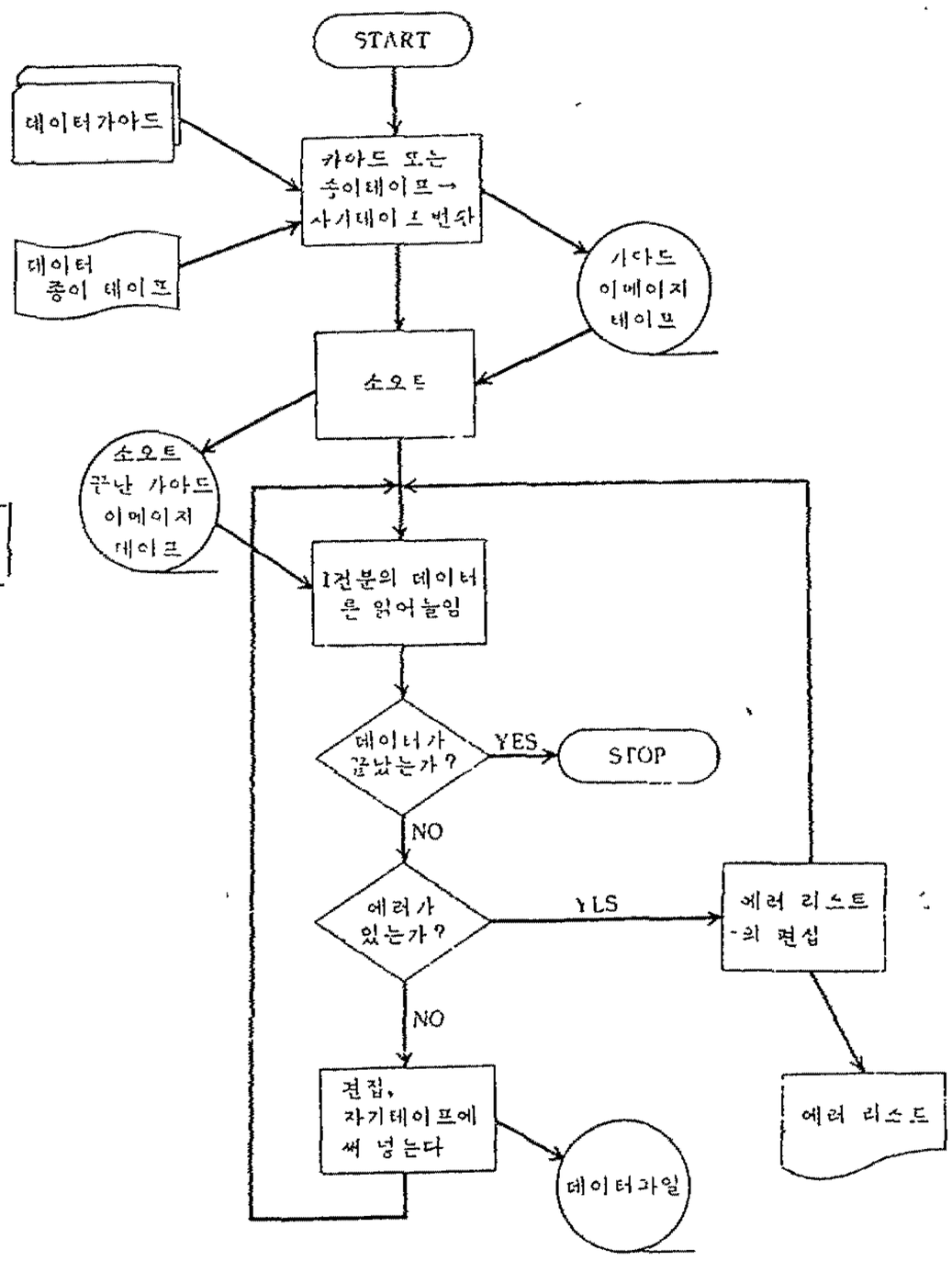
반고정장은 필요에 따라서 어떤 항목에도 장단을 자유로이 할 수 있기 때문에 문헌검색처럼 사용자수에 연결성이 없는 데이터에는 최적하다.

JICST의 BCD 파일, DERWENT 社의 Ringdoc, 日本電氣(株)의 DIA 시스템 등이 모두 이 형식이다.

가변장의 경우는 말할 것도 없이 자기 테이프의 사용효율은 매우 좋지만, 콤퓨터의 처리가 귀찮고, 효율도 나쁘기 때문에 그다지 사용되고 있지 않다.

D. 입력 프로그램

입력 프로그램은 카아드 또는 종이테이프에 펀치된 1건분의 데이터를 읽어 들여서, 에러체크(Error Check)와 편집을 행한 후 자기테이프에 수록한다. 카아드의 경우는 순서가 바뀔 수도 있기 때문에 보통 카아드를 그대로 자기테이프에 수록하고, 일단 소오트(Sort)하면서 처리한다(도-8).



도—8 입력 프로그램 제너럴 플로우 차아트

Ⅲ. 색인의 작성

A. 색인지

데이터 파일이 만들어지면, 이것을 이용하여 프로그램에 따라 여러 가지 색인을 만들 수 있다. 색인은 색인지로 인쇄하여 연구자 등의 이용자에게 배포하면 가장 손쉽게 사용할 수 있는 정보검색툴(Tool)이 된다. 또한 콤퓨터의 면에서도 색인의 작성은 정기적이며, 양적으로도 보통 일정하기 때문에 기계검색과는 달리 그것자체를 희람하든가, 목차(Contents Sheet) 코피(Copy)를 배포함으로써 많은 이용자에게 신속히, 그러면서도 이용하기 쉬운 서어비스를 할 수 있다. 누가색인지는 과거 1년분 혹은 전부를 정리한 색인지로서, 양이 많을 경우에는 분류마다 分冊으로 작성한다.

B. 색인의 종류

색인은 프로그램적으로는, 데이터 파일 중의 전항목을 색인어로써 작성할 수 있지만, 이 때는 프로그램을 하나씩 만들지 않으면 안된다. 또한 이용도를 고려하면 자연히 종류는 한정되게 된다.

흔히 만들어지는 색인지에는 다음과 같은 것이 있다.

서지(書誌) 리스트

분류 색인

키이워드색인 (표제색인, KWIC 색인, KWOC 색인, WORD 색인)

저자색인(저자색인, 저자 기관색인, 저자 리스트)

기관색인

WADEX

C. 색인작성 프로그램[6]

색인을 만드는 데는 모든 색인에 대하여 각각 다음과 같은 3 가지 프로그램이 필요하다.

① 키이워드와 이것과 함께 인쇄항목을 뽑아내는 프로그램(색인 테이프 작성 프로그램)

② 색인어를 알파벳순 혹은 가나다순으로 정리하는 프로그램(소오트 프로그램)

③ 편집·인쇄하는 프로그램[7](인쇄 프로그램)

이러한 프로그램중 소오트 프로그램은 유틸리티(Utility) 프로그램으로서 바로 사용되는 것이 있지만 색인 테이프 작성 프로그램과 인쇄 프로그램은 하나하나 만들지 않으면 안된다.

이러한 것은 컴파일러(Compiler) 언어 특히 COBOL 또는 PL/I으로써 쉽게 만들 수 있고, 또 콤퓨터가 자동 색인작성 프로그램의 패키지(Package)를 공급하고 있는 곳도 있다.

색인작성 프로그램을 만드는 데는 먼저 무엇에 대한 색인을 만들까를 결정해야 한다. 즉 색인어를 어떤 항목에서 취하며, 함께 인쇄할 항목은 무엇일까를 결정하는 것이다. 특히 몇 가지 색인을 만들 경우에는 이러한 것을 잘 조합시킬 필요가 있다.

조합에 대한 하나의 예로서는 서지 리스트와 KWIC 색인(또는 KWOC 색인)과 저자색인의 3 가지를 사용하는 방법이 있다.

이것만 있으면 표제 중의 키이워드로부터 또는 저자명으로부터도 검색할 수 있다.

다음에 인쇄형식을 정해야 한다. 보통은 실제로 타이프라이터 등으로 인쇄형식을 만들어 이용자의 반응을 보는 방법이 사용되고 있다.

D. 표제 KWIC 색인작성 프로그램

도-9는 KWIC 색인을 작성하는 프로그램의 제너럴 플로우 차아트(General Flow Chart)이다.

먼저, 데이터 파일에서 하나의 문헌에 대한 데이터를 읽어 들인다(②). 여기서 읽어 들여야 할 데이터가 없으면, ⑩으로 가지만, 아직 있을 경우에는 ④로 가서 표제 중의 최초의 낱말을 뽑아낸다. 여기서 낱말이란 딜리미트(Delimit)로서 둘러 싸인 문자의 집단을 말하므로, 2개 이상의 낱말로 되는 경우도 있지만 여기서는 편의상 낱말이라고 한다.

딜리미트란 낱말을 구분하는 기호로서 표-3과 같은 것이다.

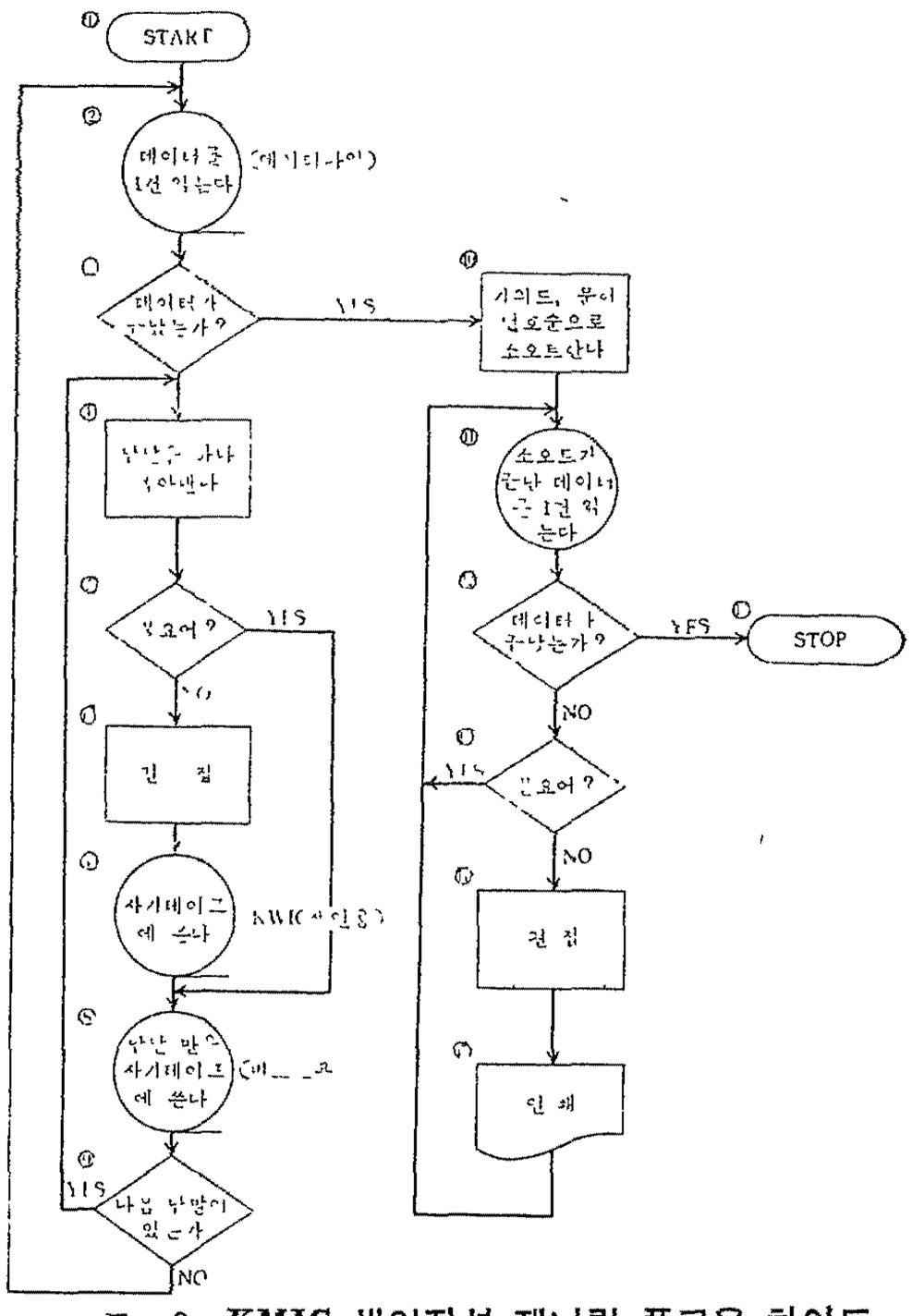
낱말을 뽑아내면 다음에 그것이 불요어[8](不要語)인가를 검사한다.

불요어는 색인어로 하지 않는 낱말로서 그것을 색인상에 써넣어도 이용자는 그 낱말로써 색인을 찾을 수 없다고 생각되는 낱말이다. 일반적으로 색인을 만들 경우 불요어를 지정하고 이 이외의 것은 모두 색인어로 하는 방법과, 반대로 색인어만을 지정하고 이 이외의 것에 대해서는 색인을 만들지 않는 방법이 있다. KWIC 색인 등은 보통 불요어를 지정하는 방법을 사용하고 있다.

불요어에는 4 종류가 있다.

(1) 문법적인 면에서의 불요어(표-4)로서 관사, 전치사, 접속사 등

(2) 1자의 것(A, B, C 등)



도—9 KMIC 색인작성 제너럴 플로우 차아트

	부호	
문을 구별하는 딜리미트	.	종 지 부
	?	의 문 부
	!	감탄부호
수식어를 만드는 딜리미트	,	apostrophe
	=	등 호
	—	연자부호 (minus)
	/	slash
기타 딜리미트	△	blank
	:	colon
	)	우 괄 호
	"	인용부호
	,	구 두 점
	(	좌 괄 호

표-3. 딜리미트 일람표

ABOUT	EACH	ITS	SEVEN	UP
AFTER	EASY	IV	SHALL	UPON
AGO	EIGHT	MAIN	SHE	UPPER
ALL	END	MANY	SHOULD	UPPWARD
ALMOST	ETC	MAY	SIX	US
ALSO	EX	ME	SO	VI
AMONG	FAR	MORE	SOME	VIA
AN	FEW	MOST	SOON	VOL
AND	FIRST	MR	ST	VS
ANY	FIVE	MUCH	TEN	WAS
ARE	FOR	MY	THAN	WE
AS	FOUR	NEW	THAT	WELL
AT	FROM	NINE	THE	WERE
BE	FURTHER	NO	THEIR	WHAT
BEEN	GOOD	NOT	THEM	WHEN
BEFORE	HAS	NOTE	THEN	WHERE
BEING	HAVE	NOW	THERE	WHICH
BEST	HE	OF	THESE	WHILE
BETWEEN	HER	OFF	THEY	WHO
BIG	HERE	OLD	THIRD	WHOLE
BOTH	HIM	ON	THIS	WHOM
BUT	HIS	ONE	THOSE	WHOS
BY	HOW	ONLY	THREE	WHY
CO	IF	OR	THROUGH	WILL
CAN	II	OTHER	THRU	WITH
COULD	III	OUR	TO	WITHIN
DID	IN	OUT	TCO	WITHOUT
DO	INC	OVER	TCWARD	WOULD
DOES	INTO	PART	TCWARDS	YET
DR	IS	SAME	TWO	YOU
DURING	IT	SECOND	UNDER	YOUR

표-4. 英字의 Stopword (문법적)

(3) 숫자로 시작되는 것(1970, 3A 등)

(4) User's Stopword(이용자가 자기의 시스팀에서는 필요로 하지 않는다고 정한 불요어)

이 중에서 (1), (2)는 대개 사용하지만 (4)는 사용하지 않는 것도 있다. 수로서는 (1)은 약 150단어, (4)는 약 1,500단어 정도이다(Chemical Titles에서는 (1), (2), (4)가 1,540단어이다).

불요어의 검사는 (1)~(3)을 ⑤의 단계에서 행하지만 이용자에 따라서는 이러한 것 중에서도 색인어로 하고 싶은 경우도 있다.

불요어가 아니면 색인어와 인쇄항목을 뽑아내어 색인 테이프에 써 넣는다(⑦). 그리고 표제에 아직 처리되지 않고 있는 낱말이 있으면 ④로 되돌아가서 다음 낱말을 취한다. 반대로 표제의 모든 낱말에 대하여 처리가 완료되면, 다음 데이터를 읽어들이기 위하여 ②로 되돌아간다. 이렇게 하여 만들어진 색인 테이프의 예를 도-10에 제시하였다.

모든 데이터의 처리가 끝나면 색인어를 알파벳순 또는 가나다순으로 소오트한다(⑩). 소오팅이 끝나면 인쇄하기 위하여 처음부터 읽어내어 그것이 User's Stopword인가를 검사한 다음(⑬) 편집·인쇄한다(⑮). 인쇄가 모두 끝나면 색인작성은 끝나게 된다(⑯).

```
                    /THE LOGICAL STRUCTURE OF SOME GEOPHYSICAL ARTICLES.=                                  FID-67-3F5
          /THE LOGICAL STRUCTURE OF SOME OF PHYSICAL ARTICLES.=                                            FID-67-3F5
THE USA.=                                   /THE DEVELOPMENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN     FID-67-3F6
                         /THE DEVELOPMENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN THE USA.=              FID-67-3F6
                    /THE DEVELOPMENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN THE USA.=                   FID-67-3F6
               /THE DEVELOPMENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN THE USA.=                        FID-67-3F6
     /THE DEVELOPMENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN THE USA.=                                  FID-67-3F6
MENT AND USE OF OPTICAL COINCIDENCE SYSTEM IN THE USA.=                          /THE DEVELOP            FID-67-3F6
RIC-RETRIEVAL SYSTEM.=                          /A NATURAL-LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAP  FID-67-3G1
IEVAL SYSTEM.=                         /A NATURAL-LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETR  FID-67-3G1
TEM.=                             /A NATURAL-LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETRIEVAL SYS  FID-67-3G1
                    /A NATURAL-LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETRIEVAL SYSTEM.=        FID-67-3G1
          /A NATURAL-LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETRIEVAL SYSTEM.=                  FID-67-3G1
LANGUAGE GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETRIEVAL SYSTEM.=                  /A NATURAL-         FID-67-3G1
GLOSSARY FOR AN AUTOMATED BIBLIOGRAPHIC-RETRIEVAL SYSTEM.=                  /A NATURAL-LANGUAGE          FID-67-3G1
INST EVALUATION OF A STUDY IN THE FIELD OF INFOR/COMPATIBILITY BETWEEN CLASSIFICATIONS AND THESAURI: F  FID-67-3G2
TUDY IN THE FIELD OF INFOR/COMPATIBILITY BETWEEN CLASSIFICATIONS AND THESAURI: FIRST EVALUATION OF A S  FID-67-3G2
 INFORMA/COMPATIBILITY BETWEEN CLASSIFICATIONS AND THESAURI: FIRST EVALUATION OF A STUDY IN THE FIELD OF  FID-67-3G2
ILITY BETWEEN CLASSIFICATIONS AND THESAURI: FIRST EVALUATION OF A STUDY IN THE FIELD OF INFORMATION STO  FID-67-3G2
ASSIFICATIONS AND THESAURI: FIRST EVALUATION OF A STUDY IN THE FIELD OF INFORMATION STORAGE AND RETRIEV  FID-67-3G2
 AND THESAURI: FIRST EVALUATION OF A STUDY IN THE FIELD OF INFORMATION STORAGE AND RETRIEVAL.=FICATIONS  FID-67-3G2
AURI: FIRST EVALUATION OF A STUDY IN THE FIELD OF INFORMATION STORAGE AND RETRIEVAL.=FICATIONS AND THES  FID-67-3G2
EVALUATION OF A STUDY IN THE FIELD OF INFORMATION STORAGE AND RETRIEVAL.=FICATIONS AND THESAURI: FIRST  FID-67-3G2
F A STUDY IN THE FIELD OF INFORMATION STORAGE AND RETRIEVAL.=FICATIONS AND THESAURI: FIRST EVALUATION O  FID-67-3G2
S.=                                   /SOME CONSIDERATION OF THE USE OF KEYWORDS AND THE THESAURU  FID-67-3G3
                         /SOME CONSIDERATION OF THE USE OF KEYWORDS AND THE THESAURUS.=                   FID-67-3G3
               /SOME CONSIDERATION OF THE USE OF KEYWORDS AND THE THESAURUS.=                             FID-67-3G3
UCTION.=                                    /THE SPECIFIC TO-GENERAL SEE REFERENCE IN THESAURUS CONSTR  FID-67-3G4
```

도-10. 표제 KWIC 색인 테이프의 내용

Ⅳ. 기계검색[9]

A. 기계검색의 특징

기계검색은 콤퓨터의 외부 기억장치 속에 축적되어 있는 데이터群에서 질문에 따라 필요한 정보를 뽑아내어, 문자 또는 도형처럼 인간이 이해할 수 있는 형태로 표시하는 검색 시스팀이다.

외부 기억장치로서는 일반적으로 자기테이프 또는 자기디스크가 사용되고 있다. 예를 들면, 자기테이프는 1권에 약 2,000만자나 기록할 수 있지만, 이것은 데이터를 1건당 400자로 하더라도 5만건을 수록할 수 있기 때문에 매뉴얼 카아드라면 두께가 약 13m나 된다.

이 정도의 데이터를 기계검색하면 수초 내지 수분 동안에 필요한 정보를 곧 뽑아낼 수 있다.

또한 이 정도의 데이터에서 KWIC 색인을 만든다면, 1건의 데이터에서 평균 6개의 키이워드를 끄집어 낸다 하더라도 30만행이 되며, 1페이지를 100행으로 잡아도 3,000페이지에 이르게 된다. 여기에 서지 리스트, 저자 색인을 추가하면 6,000페이지를 넘는 방대한 책이 되어 실용성이 극히 희박하게 된다.

기계검색의 제 2의 특징은 분류, 저자 키이워드 등 몇 개의 항목을 조합하여 탐색코자할 경우에 이용자는 별로 수고를 들이지 않고, 검색을 할 수 있다는 점이다. 또한 주어진 질문에 대해서는 모두 회답을 얻을 수 있다. 콤퓨터는 인간이 탐색할 때와 같은 융통성은 없으나 반대로 간과하기 쉬운 에러가 전연 없기 때문에 데이터 파일이 정확하고 질문이 분명하면 반드시 필요한 정보를 얻을 수 있다. 더우기 디소오러스(Thesaurus)가 있어 이것을 시스팀에 이용하면 기계적으로 처리할 수 있다.

디소오러스를 사용하는 방법에는 책자식 디소오러스(Book Thesaurus)를 통하여 필요한 키이워드를 전부 질문으로 이용하는 방법과 기계식 디소오러스(Machine Thesaurus)로서 질문을 넣으면 자동적으로 디소오러스를 참조하게 해주는 방법이 있다.

기계식 디소오러스를 사용할 수 있다는 것은 검색 시스팀을 매우 사용하기 쉽게 해준다.

B. 서어치(Search) 방식과 룩업(Look-Up) 방식

콤퓨터에 의한 검색 시스팀은 서어치방식과 룩업방식의 어느 쪽도 자유로이 사용할 수 있고, 또한 이것을 조합하여 사용하는 것도 가능하다.

일반적으로 외부 기억장치중 자기테이프처럼, 축차적인 처리를 하는 데 적합한 기억매체는 서어치방식을 사용한다. 서어치방식은 데이터를 한건한건 조사하기 때문에 시간이 걸리지만, 모든 항목에 대하여 세밀하게 비교할 수 있다. 또한 프로그래밍도 쉬워, 널리 사용되고 있다. 한편 자기디스크처럼 디렉트 액세스(Direct Access)가 가능한 매체는 룩업방식이 이용되고 있다. 이 경우, 파일은 인풋된 원래의 데이터 파일 외에, "Inverted File"이라고 불리워지는 파일이 필요하다.

Inverted File은 데이터 중에서 키이워드를 하나하나 뽑아내어, 알파벳순, 가나다순 혹은 콤퓨터의 내부 코오드순으로 정리한 파일로서, 각 키이워드에는 그것이 포함되어 있는 데이터의 문헌번호가 붙어 있다(도-11).

#1	K1, K2, K4
#2	K2, K3, K5
#3	K3, K4, K6, K7

(원래의) 데이터 파일

#1~#3 : 문헌번호
K1~K7 : 문헌에 포함되어 있는 키이워드

K 1	#1
K 2	#1, #2
K 3	#2, #3
K 4	#1, #3
K 5	#2
K 6	#3
K 7	#3

INVERTED FILE

도—11. INVERTED FILE

탐색할 때는 먼저 이 Inverted File에서 질문에 포함되어 있는 키이워드를 찾고, 다음에 이 키이워드에 붙어 있는 문헌번호로부터 필요한 데이터를 뽑아낸다.

Inverted File은 콤퓨터가 프로그램에 의하여 자동적으로 작성하지만, 데이터 중에 검색 대상항목이 많으면 많을수록 파일의 크기는 커지며, 모든 항목을 대상으로 하면, 원래의 파일보다 커지게 된다. 따라서 기억매체는 원 데이터의 2배 이상이 필요하다. 또한 프로그래밍도 비교적 어렵다. 그러나 검색시간은 매우 짧아 서어치방식의 수십분의 1 정도이다. 이 때문에 특히 온·라인(On-Line) 시스팀에서는 이것이 많이 이용되고 있다.

C. 배치(Batch) 검색과 온·라인(On-Line) 검색

기계검색의 처리방법으로서 보통 질문을 모아 두었다가 주 1회 정도의 간격을 두고 행하는 방법을 배치검색, 콤퓨터의 단말장치, 예를 들면, 타이프라이터나 CRT 디스플레이(Display)와 같은 것을 사용하여 질문을 넣고, 회답을 직접 얻는 방법을 온·라인검색이라 하여 구별하고 있다.

이것은 질문에서 회답까지의 시간, 사용기기, 서어비스 방법 등이 전연 다르다.

배치검색 시스팀은 소규모의 콤퓨터로서 행할 수 있고 비용 등도 적게 들어 문헌검색에는 가장 일반적인 것이다. 콤퓨터의 기기로서는 최저,

主記憶 16K字(K=1048)
자기테이프 4대
카아드 리이더(Card Reader) 또는 페이퍼 테이프 리이더(Paper Tape Reader) 1대
라인 프린터(Line Printer) 1대

로 좋다. 또한 1회의 검색에서 수십 가지 질문을 처리할 수 있다. 다만, 콤퓨터를 사용할 수 있는 상태가 아니면 기다려야 하기 때문에 질문에서 회답까지의 시간이 길다는 것이 단점이 되고 있다.

온·라인검색은 질문이 발생함과 동시에 콤퓨터를 이용하여 검색할 수 있기 때문에 질문에서 회답까지의 시간이 매우 짧다. 더구나 단말장치를 사용하기 때문에 질문자 자신에 의한 이용도 가능하다. 그러나 사용기기가 크고, 비용도 많이 들고, 프로그래밍도 커서 문헌검색에는 사용되고 있지 않다.

기기로서 최저,

主記憶 48K字
자기디스크 2대
타이프라이터 1대

가 검색하는 데 필요하다.

D. 기계검색처리의 흐름[10]

도—12는 배치(Batch) 처리에 의한 기계검색 시스템의 일반적인 재너럴 플로우 차아트[4]이다.

질문은 카아드 또는 종이테이프에서 인풋된다. 이때 1회의 검색으로서 될 수 있는 한 많은 질문에 대한 검색을 할 수 있도록 하는 것이 콤퓨터의 사용효율이 좋다.

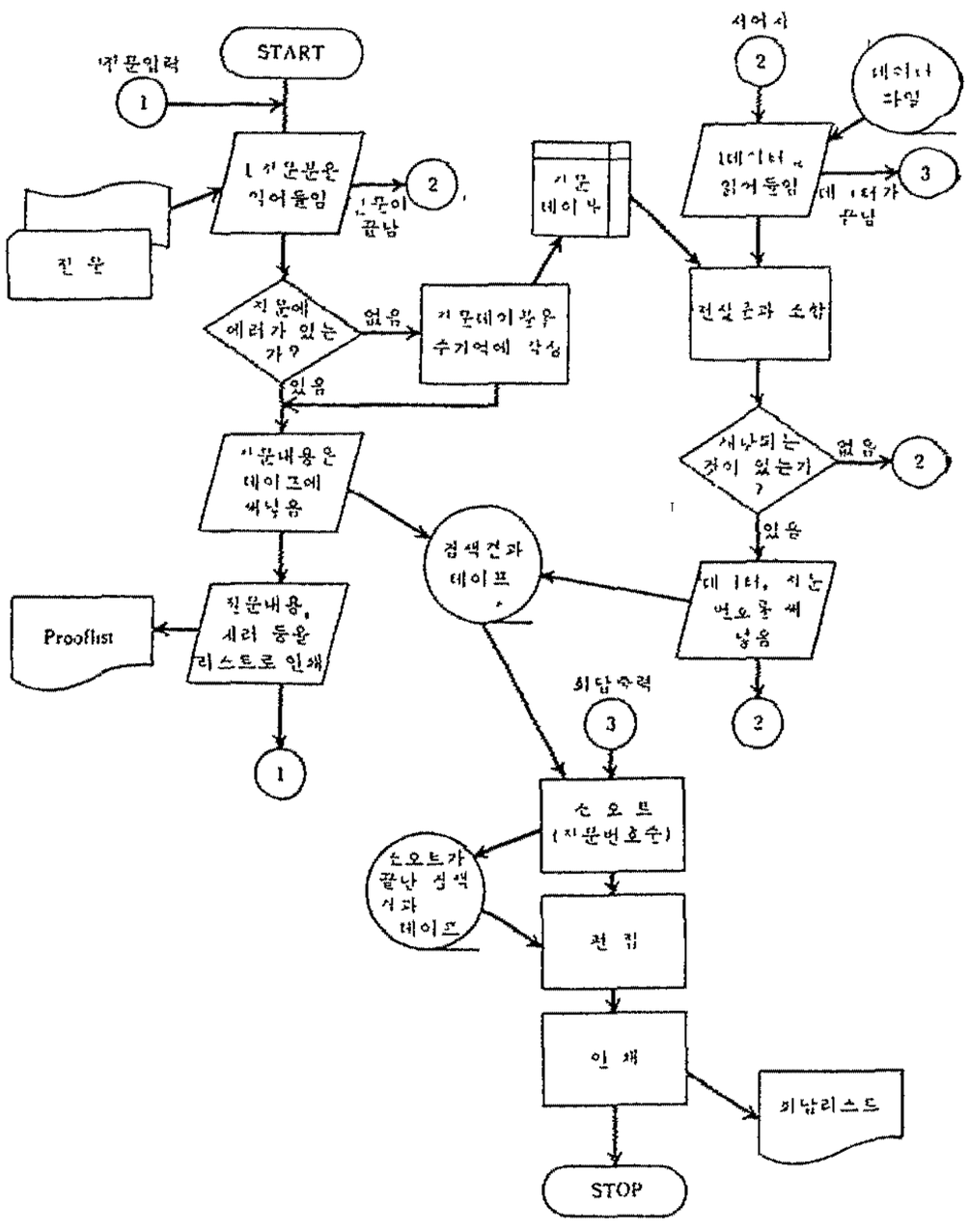
읽어 들인 질문은 분석되어, 주기억장치 내의 질문 테이블에 콤퓨터가 처리하기 쉬운 형태로 격납된다. 이 때문에 각 질문에는 다른 질문과 구별하기 위하여 질문번호를 붙여둔다. 한편 질문내용을 이용자에게 넘겨 주는 회답서에도 인쇄할 수 있도록 검색결과를 써 넣는 검색결과 테이프에도 아웃풋(Output) 해 둔다.

전질문을 다 읽어 들이거나 또는 主記憶의 질문 테이블이 완전히 메워지면 검색을 시작한다.

데이터 파일에서 1건씩 데이터를 읽어 들여서, 전질문과 조합하여 해당되는 질문이 있으면, 검색결과 테이프에 그 질문번호와 데이터를 아웃풋한다. 동일 데이터가 2개 이상의 질문에 해당하면 그 수만큼 써 넣는다.

1건의 데이터에 대하여 전질문과의 조합이 끝나면, 다음 데이터를 읽어 들여서 다시 질문을 하나씩 조합한다. 전데이터와의 조합이 완료되면 각 질문마다 검색된 데이터의 건수를 써 넣고, 탐색을 완료한다.

검색결과 테이프상에는 처음으로 전질문의 내용이 질문번호순으로 들어오고 다음에 검색된 데이터가 파일과



도—12 기계검색 제너럴 플로우 차아트[14]

같은 순서로 들어오고 끝으로 각 질문의 검색된 회답수가 질문번호순으로 들어오게 된다. 그러나 회답 리스트는 질문별로 질문내용, 회답수, 검색결과의 순으로 인쇄되지 않으면 안된다. 따라서 이와 같은 순서가 되도록 소오트하면서 편집·인쇄하여 완료한다.

온·라인 검색은 하나의 단말장치에서 1회에 1질문만 인풋한다.

질문은 분석되지만 이때는 온·라인으로서 데이터 파일과 연결되어 있기 때문에 질문 중의 키이워드를 포함하고 있는 문헌수 등을 직접 알 수 있어 질문을 개선할 수 있다. 탐색을 하고 완료하면, 희망하는 순서로 해당 데이터를 단말장치에서 얻을 수 있다.

E. 조합방식과 판단조건

기계검색의 기본은 질문과 데이터의 조합이다.

조합방법에는 다음 3 가지가 있다.

(1) 전행비교
(2) 부분비교
(3) 스캔(Scan)

전행비교는 데이터의 대상항목과 질문의 키이워드를 비교할 경우에 어느 쪽인가 행수가 짧은 쪽은 긴 쪽과 행수를 맞추기 때문에 그전 또는 후에 적당한 문자(보통 알파벳의 경우는 뒤에 공백, 숫자의 경우는 앞에 ○을

넣는다.)를 채워넣고 비교하는 방법이다. 예를 들면

데이터 : Computer

질 문 : Compute

의 경우, 질문의 Compute의 끝에 1자 공백을 넣는다. 따라서 이 경우는 일치하지 않게 된다.

이것이 유효한 것은 숫자를 대소 판별코자 할 경우이다. 또한 자연어를 사용할 경우에는 노이즈(Noise)를 방지하는 하나의 방법이 된다.

부분비교는 데이터 중의 대상항목이 질문의 키이워드보다 길 경우에 대상항목 내의 일부분과 비교하는 방법이다. 종류로서는 語頭, 語尾, 中間의 3 종류가 있다.

語頭에 의한 경우에는 항목의 좌단에서부터 비교하기 때문에 예를 들면,

Computation
Computational
Compute
Computer
Computing

등을 일괄하여 Comput로서 비교하고자 하는 방법이 바로 이것이다.

語尾의 경우는 항목의 좌단에서 비교한다.

중간의 경우는 항목의 중간의 일정 위치에서 우 또는 좌로 비교한다.

스캔은 질문의 키이워드와 같은 문자의 집단(Retrieval 처럼, 字種, 순서가 명확하게 되어 있는 것)이 데이터의 항목 내에 어딘가에 있으면 되기 때문에 부분비교의 어느 경우에도 좋다.

이 외의 조합방법으로서는 질문의 키이워드중 어떤 문자는 다른 문자로 바꾸어도 좋다고 지정하는 방법이다. 예를 들면, 온 라인과 온—라인(On-Line)은 동일하므로 한쪽만을 지정할 수 없기 때문에 사이의 공백을 하이픈으로 바꾸어 비교한다.

또한 검색 시스팀은

Color와 Colour

Center와 Centre

등도 동일한 것으로 처리할 수 있도록 하는 것이 좋지만, 곤란한 점이 많아 디소오러스를 사용하여 커버하고 있다. 조합한 결과 해당하는가 아닌가를 결정하는 것은 판단조건이다.

판단조건에는 일치(=), 부정(≠), 대소(<, ≦, >, ≧), 사이(Between) 등이 있다.

일치는 질문의 키이워드와 데이터와의 문자의 집단이 동일한 경우에 해당한다. 이것은 일반적으로 가장 많이 사용하고 있다.

부정은 동일문자의 집단이 전연 없는 경우에 해당한다. 대소는 일반적으로 수치 데이터에 사용되는 것으로 질문 키이워드와 데이터를 비교하여 대, 소, 이상, 이하가 있을 경우에 해당한다.

사이도 수치 데이터에 사용되는 것으로서 상한과 하한을 동시에 결정하는 것이다.

F. 탐색 논리

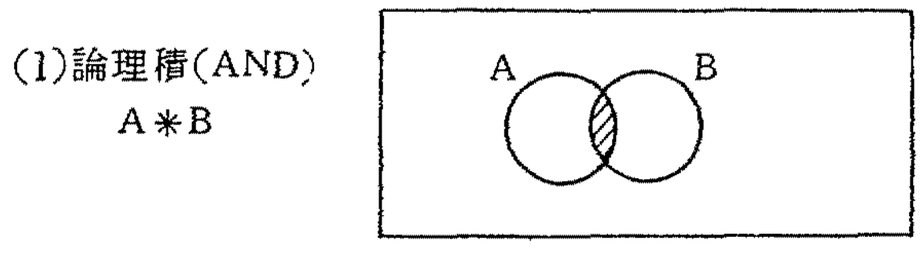
질문의 키이워드가 하나하나 해당되면, 다음에 질문전체가 데이터를 만족하는가 어떤가를 판정하지 않으면 안된다. 일반적으로 질문 키이워드는 하나의 질문에 대하여 2 개 이상이 사용되고 있다. 이런 관계를 표시하는 데는

(1) MUST, NOT, MAY

(2) Weight

(3) 부울대수를 사용한 논리식의 3가지 방법이 있다.

(1)의 방법은 오래전부터 사용되어 온 것으로서 불필요한 것에는 NOT를, 필요한 것에는 MUST를, 질문 키이워드 중 일정수만 있으면 좋다는 것에는 MAY를 사



도—13. 논 리 관 계

용한다. 이 중에서 NOT가 가장 강하고 다음에 MUST 그리고 MAY의 순으로 약하다.

이 방법은 간단하고, 보통의 질문이라면, 이것으로 충분하다. Weight에 의한 방법은 질문 키이워드 혹은 데이터 중의 키이워드에 Weight를 정해둔다. 그리고 질문에는 Weight가 정해진 질문 키이워드 외에 기준치를 정해둔다. 그리고 데이터와 조합하여 해당하는 키이워드의 Weight를 가하여 기준치를 넘는가 여부를 보아, 그 데이터를 취할 것인가 버릴 것인가를 결정한다.

부울대수를 사용한 논리식에서는 질문키이워드를 論理積(AND), 論理和(OR), 부정(NOT) 등의 논리연산자를 사용하여 연결한다. 이러한 논리연산자의 기능은 기호 논리학에서 사용되는 것과 동일하다(도-13).

일반적으로 이 방법이 가장 많이 사용되고 있다.

G. 질 문

질문은 먼저 질문자 자신이 하나의 개념으로서 갖고 있다. 이것을 구체화하기 위하여 처음에 문장화 혹은 키이워드의 나열을 행한다. 이렇게 하면 질문의 중심이 명확해지고 범위를 한정하기도 쉽게 된다. 그리고 끝으로 콤퓨터가 인풋하여 분석할 수 있는 형으로 정식화[11](定式化)한다.

정식화란 키이워드로서 사용되는 용어의 선택, 디소오러스 등에 의한 용어의 확장, 판단조건의 지정, 논리식의 조립 등 일련의 작업을 말한다.

질문 키이워드는 데이터와 직접 비교하기 때문에 이의 선택은 가장 중요하다.

먼저 데이터 중에 사용하고 있는 날말을 사용하지 않으면 안되지만, 질문 키이워드로서는 사용문자의 제한, 자수(字數)의 제한은 될 수 있는 한 없도록 하는 것이 좋다. 또한 단어만이 아니라 구로 된 키이워드도 사용할수 있어야 한다.

예를 들면, On-Line이란 구로서 검색하고 싶을 경우 이것을 나누어 On과 Line을 논리적으로서 행하면 관계없는 데이터가 나올 가능성이 충분히 있다.

질문 키이워드가 결정되면 다음에 이것과 데이터와의 관계를 표시하는 판단조건을 관계연산자로 명시해야 한다. 세번째로 질문 키이워드는 어떤 항목을 대상으로 하여 탐색할까를 결정한다. 동시에 질문 키이워드간의 탐색논리도 명시한다.

끝으로, 항목 간의 탐색논리 관계를 명확히 하여 질문의 구성을 완료한다.

이러한 질문에 질문자의 성명, 소속 또는 질문에 대한 코멘트(Comment)를 넣어 하나의 질문을 작성한다.

H. 회답 리스트

회답 리스트는 검색 서어비스의 제공 리스트로서, 이용자에게 넘겨 주는 것이기 때문에 중요하다.

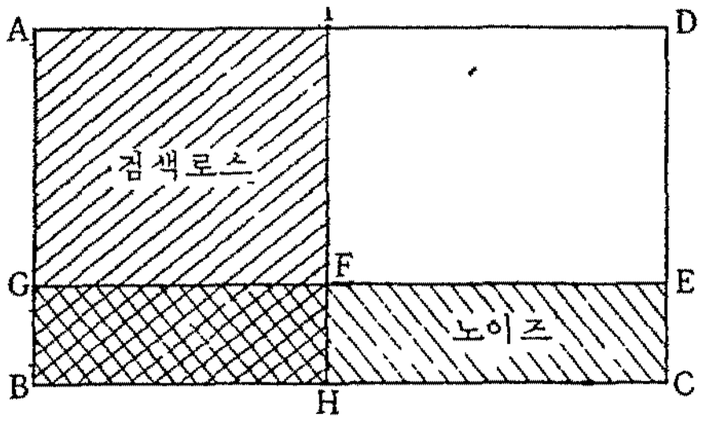
회답 리스트에는 보통 질문자의 성명, 소속 등이 먼저 인쇄되고 다음에 질문의 내용과 회답수를, 끝으로 검색결과를 아웃풋한다. 여기에 대한 레이아웃(Layout)을 결정하는 것은 중요한 일로서 실제에 있어서는 타이프라이터로 쳐서, 검토하는 것이 보통이다.

개개의 질문에 대한 검색회답의 인쇄순서는 여러 가지가 있지만, 보통은 데이터의 축적순의 최신의 것에서, 소급하여 인쇄하고 있다. 그러나 웨이트(Weight)가 정해져 있으면, 웨이트가 높은 것부터 인쇄하는 방법도 있다.

I. 검색효율의 평가

검색 시스팀을 평가하는 방법에는 검색효율, 검색시간, 검색비용 등 여러 가지가 있지만, 가장 기본적인 것은 검색효율이다.

일반적으로 검색 시스팀은 필요한 정보는 모두 뽑아내고, 불필요한 정보는 뽑지 않는 것이 필요 조건이다.



도—14. 검 색 효 율

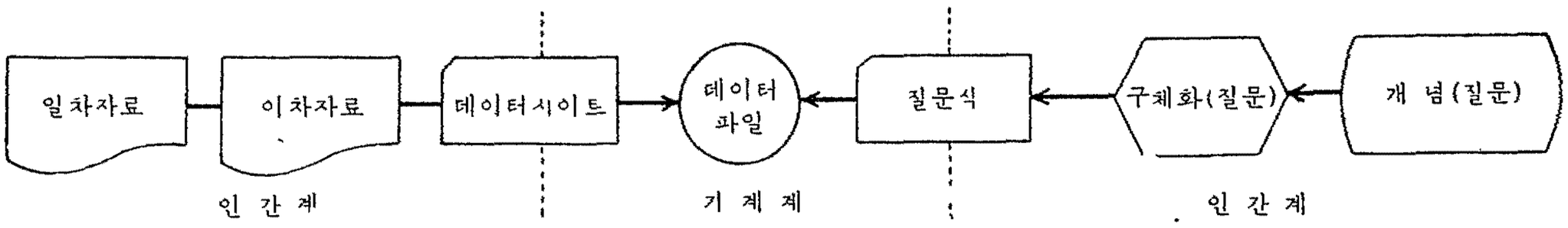
도-14에 있어서

ABCD : 축적되어 있는 전문헌

ABHI : 적합문헌

BCEG : 검색된 문헌

이라고 하면, 검색효율을 나타내는 2개의 식을 얻을 수 있다.



도—15. 정보 검색계

검색율(재현율 : Recall Factor)

$$=\frac{BHFG}{ABHI}=\frac{\text{검색된 적합문헌}}{\text{전적합문헌}}$$

적합율(Relevance Factor)

$$=\frac{BHFG}{BCEG}=\frac{\text{검색된 적합문헌}}{\text{검색된문헌}}$$

이것을 나머지 부분에서 보면,

AGFI : 검색 로스(Loss)

CEFH : 노이즈(Noise) (불필요한 정보)가 된다.

검색 로스는 이용자에게 때로는 치명적인 에러가 될 수도 있다.

노이즈는 이용자가 선택할 수는 있지만 역시 불필요한 것이다.

에러가 발생하는 것은 검색 시스팀 전체의 어딘가에 결함이 있기 때문이다.

도-15에 있어서 기계계에는 문자와 문자의 비교이기 때문에 탐색상의 에러는 없다. 가장 큰 에러는 데이터 시이트의 작성과 질문식의 작성에 있어서 용어가 통일되어 있지 않아서 생기는 것과 일관성이 없는 데이터 시이트의 작성에 기인하는 경우가 많다.

전자는 데이터시이트 작성에 있어서 용어를 통일하여, 여기에 따라서 질문을 행하든가 혹은 질문싯점에서 디소오러스를 사용하든가 하여 해결할 수 있다.

후자는 주로 형식상의 문제로서 데이터 시이트 작성 규정을 명확히 하여 두면 좋다.

이상이 주로 기계검색에 있어서의 에러이지만, 일반적인 기계검색 시스팀에서 생기는 에러의 원인은 이외에도 많이 있다.

<맺 는 말>

콤퓨터에 의한 정보검색의 대강을 해설하였지만, 기계식 디소오러스와 디렉트 액세스 메모리(Direct Access Memory)의 사용, 국어와 한자의 문제 및 코스트 문제 등 언급된 것보다는 생략한 것이 더 많다.

계속적인 연구의 필요성을 절감하면서 이 방면에 관심있는 분에게 다소나마 도움이 되었으면 한다.

인 용 문 헌

① 河野德吉編. 情報檢索の知識, 日經文庫, 日本經濟新聞社, 1968

② 林省三, 福島芳直 等 JICSTにおける ON-LINE 檢索システムの設計, 1. ファイル構成の基本的 諸問題, 情報科學技研集會發表論集〔7〕, pp. 89～98, 1971

③ 橋本昌幸, 中嶋淳. Computerによる IR (1). トクメンテーション研究, 20, 5, pp. 145～151, 1970

④ 橋本昌幸, 中嶋淳. Computerによる IR (2). ドクメンテーション研究, 20, 7, pp 203～210, 1970

기타 참고 문헌

⑤ 櫻井宣隆. 利用業務の機械化, 情報檢索の機械化. 現代の圖書館, 8, 1, 39～48, 1970

⑥ 日本科學技術情報センター編. JICST情報管理中級實務講座. 東京, 同センター, pp. B 7～18, 1971

⑦ Messina, C. G., Hilsenrath, J. EDPAC : Utility Programs Computer-Assited Editing, Copy Production and Data Retrieval NBS Tech. Note〔470〕, p. 81, 1969

⑧ Helbich, J. Direct Selection of Keywords for the KWIC Index. PB-179833 pp. 1～13, 1968

⑨ Senko, M. E. Information Storage and Retrieval Systems. Advan. Inform. Syst. Sci., 2, pp. 229～281, 1969

⑩ 高橋達郎. 情報檢索, 東洋經濟新報社, 1969

⑪ Cowles, C. C. Information Please-storage and Retrieval Systems. Data Process Mag., 12, 12, pp 29～32, 1970