지식을 이용한 특정 문서의 논리 구조 추출에 관한 연구

손영우 · 남궁재찬*

A Study on the Extraction into the Logical Structure of a Specific Document using Knowledge

Young Woo Shon and Jae Chan Namkung

〈요 약〉

본 논문은 특정문서에서 문서가 갖고 있는 일반적인 지식을 이용하여 논리적 항목을 추출하는 방법에 관한 연구이다. 먼저 입력된 문서의 영역 분할, 분리자 추출, 그리고 문자와 비문자를 구별하였다. 논리구조 추출단계에서는 구별된 요소의 상대적 크기, 위치 및 전후 블록들의 연관성에 관한 지식을 이용하여 각 블록들을 레이블링 하였고, 레이블된 항목들의 위치정보값을 이용하여 각 항목들을 자료화하였다. 마지막으로, 오분류된 항목에 대해서는 배치기술자를 이용한 검증을 통해 정정하였다. 본 논문에서 구현한 방법으로 실험한 결과 96.5%의 논리항목 추출율을 획득함으로써 그 유효성을 입증하였다.

* 광운대학교 전자계산기공학과

I. 서론

오늘날 대부분의 정보는 종이에 인쇄된 형태로 제공된다. 이러한 형태의 정보를 전자정보로 변환하여 데이타베이스화 함으로써, 원하는 정보를 편리하고, 신속하게 찾아볼 수 있도록 하는 것이 요구된다. 이를 수작업으로 할 경우 비용이 많이 들 뿐만 아니라 문자와 그림이 혼합되어 있는 복합문서(compound document)의 경우 입력이 불가능하다. 따라서 복합문서를 컴퓨터로 처리하기 위해 영상입력장치(스캐너)를 사용하여 보다 효과적으로 문서의 정보영역 분할 및 재배치를 가능하게 하는 시스템이 필요하게 되었다.

문서의 영역분할에 대해서는 많은 연구가 활발히 이루어지고 있으며[WANG & SRIHARI 1989, 가록현 외 3인 1992, 신현관 1992] 대표적인 방법으로는 RLSA(Run Length Smoothing Algorithm) 및 RXYC(Recursive X-Y Cuts)를 들 수 있으나 이 방법들은 처리속도상의 문제가 있다. 이를 개선한 것으로 문자의 좌우 경계위치를 이용하여 입력영상에 대해 단 한번의 스캐닝으로 문자와 그림을 동시에 추출하는 방법이 제안[이인동 외 2인 1991]되었지만, 다단(multi-column) 편집된 문서의 경우 문맥과는 무관한 순서로 문자를 추출하는 문제점이 있다.

최근 표준화된 문서구조인 ODA(Open Document Architecture)[ISO 1989]에서는 문서를 배치구조(layout structure)와 논리구조(logical structure)로 구분하고 있으며, 문서의 작성 및 편집 관점에서 볼 때, 논리구조에 중점을 두고 있다. 이와 같이 논리구조를 이용한 데이타베이스의 경우, 문서화상으로부터 논리구조를 추출하는 기술이 요구된다. 논리구조 해석에 관한 연구로는 배치구조와 논리구조 대응 연결 알고리즘, 인쇄된 정보에 대한 문자열의 논리적 분할에 관한 연구 등이 있지만[KISE et al. 1993, LUO et al. 1992, YAMADA et al. 1993], 다단의 장, 절, 단락구조를 상세히 해석하여 논리구조화한 연구는 지극히 미비한 실정이다.

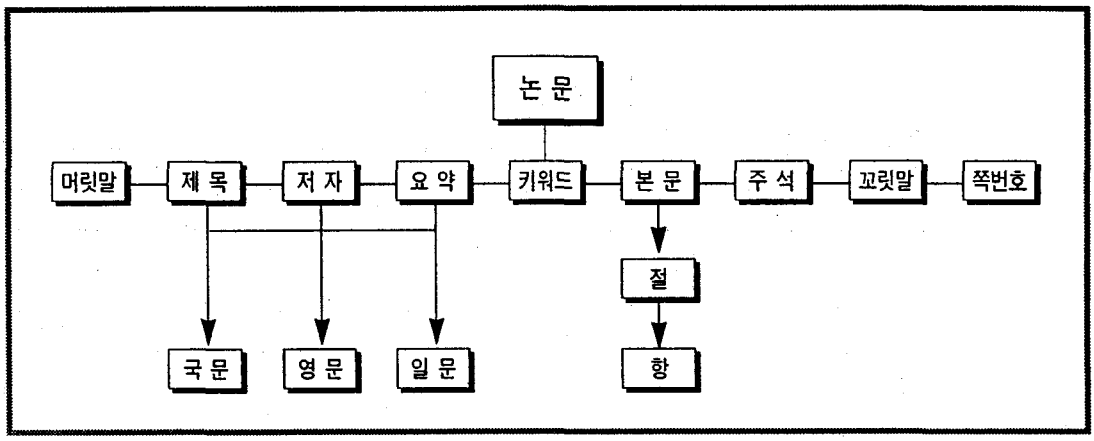
따라서 본 연구에서는 다단으로 구성된 복합문서의 논리적 구조를 추출하는 방법을 제안하였다. 문서영역을 분할하기 위한 블록화는 2단계로 이루어지는데, 제1단계에서는 백화소 열(white pixel streams)의 임계치(threshold)를 이용하여 일단 문서의 칼럼을 블록화 한 후, 제2단계에서 분리자(separator)를 찾아, 논리구조 파악단계에서 주석(footnote)부분을 인지하기 위해 나누어진 블록들에 대해서만 다시 분할한 후, 각 블록들을 병합하는 알고리즘을 제안하였다. 또한 문자열 영역(character string region)과 비문자 영역(non-character region)을 분리하는 방법은 블록화에 이용된 정보 및 본래의 문자열 영역과 비문자 영역의 특징을 이용하여 분리하였으며, 마지막으로 각 영역이 가지는 특징을 이용하여 문서의 각 영역에 대해 논리적 정보를 추출하는 새로운 방법을 제안하였다.

본 논문은 서론, 문서의 논리구조 및 배치구조 분석, 영역분할 및 비문자 영역 추출, 논리적 구조 이해, 실험결과 및 고찰, 결론으로 이루어져 있다.

II. 문서의 논리구조 및 배치구조 분석

일반적인 데이타베이스 구성항목들은(저자, 소속기관, 제목, 서지사항들(잡지명,권,호,년), 키워드, 요약 등) 학회논문지의 제 1쪽에 잘 나타나

〈그림 1〉 일반적인 논문 제1쪽의 논리구조



있다.

따라서 본 연구에서는 문서의 논리구조 및 배치구조 분석에 있어 그 범위를 각 학회논문지의 제 1쪽으로 한정하므로써 보다 명확하게 분석할 수 있도록 하였다. 학회논문이나 신문, 잡지 등의 논리구조와 배치구조를 분석하기 위해서는 문서가 가지고 있는 계층구조를 모델화할 필요가 있다. 문서의 논리구조는 장, 절 등의 논리대상을 노드로 하는 트리구조에 의해 표현이 되는데 그림 1에는 각 논문지 제 1쪽의 논리구조를 나타내었다. 또한 논리구조에는 가변적인 요소가 있을 수 있는데 그 중에는 제목 및 저자명의 갯수, 키워드의 유무 등이 논리구조에 있어서의 변동사항이 될 수 있다.

논문 제1 쪽의 배치구조는 문자열이 좌측에서 우측으로 나열되어 있으며, 구성요소는 머리말 (header), 제목(title), 저자(author), 요약(abstract), 키워드(keyword), 본문(paragraph), 분리자

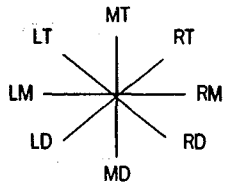
〈표 1〉 특징기술

구성요소	공 통 점	차 이 점
머릿말 (header)	문자 블럭으로 구성 문자 및 기호로 이루어짐 문서의 최상단에 존재	좌우측에 대한 위치 무관 제목과 구분하기 위해 분리기 사용함
제 목 (title)	문서내 가장 큰 글자크기 문자열길이가 일정치 않음 헤더사이에 큰 백영역 존재 중앙으로 정렬되어 있음	국내, 일본논문 영문포함 헤더가 없는 논문에서는 최상단에 존재
저 자 (author)	문자열 길이가 일정치 않음 저자명의 수가 일정치 않음 위치가 일정치 않음	국내 논문에만 영문 저자명이 존재
요 약 (abstract)	문자열의 블럭으로 구성 블럭크기가 일정치 않음	국내 논문은 영문요약 포함 미국 논문은 두 단(column)으로 구성
키워드 (key word)	대부분 없지만 있는 경우에는 요약 하단에 위치	
본 문 (paragraph)	본문 시작을 알리는 제목 있음 대부분 2개의 단으로 구성 문자, 수식, 그림으로 구성	1단으로 구성된 것이 있음 (미국논문) 제목 위치가 좌측 또는 중앙
분리기 (separator)	대부분 좌측 하단에 위치 하단에 주석 포함	분리기가 없는 논문도 있음 (미국 논문) 분리기가 상단에 위치하는 것도 있음 (국내 논문)
주 석 (footnote)	문자 및 기호로 구성 대부분 분리기 하단에 존재 문서 하단에 존재	내용의 차이가 있음 (본문설명 또는 저자 소개) 분리기 없이 존재 가능
쪽번호 (page numbe)	숫자로 구성 위치에 따라 좌, 우쪽 구분가능	쪽번호의 짝수, 홀수에 따라 좌상, 좌하, 우상, 우하, 중앙하단에 위치
꼬릿말 (footer)	문자로 구성 미국 논문에만 있음	중앙 하단이나 짝수, 홀수에 따라 좌하, 우하로 구성

〈표 2〉 위치관계와 구성요소

구 분		위치관계							
		국내				미국		일본	
		1	2	3	4	1	2	1	2
머릿말	논문지명	MT	LT, RT			LT	LT		MT
	날 짜	MT	LT, RT			LT	LT		RT
	권	MT	LT, RT			LT	LT		LT
	주	LT	LT, RT	LT	RT			LT	
	호	MT	LT, RT				LT		LT
	페이지	LT, RT	LT, RT	LD, RD	LD, RD	MD	MD	LD, RD	MD
	종합페이지	MD							
	발행처					LD, RD	RT		
제 목	제 목	MT	MT	MT	MT	MT	MT	MT	MT
	영문제목	MT	MT	MT	MT			MT	
저자명	저 자 명	MT	RT	MT	MT	MT	MT	MT	MT
	영문저자명	MT	RT	MT	MT			MT	
	소 속	LD	LD	LD	LD	MT	MT		LD
	회원종류	LD	LD	LD	LD			MT	
	접수일자	LD	LD		LD	MT	MT		
요 약	요 약	MT	MT	MT		MT	MT	MT	MT
	영문요약	MT	MT	MT	MT				
키워드	키 워 드				MT		MT	MT	
분리기	분 리 기	LD	LD	LD	LD		LD	LD, RD	LD
꼬릿말	꼬 릿 말					LD, RD		MD	

구 분	국 내	미 국	일 본
1	전자공학회	CVGIP *	전자정보통신학회
2	정보과학회	PR **	정보처리학회
3	통신학회		
4	전기학회		



* CVGIP : Computer Vision, Graphics, and Image Processing

** PR : Pattern Recognition

LT : Left-Top MT : Middle-Top RT : Right-Top

LM : Left-Middle RM : Right-Middle

LD : Left-Down MD : Middle-Down RD : Right-Down

(separator), 주석(footnote), 쪽번호(page number), 꼬릿말(footer) 등으로 이루어져 있다. 표 1에는 각 구성요소의 공통점 및 차이점에 대한 특징을 기술하였다. 머리말의 경우 문자 및 기호로 이루어진 문자 블럭으로 구성되며, 문서의 최상단에 존재한다는 공통점이 있는 반면에, 좌우측에 대한 위치에 무관하고 제목과 구분하기위해 분리차를 사용하는 경우도 있다.

제목은 문서내에서 가장 큰 글자로 문자열의 길이가 일정하지 않고 헤더사이에 큰 백영역이 존재하며, 중앙에 잘 정렬되어 있다는 공통점이 있다. 키워드는 존재하는 경우 대부분 요약하단에 위치하며, 본문은 시작을 알리는 제목이 있고 대부분 2개의 단으로 구성되며, 문자, 수식, 그림 등으로 구성된다는 공통점이 있다.

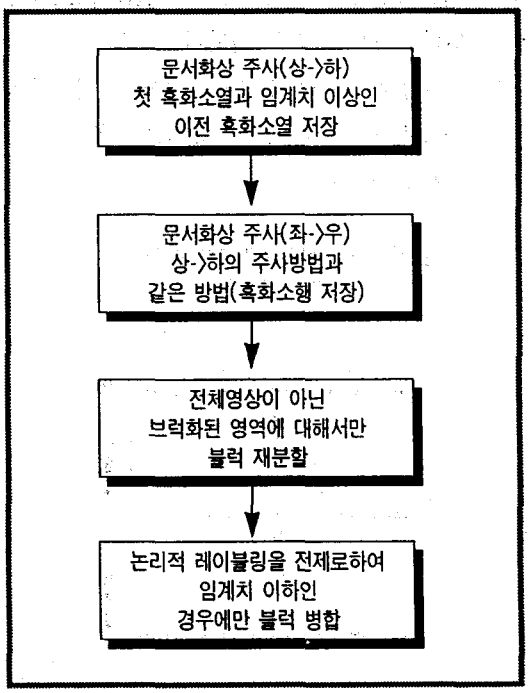
각 블록의 논리적 레이블링을 하기 위한 특징기술 특히 위치관계에 대한 것은 표 2에 나타내었다. 여기서, L은 Left, M은 Middle, R은 Right, T는 Top, D는 Down을 의미한다. 예를 들면, 머리말의 논문지 명칭이 'MT'인 경우는 'Middle-Top 위치에 논문지 명칭이 존재' 함을 의미한다.

III. 영역분할 및 비문자 영역추출

문서의 흑화소 영역을 올바르게 블록화하여 추출하기 위해서는 기울기를 정정해야 할 필요가 있으며, 본 논문에서는 문자열사이의 상호상관(correlation)을 이용[HONG 1993]하여 기울기를 정정하였다.

입력된 문서의 배치구조를 통해 논리적 레이블링(labelling)을 하기 위한 전 단계로서 문서의 각 흑화소 영역에 대한 블록화가 되어야 하는데, 본 논문에서는 CRLCA(Counted Run Length Cutting Algorithm)를 개선한 백화소 열의 임계치를 이용한 블록화방법을 그림 2에 나타내었다. 특히, 상향식(bottom-up)과 하향식(top-down)의 두 가지를 합한 혼합식(hybrid)을 이용하고 있으며, 제 1단계에서 백화소의 임계치에 의해 먼저 블록들을 큰 영역으로 구분한 후, 제 2단계에서 그 블록들을 다시 작은 블록으로 나눈다. 이를 블록 병합과정에서 각 블록들을 연결한다. 일반적으로 논문 제 1쪽의 배치구조는 요약부분까지는 일단이며, 그 이후는 이단으로 구성되어 있다.

〈그림 2〉 블록화 알고리즘

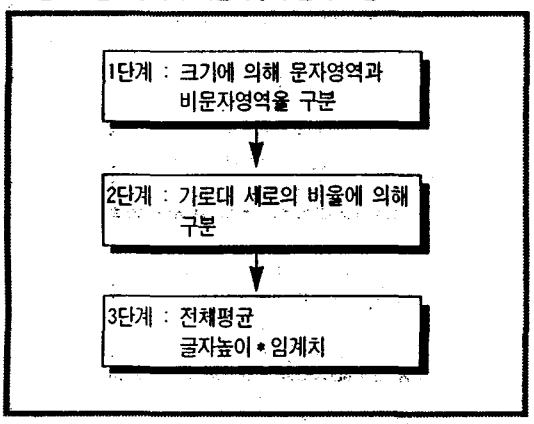


문서영상으로부터 비문자영역을 분리하는 방법은 그림 3에 보인바와 같이, 면적의 크기에 의해 영역을 대분리하고, 비문자영역 중 문자영역이 될 수 있는 후보를 가로 대 세로의 비율에 의해 중분류하였으며, 마지막으로 전체 평균글자 높이

를 이용하여 문자영역 중 비문자영역을 분리하였다. 문서분석을 위한 문자열영역과 비문자영역의 특징들은 다음과 같다.

(1) 문자열영역은 문자열로 이루어져 있고, 문자열 간격은 대부분 일정하다.
(2) 문자열은 문자열사이의 간격에 의해 구분된다.
(3) 단어는 문자로 구성되어 있고, 간격이 거의 일정하다.
(4) 문자열영역은 비문자영역보다 크기가 현저히 작다.
(5) 비문자영역의 가로, 세로비율은 일정치 않다.
(6) 비문자영역의 하단에는 문자열(caption)이 있다. (예외, 도표는 상단에 위치)
(7) 문자열영역은 한글, 영어, 일어, 한자, 기호 등으로 비문자영역은 그림, 도표, 테이블, 그래프로 구성된다.

〈그림 3〉 문자영역과 비문자영역 분리 흐름도



IV. 논리적 구조 이해

1. 논리적 구조의 정의 및 필요성

논리적 구조 이해란 문서의 기하학적 구조(geometric structure)를 논리적 구조(logical structure)로 변환하는 것을 말한다.

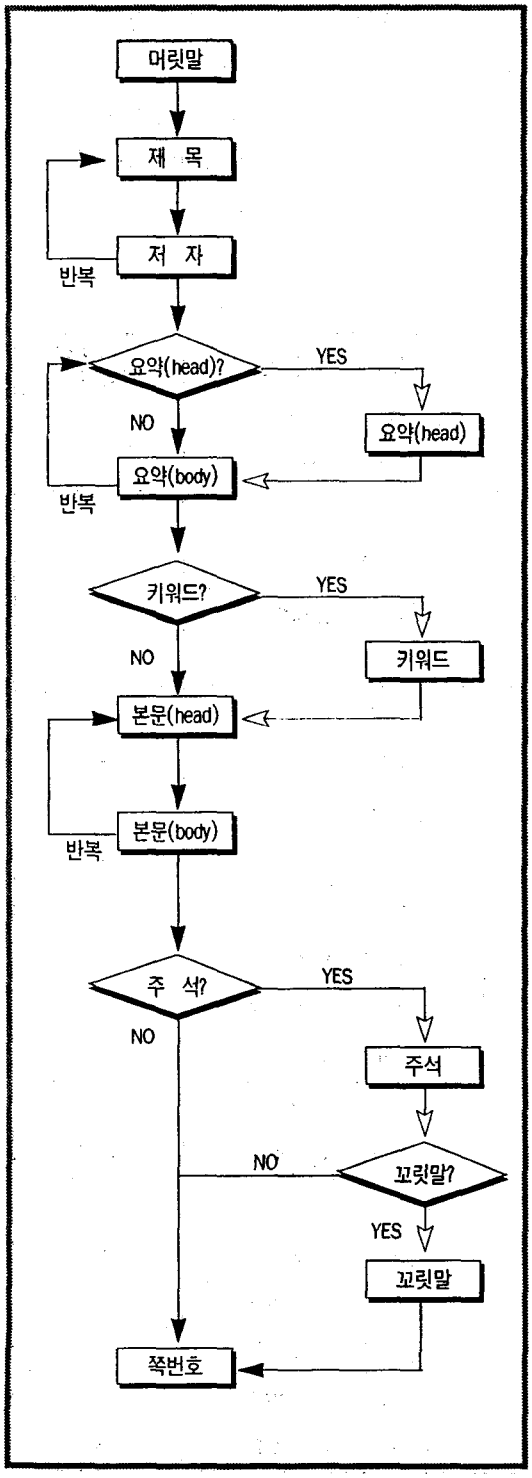
본 연구에서 실험 데이터로 학회 논문지 제 1쪽을 컴퓨터로 입력하는 목적은 관련정보를 데이타베이스화하여 선행연구를 손쉽게 찾아보기 위한 것이다. 따라서 논리적 구조를 정의하는 것(추출항목을 정하는 것)이 문서이해의 관점에서 매우 중요한 것이며 문서의 논리적 변환의 핵심적인 사항이다.

2. 논리적 레이블링

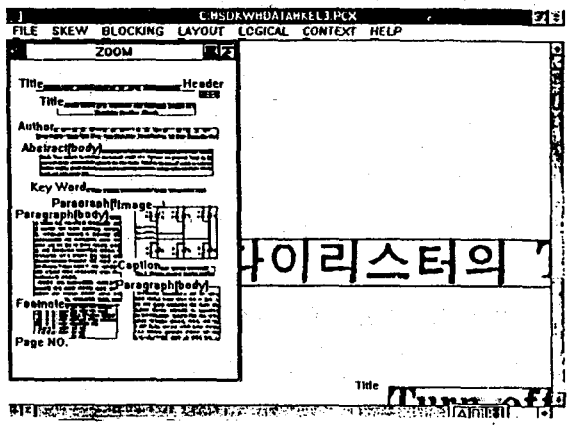
항목의 논리적 레이블링 단계는, 각 블록들의 특징들을 가지고 블록의 병합함에 있어서 완전하게 이루어지지 않은 부분에 대해 보완처리하는 단계이다. 각 블록은 두 개의 범주, 즉 머리(head)와 몸체(body) 중 하나로 분류된다. 논문 제 1쪽에서 나타나는 대표적인 것으로는 요약부분과 본문부분이며, 그림 4에는 논리적 레이블링을 위한 흐름도를 나타내었다.

이는 문서의 분석결과에 따른 논리 요소들에 대한 레이블링 흐름도로, 그림 5에 레이블링 실행결과의 예를 보였다. 기하학적 구조를 논리적 구조로 변환하기 위해 필요한 배치기술자를 표 3에 나타내었다. 여기서, TW는 Total Width of document image를, ACH는 Average Height of Character string을 의미한다. 예를들면, 블럭의 가로 배치기술자가 '〈 TW/10' 인 경우는 '블럭의 가로 폭이 전체 문서 이미지 폭의 1/10 이하'가 됨을 의미한다. 또한 블럭의 세로 배치기술자가 '〉 ACH(2/3)' 인 경우는 '블럭의 세로 높이가 문자열 평균높이의 2/3 이상'이 됨을 의미한다.

〈그림 4〉 논리적 레이블링을 하기 위한 흐름도



〈그림 5〉 블록의 논리적 레이블링 예



〈표 3〉 블록을 구분하기 위한 배치기술자

배치기술자 (종류)	표 현	배치기술자 (형태 및 범위)
블럭의 위치	LT	Left Top
	MT	Middle Top
	RT	Right Top
	LM	Left Middle
	RM	Right Middle
	LD	Left Down
	MD	Middle Down
	RD	Right Down
	MC	Middle Center
블럭의 가로	XS(smallest)	〈 TW/10
	S(small)	〈 TW/3 & 〉 TW/5
	M(medium)	〉 TW/3 & 〈 TW(2/3)
	L(large)	〉 TW(2/3)
블럭의 세로	SS(small small)	〈 ACH(2/3)
	S(small)	〉 ACH(2/3) & 〈 ACH
	M(medium)	≒ ACH
	L(large)	〉 ACH & 〈 ACH(3/2)
	LL(large large)	〉 ACH(3/2)
	XL(largest)	〉 2(ACH)
블럭의 시작과 끝점	X_start 1	〈 TW/2
	X_start 2	〉 TW/2
	X_end 1	〈 TW/2
	X_end 2	〉TW/2
Indentation (블럭 左上)	유, 무	

주) TW : 전체 문서 이미지의 폭(Total Width of document image)
ACH : 문자열의 평균 높이 (Average Height of Character string)

3. 지식구축에 의한 검증

추출된 항목들에 대해 검증하는 방법은 시각적으로 확인하는 방법과 지식을 구축하여 검증하는 두 가지 방법이 있는데, 전자는 최종적으로 항목을 확인하는 것이고, 후자는 오추출된 부분에 대해 교정을 하는 방법으로 이용하게 된다. 항목 검증에 필요한 지식들에 대해 공통이 되는 지식을 만들어 오추출된 항목들에 대해 교정을 행하며, 이때 사용되는 논리항목검증을 위한 배치기술자를 표 4에 나타내었다. 이것은 각 항목에 대해 각 배치기술자가 갖는 지식을 이용하여 추출된 항목들을 검증하게 된다.

〈표 4〉 논리항목 검증을 위한 배치기술자

배치기술자 / 항목	블럭의 위치	블럭의 가로	블럭의 세로	블럭의 시작과 끝점	Indentation (블럭 좌단)
머릿말	LT, MT, RT	S, M, L	SS, S, L	X_start 1, X_start 2 X_end 1, X_end 2	무
제목	MT	L	LL	X_start 1 X_end 2	무
저자	MT, RT	M, L	L	X_start 1, X_start 2 X_end 2	무
요약(head)	MC	XS	SS	X_start 1 X_end 2	무
요약(body)	MC	L	XL	X_start 1 X_end 2	유
키워드	MC	L	S, M	X_start 1 X_end 1, X_end 2	무
본문(head)	MC	XS	S	X_start 1, X_start 2 X_end 1, X_end 2	무
본문(body)	LD, MD, RD	L	XL	X_start 1, X_start 2 X_start 1, X_end 2	유, 무
주석	LD, RD	M	L, LL, XL	X_start 1, X_start 2 X_end 1, X_end 2	무
꼬릿말	LD, MD, RD	M	M, L	X_start 1, X_start 2 X_end 1, X_end 2	무
쪽번호	RT, MD	XS	SS	X_start 1, X_start 2 X_end 1, X_end 2	무

4. 추출항목의 표현 및 자료 구축

논리적으로 레이블링된 각각의 블록들에 대한 시작점과 끝점에 대한 값 및 번호를 찾아 추출된 항목들을 자료화하여 차기가 원하고자 하는 항목들을 이용할 수가 있으며, 하단의 표 5는 그림 5의 예를 들어 기술하였고, 추출된 항목을 표현하기 위한 자료구조는 그림 6과 같다.

〈표 5〉 데이터베이스 구축

Block Number	Logical	X_start	X_end	Y_start	Y_end
0	Title	246	1690	232	287
1	Header	1740	1913	235	350
2	Title	438	1695	405	511
3	Author	236	1898	633	728
4	Abstract	265	1851	834	1075
5	Keyword	340	1778	1200	1237
6	Paragraph (head)	531	686	1357	1391
7	Paragraph (body)	206	1002	1458	2260
8	Image	1130	1886	1368	1858
9	Caption	1134	1856	1939	2017
10	Footnote	200	960	2298	2606
11	Paragraph (body)	1112	1904	2120	2617
12	Page Number	197	226	2658	2679

〈그림 6〉 추출된 항목을 표현하기 위한 자료구조

```
typedef struct tagCHAR {
        int     blknum;
        char    name[20];
        DWORD   sx, sy;
        DWORD   ex, ey;
        POS;
```

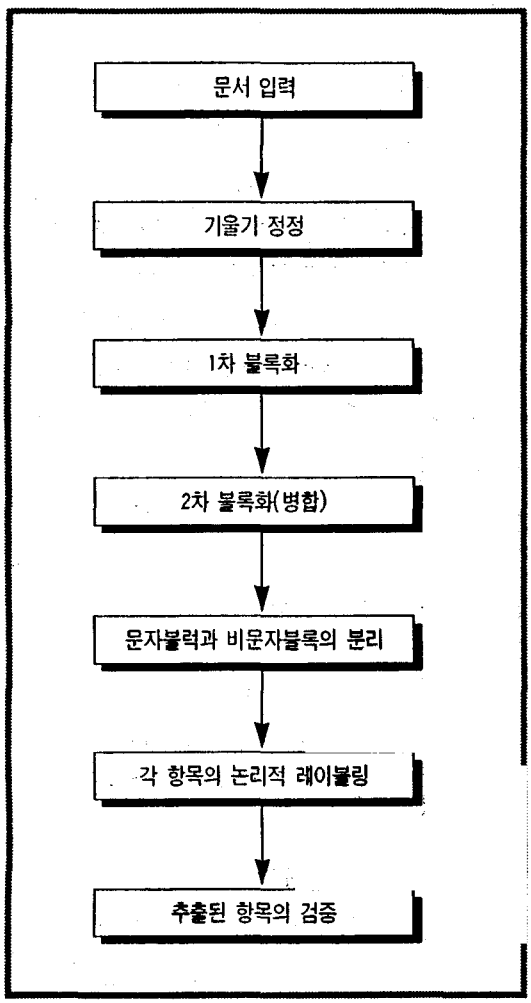
V. 실험결과 및 고찰

1. 실험결과

본 논문에서 실험데이터로 사용된 문서영상은 국내의 전자공학회, 정보과학회, 한국통신학회, 전기학회 논문지와 국외의 CVGIP(Computer Vision, Graphics and Image Processing), PR(Pattern Recognition), 日本電子情報通信學會, 日本情報處理學會의 논문 제 1쪽을 대상으로 하였으며, IBM 80486-PC 컴퓨터에서 MS-WINDOW 3.1의 SDK(Software Development Kit)를 사용하여 구현하였다.

문서영상은 HP 이미지 스캐너를 이용하여 250DPI의 해상도로 입력을 받았다. 그림 7에는 본 논문의 전체 흐름도를 보였으며, 입력 문서영상에 대한 논리항목 추출결과를 그림 8에, 최종 실험결과를 표 6에 나타내었다. 국내의 전자공학회, 정보과학회, 한국통신학회, 전기학회 논문지 등을 대상으로 실험한 결과 100%의 논리구조 추출율을 나타내었으며, CVGIP(미국)는 90%, PR(미국)과 電子情報通信學會(日本)는 91%의 추출율을, 情報處理學會(日本)는 100%로, 평균 96.5%의 논리 구조 추출율을 획득함으로써 그 유효성을 입증하였다.

〈그림 7〉 전체 흐름도



〈그림 8〉 논리항목 추출결과

〈표 6〉 최종 실험결과

학회논문지명(실험데이터)	항목개수	추출된 갯수	추출율
전자공학회	16	16	100%
정보과학회	14	14	100%
통신학회	15	15	100%
전기학회	12	12	100%
CVGIP	10	9	90%
PR	12	11	91%
電子情報通信學會	12	11	91%
情報處理學會	12	12	100%
평균 추출율	96.5%		

2. 고찰

본 논문에서는 개별문자까지 분할하여 문자와 비문자의 특성을 이용하여 구분하였으며, 논리적 구조를 추출하고 검증하는데 있어 공통적인 특징(지식)을 이용하였다. 인접한 블록들 간의 논리적 연결을 위해 각각의 실험 데이터에 대해 배치기술자를 이용하여 구분하였다. 또한, 이들에 대한 검증방법으로 지식을 구축하여 추출된 논리항목들에 대해 검증을 하였다. 국내 논문의 경우에 키워드와 요약의 몸체 부분을 구분함에 있어 단지 크기정보만을 이용해서는 곤란하며, 또한 영문논문지인 CVGIP와 일본논문지인 電子情報通信學會誌의 경우 제목과 저자의 위치 및 상대적인 크기가 일정치 않아 이를 구분하는데 어려움이 있었다.

VI. 결 론

본 논문에서는 다단으로 구성된 복합문서의 논리적 구조를 추출하는 새로운 방법을 제안하였다. 국내의 전자공학회, 정보과학회, 한국통신학회, 전기학회 논문지와 미국의 CVGIP, PR, 日本電子情報通信學會, 日本情報處理學會 논문지 등 8종의 논문지를 대상으로 특정한 구조를 갖는 논문 제1쪽의 논리적 구조를 추출한 결과 평균 96.5%의 추출율을 얻었다. 다단 문서를 대상으로 문서의 논리적 구조를 파악하였으며, 각 항목의 공통 특징정보를 조사하여 지식을 구축함으로서 그 유효성을 입증하였다. 현재 수작업으로 진행되고 있는 학회지의 데이타베이스화 작업에 스케너와 컴퓨터를 이용한 본 방법을 적용할 경우 많은 부분이 자동화 될 것으로 기대되며, 따라서 인력과 시간, 경비가 대폭 절감 될 것이다. 향후 과제로는 일반문서로의 확장, 문자인식을 이용한 검증법, ODA 및 SGML에 의한 표준화 문서 등을 연구함으로서 궁극적인 목표인 문서이해지향 시스템이 개발되어야 할 것이다.

〈참고문헌〉

가록현, 김신용, 박세진, 정동석, "CRLCA(Counted Run Length Cutting Algorithm)을 이용한 문서분할에 관한 연구", 「대한전자공학회 추계종합학술대회 논문집」 제15권, 제2호, 1992, pp.492-496.

신현관, "문서의 영역분리와 레이아웃 정보 추출에 관한 연구," 광운대학교 석사학위 논문, 1992.

이인동, 권오석, 김태균,"블록영상의 추출 알고리즘," 「한국정보과학회논문지」, Vol. 18, No.2, 1991, pp.218-226.

Kise, K., M. Yamaoka, N. Babaguchi, and Y. Tezuka, "Organizing Knowledge-Base for Structure Analysis of Document Images," 「情報處理學會論文誌(日本)」, Vol. 34, No. 1, 1993, pp.75-86.

Luo, Q., T. Watanabe, and N. Sugie, "Structure Recognition of Japanese Newspapers Based on Rule-Based Approach," 「電子情報通信學會論文誌(日本)」, Vol. J75-D-II, No.9, 1992, pp.1514-1525.

Yamada, M., "Conversion Method from Document Image to Logically Structured Document Based on ODA," 「電子情報通信學會論文誌(日本)」, Vol. J76-D-II, No.11, 1993, pp. 2274-2284.

Hong, Y., "Skew Correction of Document Images using Interline Cross-Correlation," *CVGIP*, Vol. 55, No. 6, 1993, pp.538-543.

ISO, "ISO-8613-Open Document Architecture / Open Document Interchange Format," 1989

Tsujimoto, S. and H. Asada, "Major Components of a Complete Text Reading System," *IEEE(USA)*, Vol.80, No.7, 1992, pp.1133-1149.

Wang, D., and S.N. Srihari "Classification of Newspaper Image Blocks Using Texture Analysis," *Computer Vision, Graphics, and Image Processing*, Vol.47, No.3, 1989, pp.327-352.