내용분석을 통한 논문지의 목차분류 시스템의 구현

권 영 빈†

요 약

본 논문에서는 논문지 정보를 데이터베이스화하는 시스템의 구축에 있어 논문지 정보를 입력하고 색인을 구성하는 데 드는 노력을 줄이기 위해 목차를 자동으로 생성하기 위한 방법을 제안하고 있다. 기존의 문서분석방법으로는 예외적인 부분이 많은 목차 형식을 효과적으로 분석할 수 없었으며 우리가 원하는 부분을 효과적으로 추출할 수가 없었으므로 본 논문에서는 논문지 목차의 효율적인 인식을 위한 구조적인 분석방법을 제안하고 있다. 논문지 목차에서 표현하고자 하는 가장 필수적인 요소는 논문지에 실린 논문의 제목, 저자, 페이지 등 세 항목이므로 이 세 가지 항목을 추출하기 위하여 모델링하고 특성을 분석하고 있다. 제안한 모델링 방법에 따른 목차 인식 시스템을 구현하여 제목,저자, 페이지 등의 순서를 갖는 논문 목차를 대상으로 660편에 대하여 측정한 결과 91.5%의 논문추출 성공률을 얻었다.

키워드 : 논문지 목차 분류, 구조적 분석, 논문지 목차 모델링, 문서 구조 분석

Implementation of a Journal's Table of Contents Separation System based on Contents Analysis

Young-Bin Kwon†

ABSTRACT

In this paper, a method for automatic indexing of contents to reduce effort for inputting paper information and constructing index is considered. Existing document analysis methods can't analyse various table of contents of journal paper formats efficiently because they have many exceptions. In this paper, various contents formats for journals, which have different features from those for general documents, are analysed and described. The principal elements that we want to represent are titles, authors, and pages for each papers. Thus, the three principal elements are modeled according to the order of their arrangement, and their features are extracted. And a table of content recognition system of journal is implemented, based on the proposed modeling method. The accuracy of exact extraction ratio of 91.5% on title, author, and page type on 660 published papers of various journals is obtained.

Key Words : Journal's Table of Contents Separation, Structural Analysis, Modeling of Journal's Table of Contents, Document Layout Analysis

1. 서 론

오프라인 문자인식 연구는 상용컴퓨터가 보급된 초기부터 40여 년 간 수행되어 현재는 상용화 수준에 도달하였다. 오프라인 문자인식 기술은 낱자로 분리된 문자만을 인식 대상으로 하므로 멀티미디어 구성 요소를 갖는 실제의 여러 가지 인쇄물에 대하여 인식률이 많이 저하되고 있다. 문서구조분석은 이러한 문자인식에 선행하여 여러 종류의 문서로부터 문자, 그래픽, 그림 등을 분리해내고 그 중 문자를 문자인식기로 처리하도록 하는 역할을 수행한다. 즉, 문서구조분석 방법은 임의의 형태를 갖는 문서로부터 문자, 도표, 그래픽 등을 분리하여 문자부는 문자인식기로 인식하여 코드화된 데이터로 저장하도록 하고, 나머지 부분인 그림과 그래픽은 압축하여 저장하는 분야를 의미한다[11][22].

최근 디지털화에 힘입어 데이터베이스 시스템들도 단순히 자료 저장과 관리에 만 그치지 않고 다양한 서비스를 제공하는 방향으로 변화하고 있는데, 예를 들어 논문 DB 시스템만 하더라도 예전처럼 단순히 논문자료들을 저장하는 데에만 그치지 않고 색인 및 검색 등 추가 서비스를 제공하고 있는 형편이다. 특히 ISI의 SCI 등은 전 세계를 대상으로 우수 논문들에 대한 검색 및 열람서비스를 제공하여야 하므로 효율적인 논문 입력 및 검색 시스템이 필수적이다.

본 논문에서는 이러한 흐름에 발맞추어 갖가지 형태의 논문지 정보를 수집하여 입력, 저장, 관리, 분석 표현하기 위해 구축되는 시스템의 입력부분을 담당하는 인식 모듈 중 색인

※ 본 연구는 중앙대학교 2005년도 교비연구비에 의하여 이루어졌습니다.
† 종신회원 : 중앙대학교 컴퓨터공학부 교수
논문접수 : 2007년 1월 31일, 심사완료 : 2007년 8월 7일

및 검색을 위한 논문지의 목차 (table of contents, TOC) 인식에 관한 내용을 연구하기로 한다. 일반적으로 논문 데이터베이스는 입력, 저장, 검색, 그리고 출력의 네 부분으로 이루어져 있는데, 그 중에서 가장 힘들고 시간을 요하는 부분이 입력 부분으로 알려져 있다. 논문 정보의 입력은 사람에 의한 직접 입력 방법과 이미 출판되어 있는 논문지 페이지들을 스캔 입력받아 이미지나 PDF 등의 영상 자료로 저장하는 방법, 그리고 논문지를 스캔한 입력 영상으로부터 자동 인식하는 방법 등이 가능하다. 그런데 사람에 의한 직접 입력 작업은 사람이 일일이 그 논문 자료를 보고 직접 타이핑을 해 넣기 때문에 오차 발생율이 적고 텍스트 자료로써 저장을 할 수 있어 원하는 DB 시스템에 가장 적합한 결과를 얻을 수는 있지만 너무 많은 시간과 노력을 필요로 하며 인건비의 상승에 따른 비용이 많이 드는 단점이 있다. 반면 스캔된 입력을 통해 영상자료로써 저장하는 방법은 원문을 그대로 이미지화하여 저장하게 되므로 원문 형식을 가장 충실하게 보존할 수 있고 입력 및 보관이 편리하다는 장점이 있으나 데이터가 이미지로써 저장되므로 본문의 검색이나 일부분의 발췌 및 수정이 불가능하다. 마지막으로 스캔된 입력을 통한 자동 인식 방법은 시간과 노력을 줄일 수 있으며 데이터가 텍스트로 저장되므로 위의 두 가지 방법의 장점을 두루 갖춘 가장 효율적인 방법으로 판단되고 있다. 그러나 논문지 형식이 너무 다양하므로 일반적인 문서영상구조 분석 방법으로는 일반화되기 어려운 문제점이 존재한다. 그러므로 본 논문에서는 자동 인덱싱을 위하여 논문지의 목차로부터 정보를 자동으로 추출하는 방법을 고려해 보기로 한다. 다양한 논문지 형식을 일반화하기 위해 실제의 논문지의 표지를 수집하고 형식을 분석하여 형태 분류를 수행한 후 형태별 특성을 분석하여 자동인식을 위한 방법을 제안하여 실용성이 있는 논문지 목차 입력 시스템을 구성하도록 한다.

논문의 구성을 살펴보면 다음과 같다. 2장은 문서구조분석에 대한 기존의 연구를 살펴보고, 본 연구에서 해결해야 할 문제들을 정의하도록 한다. 3장에서는 논문지 목차 형식을 분석하기 위한 모델링 과정에 대해서 언급하며, 4장에서는 다양한 논문지 목차를 추출하기 위한 방법을 제안하여 이를 구체적인 시스템으로 구현하기로 한다. 그리고 5장에서는 본 논문에서 제안하는 방법에 대한 실험결과 및 분석을 하고, 마지막으로 6장에서 결론 및 앞으로의 과제에 대하여 설명하기로 한다.

2. 문서 구조 분석에 관한 연구의 분석

2.1 문서 구조 분석

문서구조분석이란 임의 형태의 문서로부터 문자영역, 그림영역, 도표, 그래픽영역 등을 분리해 내는 기법이다. 이러한 문서구조분석의 방법에는 영상 데이터의 작은 요소를 찾아내어 이를 기반으로 문서 전체의 구조를 분석해 나가는 상향식 방법과 문서 전체를 작은 영역으로 분리하는 하향식 방법이 있다[4][22]. 이 두 가지를 병행하여 분석하는 방법이 사용되기도 한다. 상향식 방법의 대표적인 예로 연결요소분석(Connected Component Analysis)방법[12]이 있다. Fletcher와 Kasturi는 제약 조건을 갖는 회로 문서의 분석을 수행하였으며[13], 장명욱등은 연결요소를 이용하여 문서를 문자정보와 그래픽정보로 분류하고 문자분할 오토마타를 이용하여 낱자인식을 수행하였다[7]. 연결요소와 영역확장을 이용하여 문서영상을 분할하는 연구 [1][5][6][8]와 테이블 영역에서의 단어 추출 연구[9], 부분 매칭을 이용한 서식 문서 분류[2] 등이 수행되었다. 하향식 방법의 대표적인 방법으로는 런길이 평활화(Run Length Smoothing) 방법[14]이 있다. Whal은 문자, 선, 그래픽 및 하프톤 영상을 분리하는 방법을 제안하였으며[22] Tsujimoto는 문서의 분리뿐 아니라 접촉문자의 분리까지 수행하는 방법을 제안하고 있다[20]. Hirayama는 런길이 평활화와 연결요소 분석을 혼용하여 문자와 그림 블록을 분리하는 방법을 제안하고 있다[15]. 구문론적인 구문분석[3]을 이용한 방법도 존재한다. 또 다른 하향식 방법으로는 본 논문에서 사용한 프로젝션 방법(Projection Method)[18-19]이 있는데 이 방법은 앞에서 설명한 두 가지 분석방법과 병행하여 사용할 경우 큰 효과를 얻을 수 있으며, 상황에 따라 선택적인 적용이 가능하다는 장점을 갖는다.

런길이 평활화 방법은 구현이 간단하고 처리속도가 빠르지만 복잡한 문서에서 다양한 구조를 세밀하게 분석하는 능력이 떨어진다. 반면 연결요소분석 방법은 구현이 복잡하고 처리속도가 상대적으로 느리지만 세밀하게 구조를 분석할 수 있는 장점을 가지고 있다[22]. 그리고 프로젝션 방법은 가로축 방향이나 세로축 방향에 대해서 프로젝션을 하게 되므로 속도가 빠르며 전체적인 구조 및 부분적인 영역분포나 상대적인 위치 등을 세밀하게 분석하는데 뛰어나다, 그러나 프로젝션 방법은 단독으로 수행되었을 경우에 정확한 라인 정보를 얻기 힘들기 때문에 런길이 평활화 방법이나 연결요소 분석 방법과 함께 사용하는 것이 좋다. 본 연구에서 분석대상으로 하는 문서가 논문지 목차 형식이고 이는 수식을 제외한 인쇄체 한·영 혼용문서이며 일반적인 문서와 달리 특수구조를 가지고 있으므로, 이를 구조적인 특징을 바탕으로 세밀하게 분석하기 위하여 프로젝션 방법을 주로 사용하고 부분적으로 런길이 평활화와 연결요소 방법을 혼용하는 것을 고려하고 있다. 또한 본 논문에서는 대부분이 논문의 표지에 나타나는 목차의 구조 분석을 대상으로 하고 있으므로 표지는 논문지의 맨 앞 부분에 해당하므로 스캔을 수행하는 경우 스캐너의 상단에 위치하도록 하므로 상대적으로 기울어짐이 적게 발생하고 있다. 그리고 스캐너를 사용할 경우 해상도가 과거에 비하여 개선이 되었으므로 잡영이 발생하는 경우가 현저하게 감소되었다. 그러므로 본 논문에서는 스캔시의 기울어짐이나 잡영에 대한 문제는 없는 것으로 가정을 하고 간단한 필터링만 수행하는 것으로 가정하고 있다.

문서 구조 분석 중에서 논문지 목차의 분석(TOC)에 관한 연구가 형태 분석의 특수성 때문에 다수 이루어지고 있다.

Belaid 등은 목차에 대한 선행 지식 모델을 사용하여 자동 인식을 수행하고 라벨을 생성하는 방법[10]을 제안하였으며, Mandal은 137개의 목차를 구분하기 위하여 목차에 대한 선행 지식을 통해 구조적인 정보를 추출하는 방식을 제안[17]하고 있다. Tsuroka 등은 구조 분석을 통하여 서적의 목차를 XML 문서로 변환하는 방식을 제안[21]하고 있으며 Lin은 텍스트와 페이지 번호의 내용 결합에 의한 목차 분석 방식을 연구[16]하였다. 이러한 방법은 모두 OCR과 연계하여 결과의 정확도를 다시 검증하고 있다. Belaid와 Lin의 연구에서 논문의 목차로부터 목차 정보를 추출하는 아이디어를 얻을 수 있다. 이들은 OCR로부터 페이지 번호를 추출한 후 필요 정보를 추출하는 방식을 채택하고 있는데 비하여 본 논문에서는 OCR에 의존하지 않는 방법을 제안하여 제시된 방법의 결과와 비교해 보기로 한다.

문서구조분석 뒤에는 흔히 문서구조 이해가 뒤따르기도 한다. 문서구조분석은 문서를 영상데이터로 입력받아 그 형태를 분석하는 것임에 반하여, 문서구조이해는 문서구조분석의 결과로 얻어진 문서의 물리적 구조로부터 사람이 인지할 수 있는 논리적 구조로 도출하는 것을 의미한다. 이러한 문서구조이해는 문서에 대한 사전지식이 없이는 수행될 수 없으므로 문서형태 정의 언어(Form Definition Language)로 미리 정의된 정보가 있어야 하며, 문서구조이해 과정을 거친 문서영상은 논리적인 구조에 의해서 체계적으로 데이터베이스로 변환될 수 있다[3][11][22].

2.2 문서 구조 분석의 문제점

문서 인식이나 문자 인식에 있어서 고려하는 방법은 크게 다르지 않다. 가장 커다란 차이를 보이는 부분은 문서상에서의 영역분할이다. 문서를 인식하는데 있어서는 문서의 구조를 분석하여 그 영역을 각기 다른 속성을 갖는 블록들로 분할해야 하는 것이 문자인식과는 근본적으로 다른 점이다. 문자인식이 문자열로 판별된 영역에 대해서 낱자분리나 자소분할 및 합병을 하는 방법과 문서인식이 각 영역후보에 대해서 분할 및 합병을 하는 방법은 유사하다. 그러나 문서인식에 있어서는 영역의 범위를 정확히 판별하는 것 이외에 영역의 특성을 파악하는 것 또한 중요하다. 문자인식에 앞서 문자열이나 그림, 도표 등으로 문서영역이 구분이 되어야 각기 특성에 맞게 정확한 인식이 가능한 것이다. 하지만 문서는 그 형식상의 다양함으로 인해 모든 부분에 대한 인식이 불가능하며 주로 나타나는 표나, 그림에 대해서만 판별이 가능한 수준이다. 그리고 문서인식에 있어 궁극적으로 필요한 것은 영역들의 의미를 판별하는 것이다. 현재 문자인식 모듈들은 일반적인 단어와 사람이름을 구분하는 것이 불가능하다. 단어사전 정보를 바탕으로 조정하여 인식률을 향상시키는 시스템이 존재하지만 그 효율이 그다지 높지 못한 편이고 이 경우에도 사람의 이름에 대해서는 여전히 오인식 문자열로 처리하게 된다.

지금까지 살펴본 바와 같이 문서 구조 분석은 그 형태상의 특성 때문에 주로 비교적 보편적인 분야에 대해서 연구가 진행되고 있는 실정이다. 앞서 살펴본 대부분의 연구는 주로 문서상에 존재하는 그림이나 표, 문자영역에 대한 구분과 그 후의 문자인식에 관한 연구가 주된 내용이었다. 그러나 현재까지의 문서인식 방법으로는 문서상에 나타나는 영역들의 각각의 의미에 대한 판단이 불가능하며, 보편적인 문서의 구조를 많이 벗어날 경우 그 인식률에 한계가 있었다. 따라서 본 논문에서는 그 인식의 대상 분야를 논문지의 목차로 한정시켜 분석을 행함으로써 문자인식에 선행하여 구조분석만으로 각 영역요소에 대한 의미를 판단할 수 있게 하고 또한 이를 바탕으로 문자 인식시에 각 문자열 후보들이 의미를 갖도록 추출하는 방법을 제안하였다. 이를 통해 논문 데이터베이스에서의 원문 입력시 인식기가 논문정보를 자동으로 지각하여 입력함으로써 데이터베이스 상에서의 논문 색인 및 검색을 가능하게 하는 것이 목표가 되고 있다.

3. 논문지 목차 형식의 모델링

3.1 논문지 목차 형식 모델링의 필요성

논문지는 각 학회별로 우수한 논문을 게재하여 일정기간마다 발행하는 것이 보통이다. 현재 학문적인 연구가 진행되는 모든 분야에 학회나 단체가 존재하며 각 분야마다 제각기 특성이 다른 단체들이 무수히 많이 존재하고 있다. 여기에 시간이 흐를수록 더욱 새로운 학문분야가 탄생하고 또 그에 따라 신생 학회도 늘어날 것이기 때문에 각 학회나 단체에서 발행하는 논문지는 계속 증가하고 있다. 그런데 이러한 논문지는 정형화된 형식이 없기 때문에 관습적으로 사람들이 받아들이고 있는 정도의 형식만 지켜줄 경우 각기 조금씩 다른 형태를 지니게 된다. 논문지는 발행하는 학회에 따라 다양한 형식을 지니고 있으며 목차도 또한 마찬가지이다. 그러나 공통적인 것은 논문지에 실린 논문의 제목과 그 논문을 저술한 저자, 그리고 논문지 상에서 그 논문이 실린 페이지 정보 등이다. 논문지는 대부분이 인쇄매체로 발행되어 보급되고 보관하는 것이 보편적이다. 그러나 누적된 논문의 양이 방대해지고, 보관상의 어려움을 극복하고자 전자문서화 방안이 제시되고 있다. 논문의 보관 및 검색 또한 예전의 도서관에서 보관하던 인쇄매체를 통하던 수준에서 벗어나 데이터베이스 시스템을 통한 논문 서비스가 급속히 확산되고 있는 편이다. 얼마 전만 해도 데이터베이스를 통한 검색에서도 초록만을 서비스해주는 수준에 그쳤으나 현재는 원문 서비스도 보편화되고 있다. 이로 인해 과거의 논문지 정보나 또는 현재, 그리고 이후의 논문들을 전자문서로 변환하여 데이터베이스에 입력하고 관리하는 것이 중요한 비중을 차지하게 되었다. 그러나 그 작업의 특성상 입력에 상당히 많은 인력과 시간을 필요로 하는 것이 문제점이다. 세계적으로 권위있는 Index의 경우 전 세계를 대상으로 우수 논문들에 대한 검색 및 열람서비스를 제공하여야 하므로 효율적인 논문 입력 및 검색 시스템이 필수적이라 생각된다. 그러나 데이터베이스 상에서의 검색을 위해서는 논문들에 대한 색인과 초록 등의 정보가 필요한데, 기존의

인쇄문서나 전자문서화 되지 않은 문서의 경우, 그리고 전자문서화 되었으나 이를 디지털화하여 저장하지 않은 경우에는 사람이 지각을 하여 수동적으로 추출을 해내야 하므로 데이터베이스를 구축하는데 있어 2중의 수고를 피할 수가 없게 된다. 이에 따라 논문 영상을 자동으로 인식하는 시스템들은 많이 연구가 진행되었으나 아직 목차에 대한 연구는 국내에서 많이 이루어지지 않고 있으므로 이 부분에 대한 연구가 본 논문에서 고려하는 주된 목표가 되고 있다.

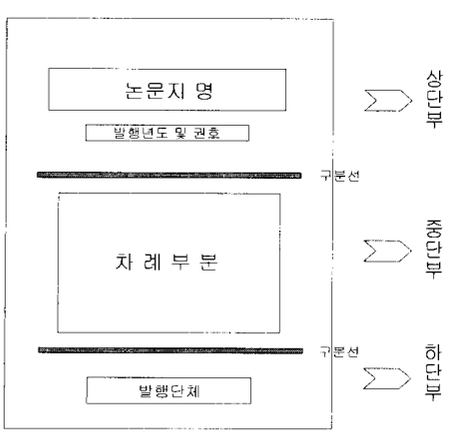
논문지 목차에 대한 인식을 할 경우에 우선 논문지 목차 형식에 대한 선행지식이 필요하다. 대부분의 인식모듈은 선행적인 구조지식에 많이 의존하며 이 과정이 인식결과에 있어 많은 영향을 주게 되므로 보다 정확하고 효율적인 인식을 위해 논문지 목차 형식의 분석 정보를 통한 모델링이 필요하게 된다.

3.2 논문지 목차형식의 분석

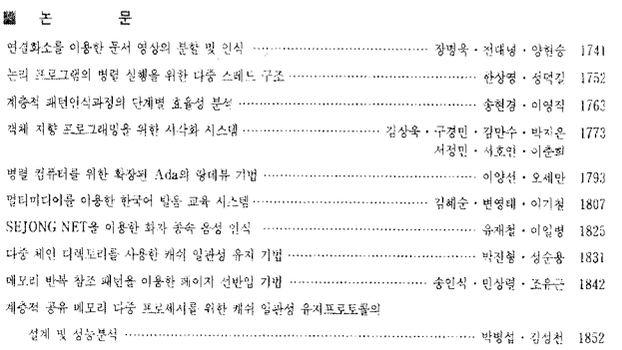
3.2.1 논문지 목차 형식의 배치구조 분석

논문지 목차를 분석하는데 있어서 가장 중요한 것은 논문지 목차가 갖고있는 정보이다. 분야나 발행단체에 따라 논문지 목차에는 그림이나 로고, 발행기관 홍보문구 등 수없이 다양한 내용이 분포되어 있다. 그러나 공통적으로 논문지 목차가 나타내려고 하는 것은 논문지에 실린 논문 제목과 저자 및 페이지 등과 같은 논문에 대한 색인 및 차례정보이다. 본 연구는 목차형식의 특징을 구조적으로 분석하여 필요한 정보를 효율적으로 추출하려고 한다. 이를 위하여 논문지 목차 형식들을 분석하기로 한다. 현재 대부분의 논문지는 편집을 통하여 파일로 보관되고 있으므로 본 논문에서 처리하고자 하는 논문지는 파일로 보관되고 있지 않은 과거의 논문지가 대상이 되고 있다.

예를 들어, 정보처리학회 논문지는 비교적 간단한 목차 형식에 속하며 구조 분석에 적합한 형태를 갖고 있는 편이다. 특별한 배경 그림 및 로고가 없으며 목차에 속하지 않는 구역은 상단과 하단으로 나뉘어져 있어 원하는 목차 영역을 취득하기에 좋은 특성을 지니고 있다. 상단부에는 논문지 제목 및 발행 부수, 년도 등의 내용이 인쇄되어 있고 구분선 다음에 논문지에 실린 논문들의 차례가 이어지고 있



(그림 3-1) 논문지 목차의 기본 구조


■ 논 문

연결화소를 이용한 문서 영상의 분할 및 인식 ····· 장병욱 · 전대성 · 양현승 1741
논리 프로그램의 병렬 실행을 위한 다중 스페느 구조 ····· 한상영 · 성덕길 1752
계층적 패턴인식과정의 단계별 효율성 분석 ····· 송현경 · 이영직 1763
객체 지향 로프그래밍을 위한 시각화 시스템 ····· 김상욱 · 구경민 · 김만수 · 박지은 1773
서정민 · 서호원 · 이춘희
병렬 컴퓨터를 위한 확장된 Ada의 랑데뷰 기법 ····· 이양선 · 오세만 1793
멀티미디어를 이용한 한국어 탈을 교육 시스템 ····· 김희순 · 변영태 · 이기철 1807
SEJONG NET을 이용한 화학 종속 음성 인식 ····· 유재철 · 이일병 1825
다중 체인 디렉토리를 사용한 캐쉬 일관성 유지 기법 ····· 박진철 · 성순용 1831
매모리 반복 참조 패턴을 이용한 페이지 선반입 기법 ····· 송인식 · 민상렬 · 조유근 1842
재승적 공유 메모리 다중 프로세서를 위한 캐쉬 일관성 유지프로토콜의
설계 및 성능분석 ····· 박병섭 · 김성천 1852

(그림 3-2) 차례 부분의 구조

다. 그리고 마지막으로 구분선으로 마무리를 하고 아래에 발행 단체의 로고와 이름이 인쇄되어 있다. 이런 스타일의 경우 전체적인 구조 분석에는 상당히 유리하다. 실질적인 목차 영역을 찾아내는 데에 특징적인 구분자가 있어 구조 분석에서 도움을 얻을 수 있다. 이를 그림으로 나타내면 (그림 3-1)과 같다.

위 그림은 차례 부분을 일부 발췌한 것이다. 이를 살펴보면 먼저 '■ 논 문'이라는 특정 지시어가 있고 그 뒤로 논문들의 차례가 열거되어 있다. 또한지시어가 있는 자리에 세부 분야를 명시하는 경우가 많다. '■ 논 문'이라는 위치에 존재하는 지시어들을 포괄적으로 'Sub Area' 구분자로써 취급하기로 한다.

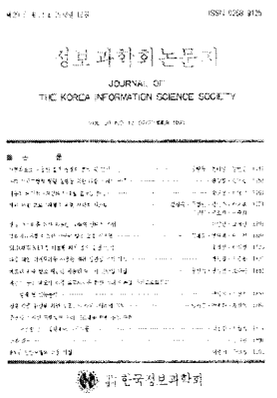
이제 논문의 정보를 담고 있는 문자열의 내용을 보면 먼저 제목 그리고 점선, 저자, 게재 페이지 순으로 되어 있으나 현재 상용화되어 있는 소프트웨어로 분석을 해보아도 분석이 되지 않는다. 일반적인 구조 분석법에서는 문자들이 모여 있는 문자열 또는 문단 영역, 그림이나 그래픽 영역, 도표나 차트영역, 이렇게 크게 세 가지로 분류를 하게 되고 이 중 문자열이나 문단이라고 인식된 영역에 대해서 다시 오프라인 문자인식을 행하게 된다. 그러나 이를 기준으로 위의 차례 형식을 분석하려 할 경우에는 우리가 의도한 것과는 전혀 다른 방향으로 분석이 이루어지게 된다. 전체적으로 처음 논문지 제목과 같은 것만 보더라도 일반적인 문서인식기는 이것을 텍스트로 간주하지 않게 된다. 그 이유는 이 영역이 일반적인 문자와 비교해 볼 때 비정상적으로 크고 또 상대적으로 현재 페이지가 포함하고 있는 문자열들과 비교해 봐도 균형이 맞지 않기 때문이다. 실제 차례 영역으로 들어가 보면 문자, 점열, 다시 문자, 숫자 등이 혼합되어 인식이 거의 불가능한 구조를 지니고 있다. 그 이유를 들면 우선 점열의 존재이다. 일반적인 문서에서는 위의 경우와 같이 비상식적으로 점이 반복되는 경우가 거의 없다. 문자열 상에서 ',' 나 '.', '·'가 나타나긴 하지만 이것은 문맥상의 구분이나 문장의 구분, 또는 단어의 구분을 위해서 독립적으로 나타날 뿐이지 여러 개가 연속적으로 나타나지는 않는다. 인식기는 점열을 당연하게 그래픽 영역으로 처리하게 된다. 추가적으로 5행과 11행을 보면 한 논문지에 대한 정보가 두 줄로 이루어져 있다. 논문제목이 길거나 저자가 여러 명일 경우 흔히 일어날 수 있는 상황으로 독립적

〈표 3-1〉 항목의 배열순서에 따른 모델 구분

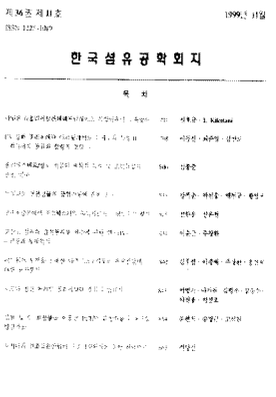
Type	항목의 배열 순서
I	논문제목(Title) - 논문저자(Author) - 게재페이지(Page)
II	논문제목(Title) - 게재페이지(Page) - 논문저자(Author)
III	게재페이지(Page) - 논문제목(Title) - 논문저자(Author)
IV	게재페이지(Page) - 논문저자(Author) - 논문제목(Title)
V	논문저자(Author) - 논문제목(Title) - 게재페이지(Page)
VI	논문저자(Author) - 게재페이지(Page) - 논문제목(Title)

인 문단이 아니며 단지 영역 분류상 두 줄로 나뉜 것이다.

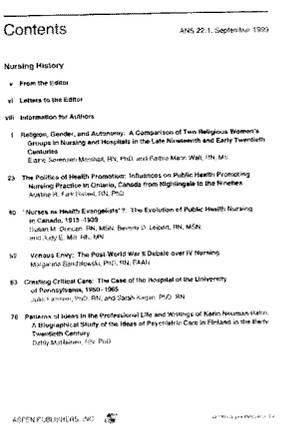
논문지 목차에서 우리가 필요로 하는 정보는 주로 논문제목, 논문저자, 게재페이지 등이므로 이를 중심으로 형태를 분류하게 된다. 우선 제목, 저자, 페이지의 세 가지 항목을 배열 상태에 따라 크게 구분해 보면 순서에 따라 6가지의 경우가 존재하게 된다. 이 세 항목을 기준으로 형태를 분류해 볼 경우 대부분의 논문지 목차 형식이 이를 크게 벗어나지 않는다. 실제 논문지 목차 형식에서 차례 부분에 나타나는 내용들은 기본적인 세가지 항목 이외에도 세부 분야 지시어, 특수기호 및 지시어 등이 있으나 이는 규칙적인 패턴이 없고 또한 항상 존재하거나 꼭 필요한 항목들이 아니다. 그래서 나머지 항목들은 형태를 분류하여 모델을 선정하는데 있어 기준자료로써의 활용가치가 적다.



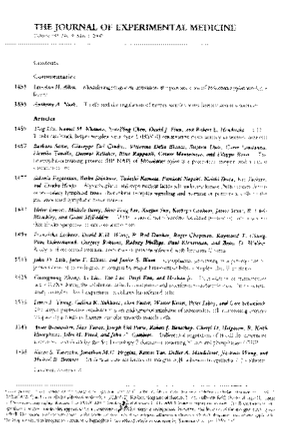
① Type I : T-A-P 형태



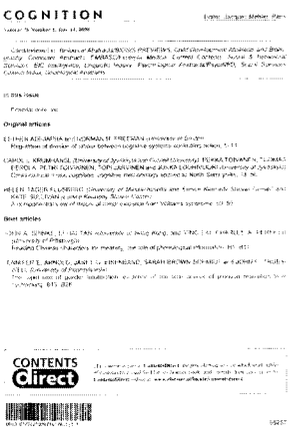
② Type II : T-P-A 형태



③ Type III : P-T-A 형태



④ Type IV : P-A-T 형태



⑤ TypeV : A-T-P 형태

⑥ Type VI : A-P-T 형태

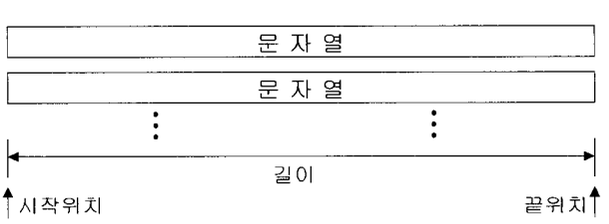
(그림 3-3) 6가지 형태에 대한 목차 예

3.2.2 논문지 목차형식의 구조적인 특성 분석

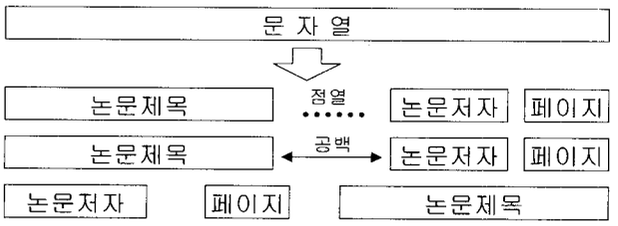
지금까지 살펴본 정보를 바탕으로 논문지 목차 형식에 대한 분석결과를 그림으로 나타내면 앞의 (그림 3-1)과 같다. (그림 3-2)에 나타난 차례 부분의 Sub Area 항목을 분석하여 그림으로 나타내면 (그림 3-4)와 같다. 대개 시작 위치는 차례부분에 나타나는 다른 문자열들과 같으며 폰트 크기도 다르지 않다 다만 특징적인 것은 지시어의 성격이 강하기 때문에 문자열의 길이가 다른 것에 비해 짧다는 것이다. 그리고 논문 정보를 나타내는 문자열의 경우는 시작위치와 마찬가지로 끝나는 위치도 일정하다. 그러므로 문자열의 길이가 전체 문자열의 평균 길이에 비해 2/3 이하로 짧고, 끝나는 위치도 전체 영상에서의 너비의 2/3 이하이면 Sub Area 항목으로 볼 수 있다.

다음으로 실제 논문의 정보가 실려 있는 차례를 보면 이 영역에서 나타나는 문자열은 일정한 패턴을 지니고 있다. 한 문자열 내에서 표현하고자 하는 정보를 모두 충족시킬 경우 각각의 문자열마다 시작위치가 일정하며, 종료 위치도 일정하다. 그러므로 시작과 종료 사이 문자열의 길이를 측정하면 일정하게 나오게 된다. 이러한 성질은 차례 형식이

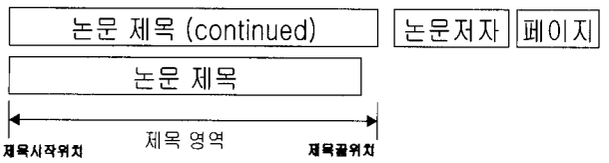
(그림 3-4) Sub Area 영역이 갖는 특징



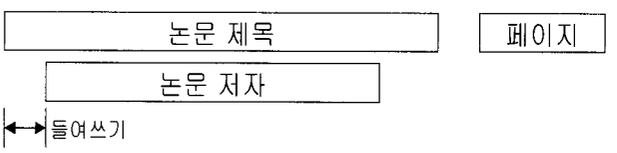
(그림 3-5) 차례부분에 나타나는 문자열의 특징



(그림 3-6) 차례에서 논문정보를 담고 있는 문자열의 특징



(그림 3-7) 멀티라인에서의 제목의 이어짐



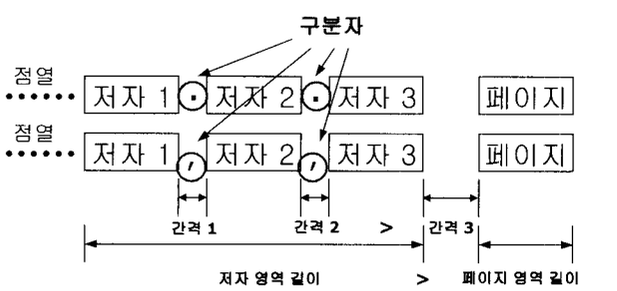
(그림 3-8) 멀티라인에서의 저자영역의 판별

가지는 정형성으로서 이 조건에서 벗어날 경우에는 Sub Area 지시어일 때와 논문정보의 내용이 길어 한 줄에 다 표현하지 못하는 경우를 예상하면 된다.

개별적인 논문의 정보를 싣고 있는 각각의 문자열은 세부적인 항목들로 다시 구분이 될 수 있다. 게재된 논문의 제목, 논문의 저자, 그리고 논문이 게재된 페이지 등이 바로 그것이다. 이들 세부요소는 우리가 목차 인식을 통해 논문에 대한 요약정보를 추출하고 자동 색인이 가능하게 하려는 대상이 되는 항목들이다.

(그림 3-7)과 (그림 3-8)은 한 논문의 정보가 한 줄에 다 표현되지 못하고 여러 줄에 나뉘어 표현될 때의 그 구분을 위한 특징을 나타내고 있으며, (그림 3-9)는 저자영역과 페이지영역이 인접해 있을 경우 이의 구분을 위한 특징들을 나타내주고 있다. 다음절에서는 세 항목에 대한 이러한 특성들을 바탕으로 모델을 세워 일반화시킴으로써 인식에 도움이 되고자 한다.

이제까지 살펴본 바에 따르면 논문지 목차와 같은 특수성을 띠는 문서에 대해서 일반적인 문서 구조 분석 방법은 통



(그림 3-9) 저자영역과 페이지영역의 구별

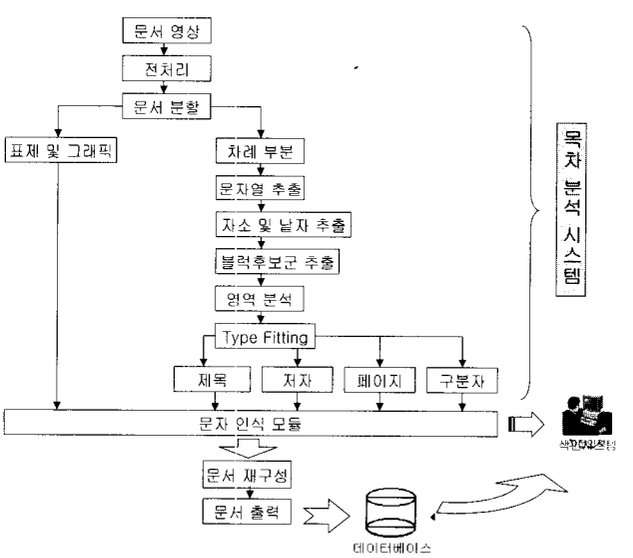
용되지 않는다. 이에 본 연구에서는 논문지 목차 형식을 분석하고 대략적인 특징을 추출하여 이를 형태별로 일반화함으로써 인식기가 목차 형태를 판단하고 자동으로 차례 내용을 인식하는데 있어 도움이 되고자 하며 분석 정보를 바탕으로 논문지 목차 자동 인식기를 구현하여 이를 검증하고자 한다.

4. 분석에 근거한 목차 인식 시스템의 구현

4.1 제안하는 시스템의 구성

본 논문에서는 앞장에서 분석한 논문지의 목차구조정보를 바탕으로 논문지 목차 형식의 항목들에 대해 인식시스템이 각각의 항목들이 갖는 의미를 자동으로 인식하여 논문 데이터베이스 시스템에서 논문정보에 대한 요약과 검색 및 색인 추출을 용이하게 하는 시스템을 제안하여 기존의 문서구조 분석 방법에서 처리하지 못하는 TOC의 처리가 가능하도록 하는데 목표가 있다.

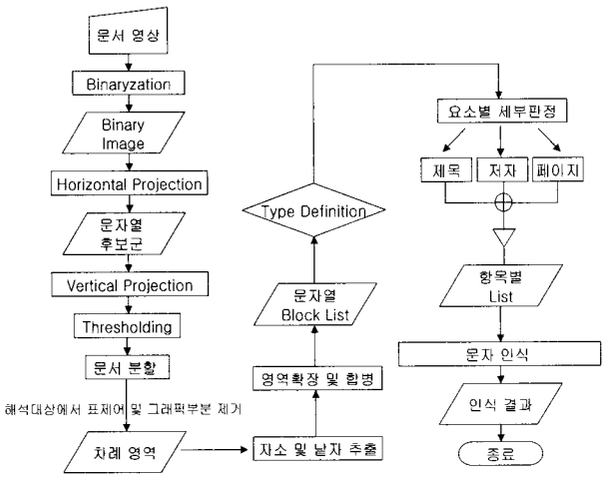
제안하는 시스템의 전체적인 구성은 (그림 4-1)과 같다. 입력된 문서영상은 먼저 표제 및 그래픽 부분과 차례 부분으로 분리되어 따로 인식이 되게 되며, 차례 부분은 다시 세밀한 분석절차를 거쳐서 문자열에서 각각의 항목들을 추출해내게 된다. 이때 앞 장에서 살펴본 논문지 목차의 구조분석 정보가 기준 자료가 되며, 분류했던 형태를 적용시켜 최종적으로 인식을 하게 된다. 이 때 각각의 항목들은 자신들의 항목에 해당하는 의미를 지니게 되므로 문자인식 모듈을 거쳐 색인 시스템에 자동적으로 추출될 수 있도록 한다.



(그림 4-1) 목차 인식 시스템의 구성

4.2.1 목차 분석 시스템

입력 자료로는 8bit gray scale 형식으로 된 300dpi의 해상도 이미지를 사용하여 처리하게 된다. 문서 이미지를 입력으로 받아들이면 목차인식기는 첫 단계로 이 문서에 대해 전처리를 하여 스캔하면서 일어난 이미지의 손실을 보완하



(그림 4-2) 목차 인식 과정

고 이진화를 수행하게 된다. (그림 4-1)의 목차분석 시스템에서 목차인식부분에 대해 구체적으로 살펴보면 (그림 4-2)와 같은 순서로 그 분석과정이 진행되게 된다.

첫 단계에서 잡영이 제거되고 이진화 된 이미지 파일은 두 번째로 가로 방향에 대하여 즉, Y축을 기준으로 수평 프로젝션(horizontal projection)을 행하게 된다. 이렇게 가로 방향에 대하여 프로젝션을 하면 대략적으로 라인 분포를 알아 낼 수 있다. 다음으로 검출된 라인에 대해 2차로 X축에 대해서 세로방향으로 수직 프로젝션(vertical projection)을 하면 각 라인별로 자소 및 낱자의 분포까지도 알 수 있게 된다. 그 다음 문서분할을 하여 차례 영역만을 인식 대상으로 정하고 세부 분석을 하여 2차 프로젝션 단계에서 구분했던 결과를 바탕으로 자소 밑 낱자에 대한 외접 사각형을 연결 요소로 구성한다. 그리고 영역확장 및 합병을 통해 이들 낱자영역을 다시 어절단위의 영역으로 확장한다. 이렇게 생성된 그룹들에 대해 앞장에서 분석한 차례 형식의 특징을 적용하면 각각의 항목별 후보 블록들을 추정할 수 있게 된다. 여기에 목차 형태별 모델 형태를 추정하여 후보 블록들에 대해 세부요소별 판정이 가능해진다. 제목, 저자, 페이지로 판별된 각각의 항목들에 대해서 리스트를 작성하여 이를 문자인식 모듈을 통해 인식할 수 있게 하고 인식된 결과를 논문 정보 색인에 이용할 수 있게 한다.

4.2.2 라인 후보 추출

두 번째 단계에서는 문서 영상에서 구조분석의 대상이 되는 후보군들을 추출하는 단계이다. 전처리된 영상에 수평 프로젝션(Horizontal Projection)을 행하여 가로방향에 대한 흑점들의 누적분포도를 구하게 된다. 줄과 줄 사이의 간격이 일정하게 유지되는 목차 형식의 경우 수평방향의 프로젝션 방법만으로 손쉽게 행 추출까지도 가능하다. 논문지 형식에 있어 구조분석에 필요한 윤곽정보를 알 수 있게 되는 것이다. 프로젝션 후에 결과값을 분석하여 상대적으로 일정 임계값을 넘는 부분은 문자열이 있는 라인으로써 판별하고 나머지 부분은 줄 간 여백으로 간주하게 된다. 이 때 기울

(그림 4-3) 가로 + 세로 구분선을 그었을 때

어짐(skew)이 심하지 않아야 하는데, 프로젝션 방법의 특성상 대상 영상의 기울어짐에 따른 결과값의 변화가 민감하기 때문이다. 일반적인 문서 구조 분석에 있어서 연결요소 추출방법을 중심으로 구조를 분석해가게 되는데, 본 연구에서는 인식대상이 논문지 목차로 한정되어 있으므로 연결요소 추출법보다 프로젝션 방법이 오히려 문서에 대한 1차적인 정보를 추출하는데 있어 더 효율적인 방법이 되는 것이다. 본 연구에서 인식대상으로 삼는 논문지 목차형식의 경우에 대상 문서영상이 줄과 줄, 특히 의미를 갖는 블록 간에 반드시 일정 간격 이상의 공백이나 구분자가 존재한다는 가정에 기반을 두고 있으며, 이러한 형식상의 특징 때문에 일반적으로 구조분석에 쓰이는 연결요소 추출방법이나 런길이 평활화 방법보다 프로젝션 방법이 더 효율적인 방법으로 채택되었다.

수평 프로젝션을 바탕으로 임계값을 적용하여 후보군이라고 추정되는 영역에 대한 경계값들을 저장해 놓는다. 임계값을 적용하는 이유는 잡영을 제거하기 위함이다. 가로 방향에 대해서 매우 미세한 픽셀의 분포는 의미가 없으며 이는 잡영으로 판단하여 제거할 수 있다. 가로 방향의 프로젝션에서 얻어진 라인 영역에 대해서 각기 개별적으로 세로방향의 프로젝션을 하게 된다. 즉, 라인 영역이라고 판별된 곳에 대해서만 2차 분석을 위해 다시 가로축, 즉 X축을 기준으로 개별적인 수직 프로젝션(vertical projection)을 행하게 된다. 픽셀 분포에 대해 연산을 하면 모든 독립적인 요소에 대해 낱자별 구분이 가능해지게 된다. 2차 프로젝션 결과를 바탕으로 세로로 된 구분선은 (그림 4-3)과 같으며 이를 보면 각 요소 별로 영역이 판별됨을 알 수가 있다.

4.2.3 영역해석 및 문서분할

구분된 라인들에 대해서는 우리가 원하는 관심영역(Region of Interest, ROI)인 차례 부분을 판별하기 위해 첫 번째로 주어진 영상 중 그림 및 표제 부분을 해석대상에서 제거하는 단계를 거치게 된다. 1차 프로젝션에서 구해진 각 라인들에 대해 평균 높이를 구한다. 그리고 각 라인에서의 2차 프로젝션에서 구해진 연결요소들에 대해서 평균 너비와 평균 간격, 문서상에서의 위치를 구한다. 마지막으로 1차 프로젝션에서의 누적분포도 정보와 연관하여 차례 부분이 가지는 특성에 어긋날 경우 이를 표제어 및 로고 등의 그래픽 부분으로 간주하여 이를 해석대상에서 제거한다. 남은 영역에서 일정한 패턴이 나오는 영역을 차례 부분으로 추정하고 이 영역을 차례 영역으로 추출한다. 실험 결과 대개의 경우 양호한 분석률을 보였지만 특수 기호나 예외 형식이 나타날 경우 정확히 판별하지 못하는 것도 있었다. 구현한 시스템에서는 정확한 입력을 위하여 사용자가 수동으로 ROI 영역

을 지정하는 기능을 추가하였으며 잘못 판별하였을 경우 이와 같은 피드백(feedback) 기능을 통해 오류를 정정할 수 있도록 하였다.

4.2.4 영역확장 및 요소분석

낱자 정보만으로는 원하는 정보를 찾아내기 어려우므로 이들 낱자 영역을 연계된 후보끼리 합병 및 확장하여 문자열 후보로 도출할 수 있도록 한다. 이 때에는 낱자 영역간 거리, 즉 간격정보를 기준으로 하여 어절단위로 확장해 나가게 된다. 각 문자열별로 측정된 낱자별 평균 간격을 계산하여 이 거리보다 작을 경우 단어간 간격보다는 자소간 간격 내지는 낱자간 간격으로 판단하고 서로의 영역을 합병하게 되는 것이다. 이 과정에서 앞부분이나 뒷부분에 상당히 동떨어져 존재하는 작은 영역이 있을 경우 이를 제거하게 되므로 또 한번의 잡영 제거 효과를 볼 수 있다. 이 단계를 거치게 되면 어절 단위로 구분이 가능하게 된다. 어절 단위의 외접 사각형을 구하는데 있어 연결요소 추출이나 런길이 평활화 방법을 이용하지 않고 두 번의 프로젝션 만으로 찾아냄으로써 정확도를 떨어뜨리지 않고 시간적으로 많은 단축을 꾀할 수 있다는 것이 큰 장점이다. 실제 동일한 윤곽정보를 얻기 위해 시험을 해본 결과 연결요소 추출법에 비해 약 1/20, 그리고 런길이 평활화 방법에 비해서는 약 1/6 정도의 시간이 소요되었다.

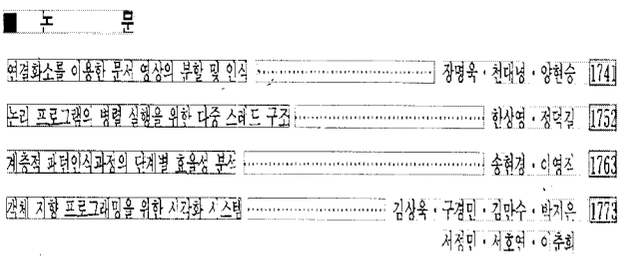
어절 단위로 판별된 그룹에 대해서 이제 다음과 같은 조건을 적용하여 재확장 및 병합을 실시하며 이렇게 제2의 블록 후보군으로 판별된 그룹들을 Component로서 연결요소화하여 리스트로 저장해 놓는다. 블럭 단위 구분이 이루어지면 원하는 의미 있는 블록의 형태로 구분하기 위해 패턴에 대한 분석이 되어야 한다. 차례 형식에서 찾아볼 수 있는 첫째 특징은 바로 문자열의 시작위치이다. 한 줄에 다 표현이 되어 있다고 가정할 경우 시작위치는 항상 일정하다. 마찬가지로 끝나는 위치도 일정하다. 이를 가지고 한 논문에 대한 글이라는 것을 추측할 수 있으며 들여쓰기 등으로 어긋나는 경우에는 문자열이 위나 아래의 블록에 연관된 것이라는 것을 추정할 수 있다.

Sub Area인지 아닌지를 판단할 때는 폰트의 차이나 위치정보를 우선적으로 분석하여 판단하게 된다. 대개 Sub Area일 경우에는 이탤릭체나 고딕체를 사용하고 논문제목과 구별하기 위하여 폰트크기도 약간 크게 사용하게 된다. 그리고 간결한 지시어이므로 그 위치가 전반부에 걸쳐있는 게 보편적이다. 이 사실을 근거로 Sub Area 영역을 먼저 추출해 낸다. 본 연구에서 제안한 프로젝션 방법으로는 글씨체의 종류를 판별하기가 불가능하다. 굵은 글씨나 이탤릭체에 대한 정보를 픽셀분포만으로 판단해 내기는 어렵기 때문이다. 그러므로 폰트크기와 위치 정보를 활용하게 된다. 차례 영역으로 판별된 부분에서 각 문자열들이 일정한 높이를 가지는 것을 계산적으로 알아낼 수 있다. 같은 레벨을 가지는 논문 내용에 관해서는 일정 폰트로 인쇄하기 때문이다. 이를 벗어나는 높이를 가지는 문자열 영역이 있을 경우 1차적으로 후보로써 추정하고 이 후보들에 대해서 다시 가로축 프로젝션 정보를 가지고 위치를 판별하게 된다. 높이도 상대적으로 크고 위치도 상대적으로 전반부에 있을 경우 이를 Sub Area로 판단한다.

점열이나 공백을 기준으로 양쪽으로 블록을 나눈 후에 다시 앞쪽과 뒤쪽 블록을 세분화 시킬 수 있는 여부를 판단하게 된다. 형태에 따라 앞쪽은 제목, 뒤쪽 블록이 저자와 페이지로 나뉠 수도 있고, 아니면 앞쪽이 제목과 페이지 뒤쪽이 저자로 나뉠 수 도 있다. 이러한 분류는 기본적으로 세 가지 항목이 항상 나타난다는 가정에서 이루어지는 것이다. 양쪽 블록으로 나누면 먼저 분류했던 논문지 목차 형식의 형태에 따라 추정을 하게된다. T-A-P 형태의 경우 앞에 T-A, 뒤에 P 또는 앞에 T, 뒤에 A-P가 될 것이다. 이러한 전제를 가정하고 각 블록 상에서 구분이 가능한 영역이 존재하는지를 판단하게 된다. 일반적으로 항목 사이에는 단어 사이의 띄어쓰기와 구분할 수 있을 정도의 공백이나 구분자가 들어가게 되고 저자 영역이 페이지 영역에 비해 블록사이즈가 크기 때문에 이를 분석하게 되면 각각을 세부적으로 다시 분류할 수 있게 된다. 만일 두 줄에 나뉘어 겹치는 경우가 발생할 경우 일정한 시작위치를 찾아 들여쓰기가 된 행이 시작위치에 비해 위쪽인지 아래쪽인지에 대해서 판단을 하게 된다. 행이 나뉘더라도 대개 문자열의 위치는 그 문자열이 속하는 블록에 가까운 쪽에 위치하기 마련이므로 위치 정보를 파악하여 아래나 위의 라인에 합병하게 된다.

4.2.5 형태의 검증

영역 분석 단계까지 거치게 되면 목차 형식에 대한 사전 레이아웃이 나오게 된다. 분석된 블록후보군들의 특징과 사전 레이아웃 정보를 바탕으로 목차의 형태를 결정하게 된다. 모델이 결정되면 모델별 형태 특성을 적용하여 요소별 세부 판정을 하게 되며, 이 단계에서는 모델에 따른 항목들의 배열 순서를 알게 되므로 앞에서 해석 못한 블록들에 대해서도 추가적인 보정이 가능해진다. 각 항목에 대한 판정이 이루어지면 최종 레이아웃이 생성되게 된다. 문자열의 내용에 대해 의미가 정해졌으므로 각각의 항목별 성분으로 리스트화 한다.



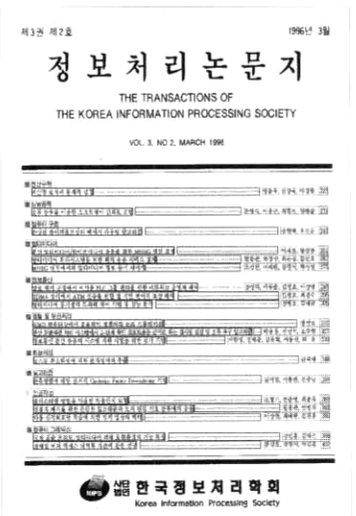
(그림 4-4) 차례 부분의 구조 분석 결과

(그림 4-4)는 분석된 Sub-Area, Title, Dot, Page 등에 대해서 다른 색상을 가지고 Out Linc을 표시해주어 인식기가 어떻게 판단했는지에 대한 정보를 보여준다. 본 연구에서

구현한 시스템은 T-A-P 형태에 대해서 실험적으로 구현이 되어 있다. 그래서 T-A-P 형태에 대해서는 자동으로 인식이 가능하도록 구성하였으며 실험결과는 다음 장에 제시되어 있다.

5. 실험 결과 및 분석

실험에서는 T-A-P 형식을 갖는 컴퓨터 관련 학회의 논문지와 영문 논문지, 기타 일반 논문지 등 다양한 종류의 논문지를 스캔한 영상을 대상으로 구조분석을 수행하여 논문제목, 논문저자, 페이지 정보 등을 추출하였다. (그림 5-1)과 (그림 5-2)는 각기 다른 논문지의 표지로부터 T-A-P 정보를 추출한 결과를 보이고 있다. 그림에서 파란색으로 표시한 영역은 제목을 뜻하고, 녹색으로 표시한 영역은 저자이며, 붉은색 영역은 구분자인 점열, 자주색 영역은 페이지로 인식된 결과를 나타내고 있다.


제3권 제2호 1996년 3월

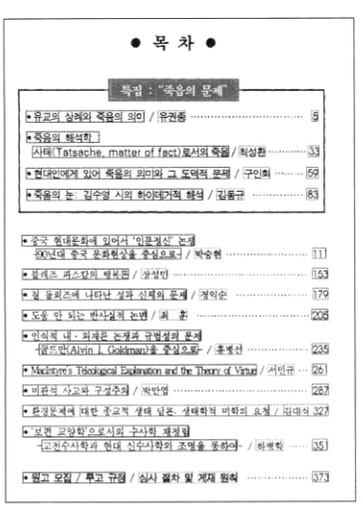
정보처리논문지

THE TRANSACTIONS OF
THE KOREA INFORMATION PROCESSING SOCIETY

VOL. 3, NO 2, MARCH 1996

한국정보처리학회
Korea Information Processing Society

(그림 5-1) 정보처리학회지 목차 분석 결과


● 목 차 ●

특집 : "죽음의 문제"

- 유교의 실제와 죽음의 의미 / 유권종 5
- 죽음의 해석학 : 사태(Tatsache, matter of fact)로서의 죽음 / 최성환 33
- 현대인에게 있어 죽음의 의미와 그 도덕적 문제 / 구인회 59
- 죽음의 눈 : 김수영 시의 하이데거적 해석 / 김동규 83
- 중국 현대문화에 있어서 '인문정신' 논재 : 80년대 중국 문화현상을 중심으로 / 박승현 111
- 플라즈 코스칼의 생보론 / 장성민 153
- 칼 슬리츠에 나타난 성과 신체의 문제 / 정익순 179
- 도중 안 되는 반사실적 논변 / 최 훈 208
- 인식적 내 - 외재론 논쟁과 규범성의 문제 : 골드먼(Alvin I. Goldman)을 중심으로 / 홍병선 235
- Mackenzie's Teleological Explanation and the Theory of Virtue / 서민규 261
- 비판적 사고와 구성주의 / 박만엽 287
- 환경문제에 대한 종교적 생태 담론: 생태학적 비판의 요청 / 김대의 327
- '보편 교양학'으로서의 수사학 재정립 : 신전수사학과 현대 신수사학의 조명을 통하여 / 하병학 351
- 필고 모집 / 투고 규정 / 심사 절차 및 게재 원칙 373

(그림 5-2) 일반 논문지 분석 결과

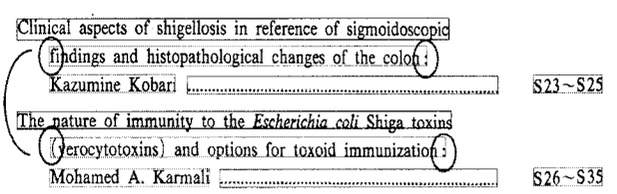
〈표 5-1〉 T-A-P 형태 논문지의 실험 결과

	게재 논문 수	게재 논문 정보 추출 편수 (추출성공률)
정보관련학회논문지	288편	274편 (95.1%)
기타 논문지	372편	330편 (88.7%)
평 균	660편	604편 (91.5%)

본 연구에서 제안한 목차구조 분석방법을 바탕으로 구현된 목차 분석 시스템에 대한 실험결과는 <표 5-1>과 같다. 실험 대상으로 쓰인 정보처리학회 논문지와 정보과학회 논문지 및 기타 일반논문지에 대해서도 같은 형태에 속하는 목차구조를 가진 것 중 비전산계열의 것을 무작위로 선정하였다. 구현된 목차 분석 시스템은 본 연구에서 목차구조의 자동 인식을 위해 일반화 방법으로 제안된 전체 모델 중 T-A-P 형태에 대한 검증을 위한 것이므로 이를 만족시키는 것만을 실험 대상으로 한 것이다.

입력정보는 정보처리학회 논문지와 정보과학회 논문지 등 2종의 컴퓨터 관련학회 논문지에서 288편의 목차를 대상으로 하였으며 국내에서 발간된 영문 논문지인 Journal of EEIS, 한국부식학회지, 인문과학 논문지 등에서 373편을 추출하여 총 660편을 대상으로 분석을 수행하였다. 예를 들면 정보처리학회 및 정보과학 논문지에 게재된 논문의 표지에 기록된 논문이 총 288편이었으며 이를 목차 분석 시스템으로 실험한 결과 논문지에 실린 논문 중 95.1%인 274편의 논문에 대해서 항목별 정보(T-A-P)를 정확하게 추출하는데 성공했음을 나타내고 있다. 3가지의 구성 요소 중에서 하나라도 잘못 추출된 경우는 추출 성공률에서 제외하였다. 실험 결과에서 볼 수 있듯이 기타 논문지의 경우 성공률이 상대적으로 낮게 나타났는데 이것은 같은 형태에 속하는 것이라 하더라도 목차의 구조가 복잡하고 예외적인 요소가 많아 구조해석에 어려움이 따랐기 때문이다. 이 결과를 비교하기 위하여 Lin의 연구 결과를 보면 OCR의 도움을 받아 10종의 다른 논문지에 대하여 실험을 한 결과 94%의 목차 분석 성공률을 얻고 있다. 그러므로 본 연구에서 제시한 방법은 OCR을 사용하지 않고 목차의 특성 분석에 의한 a priori 정보만을 이용하였음에도 불구하고 이를 상회하는 분석 결과를 얻음으로써 제안된 방법이 우수하다는 것을 보이고 있다.

분석에 실패한 내용을 분석해 보면 다음과 같다. (그림 5-3)에서는 발견된 문제는 저자 영역의 오인식이다. 위의 예에서는 논문에 대한 정보가 세 줄로 나뉘어 표시되고 있는데 첫 행에 이어 논문제목의 나머지를 나타내는 두 번째


Clinical aspects of shigellosis in reference of sigmoidoscopic findings and histopathological changes of the colon :
Kazumine Kobari ..... S23~S25

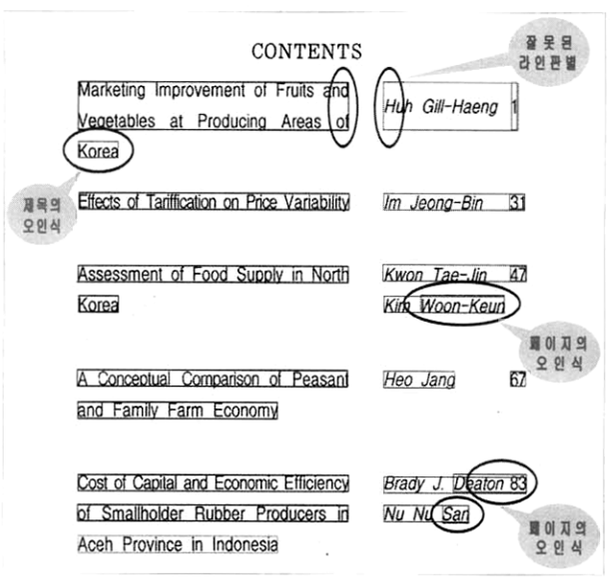
The nature of immunity to the *Escherichia coli* Shiga toxins (verocytotoxins) and options for toxoid immunization :
Mohamed A. Karmali ..... S26~S35

(그림 5-3) 구조분석 실패의 예

문자열이 들여쓰기가 되어있으므로 저자영역으로 잘못 판별되었다. 이와 같은 문제를 해결하기 위해서는 ':'를 구분자로써 활용하는 방법이 필요할 것으로 보인다.

(그림 5-4)에서는 복합적인 문제점들이 나타나고 있다. 우선 첫 번째로 나타난 저자 영역이 논문제목을 나타내주는 행과 행 사이에 걸쳐 나타나고 있다. 본 연구에서 제안한 방법에서는 목차구조를 분석하는데 있어 프로젝션 방법을 주로 사용하고 있으므로 이와 같은 경우가 발생하면 영역을 검출하는데 어려움이 따른다. 위의 예에서는 저자영역이 논문제목을 나타내주는 행과 행 사이에 겹쳐서 걸침으로써 다행히 이에 대한 프로젝션 결과가 하나의 행으로 처리되고 있다. 그러나 행과 행 사이에 걸치지 않고 완전히 독립적인 행으로 나타날 경우 이를 세부 항목으로 판별하는데 문제점이 발생할 수가 있다.

두 번째로 3행에서 나타난 Sub Area의 오인식이다. 이것은 첫 번째로 발견된 문제와 연관되어 있다. 논문정보가 여러 줄에 나타나는 경우에는 위나 아래의 문자열에 대해서 조건을 검사하여 합병을 하게 되는데, 이 경우에 위쪽의 1, 2행이 합쳐져 행의 높이가 커짐에 따라 그 다음에 나타난 'Korea'가 다른 높이를 갖는 별개의 영역요소로 처리된 것이다. 마지막으로 6, 9, 10행에서 나타난 페이지 영역의 오인식이다. 6행에서는 제목과 저자 등의 일정 패턴을 충족시키는 요소들이 나타남으로 인해서 페이지 영역을 찾으려 하는 잘못된 결과를 생성하였다. 더욱이 9행의 경우 저자의 이름을 나타내는 단어에서 저자영역과 페이지 영역의 구분이 되는 간격보다도 큰 공백이 나타남으로 인해서 올바른 페이지 영역이 존재함에도 불구하고 잘못 분류를 하게 되었다. 이렇게 복잡한 구조를 갖는 목차 형식의 경우에는 그 특징을 일반화하기가 무척 어려울 것으로 판단되므로 다양한 패턴을 가진 목차형식에 대한 많은 분석과 일반화를 위한 연구가 필요하다.



(그림 5-4) 구조분석 실패의 두 번째 예

오인식 결과들에 대한 분석으로부터 찾아낸 문제와 본 연구의 궁극적인 목표인 데이터 베이스 시스템과의 연계를 생각한다면 몇 가지 대안적인 방안도 필요할 것으로 보인다. 매우 다양한 논문지를 데이터베이스에 입력하는 경우, 시스템의 완전 자동화는 불가능하므로 이에 대한 메타 데이터를 도입하는 방안을 고려해 볼 필요가 있다. 논문지 종류가 다양하다 하더라도 한 종류에 해당하는 논문지의 권수는 상당한 분량일 것이 당연하므로 논문지 목차 형식에 대해서 어느 정도의 메타 정보를 목차 분석 시스템에 준다면 반자동화의 형식을 지니게 되지만 인식률 및 정확성을 비약적으로 향상시킬 수 있을 것이고 데이터베이스 시스템도 보다 효과적으로 논문 정보를 색인에 이용할 수 있을 것이다. 실제로 자동화를 꾀하는 데이터베이스 시스템들이 효율을 높이기 위해 이와 유사한 방법으로 메타데이터를 활용하고 있고, 최신의 데이터베이스 시스템이라 하더라도 정보의 입력은 아직 대부분이 사람에 의한 수작업으로 이루어지고 있는 것을 보면, 이러한 대안은 실제적으로 본 연구의 가치를 높이는데 큰 도움을 줄 것으로 생각된다.

6. 결론 및 향후과제

문서의 형태로 저장되어 있는 산적한 정보를 전산화하기 위해서 문자인식기수과 더불어 문서분석기술은 반드시 필요한 기술 분야이다. 단순하게 낱자만을 인식할 수 있는 문자인식기에 문서분석기술을 도입하면 다양한 문서로부터 낱자를 추출하여 인식기로 처리할 수 있을 뿐만 아니라, 문서내의 여러 가지 요소를 세밀하게 분석하여 체계적으로 데이터베이스화 할 수 있다.

본 논문에서는 일반적인 문서와 상이한 특성을 갖는 논문지 목차 형식을 구조적으로 분석하였으며, 이 결과를 바탕으로 여러 가지 목차 형식을 6가지 모델로 분류하고 특성을 분석함으로써 인식에 도움이 되도록 하는 방안을 제시하였다. 그리고 제안한 모델링 방법을 기준으로 목차 인식 시스템을 구현하여 T-A-P 형식에 대하여 관련 논문지를 대상으로 검증하였다.

기존의 문서분석방법으로는 예외적인 부분이 많은 목차 형식을 효과적으로 분석할 수 없었으며 우리가 원하는 일부분만을 추출할 수가 없었다. 본 연구에서는 목차의 효율적인 인식을 위해 다양한 논문지 목차의 형태학적인 성질과 지정학적인 성질을 분석하여 기술하였다. 논문지 목차형식에서 궁극적으로 표현하고자 하는 가장 필수적인 요소는 그 논문지에 실린 논문의 제목과, 저자, 페이지 등의 세 항목이므로 이 세 가지 항목을 배열순서에 따라 모델화 하여 이들 모델에 따른 특성을 일반화하였다. 마지막으로 본 연구에서 제안된 방법을 검증하기 위해 구현된 목차인식 시스템에서는 목차의 구조적인 특징을 분석하는데 효과적인 프로젝션 방법과 지정학적인 특성을 이용하여 각 항목을 구분하였으며 각 항목이 갖는 의미까지도 구별하여 논문 색인 정보로써 자동 추출할 수 있도록 하고 이를 문자로써 인식하는 데

에도 도움이 되도록 하였다.

본 연구에서 제안하는 방법을 바탕으로 구현된 목차 분석 시스템은 T-A-P 형식을 대상으로 하였으므로 나머지 5가지의 형태를 포함한 시스템으로 확장할 필요가 있다. 또한, 메타 데이터를 사용하여 시스템이 특정 목차 형식에 대한 정보를 얻을 수 있도록 하거나, 검증 단계를 두어 사람이 개입할 수 있도록 함으로써 효율의 향상과 정확도의 개선이라는 두 가지 효과를 함께 얻도록 할 필요가 있다. 또한 문자인식을 추가적으로 적용하면 개선된 논문지 목차 형식 분석 결과가 얻어질 것으로 예측된다. 이 연구에서 얻어진 결과는 정보처리학회의 논문지와 같은 형식을 갖는 도서관에 보관된 논문지의 스캔에 따른 인덱스 구축 작업에서 효율적으로 활용이 가능하다.

참 고 문 헌

[1] 김병기, "연결요소와 색상정보를 이용한 실제적 문서영상 분할", 한국정보처리학회 논문지 A, Vol.7, No.1, pp.273-285, 2000.

[2]변영철, 최영우, 김경환, 이일병, "부분 매칭 방법을 이용한 효율적인 서식 문서 분류", 한국정보처리학회 논문지 B, Vol. 8-B, No.1, pp.1-9, 2001.

[3]이경호, 최윤철, 조성배, "문서 영상의 논리적인 구조 분석을 위한 구문론적인 접근 방식", 한국정보과학회 논문지 B - 소프트웨어 및 응용, Vol.28, No.7, pp.524-536, 2001.

[4] 이성환, 문자인식-이론과 실제, 홍릉과학 출판사, pp. 87-108, 1993.

[5] 장대근, 오원근, 양영규, "연결요소와 영역확장을 이용한 문서영상 분할", 한국정보처리학회 제12회 추계학술대회 발표논문집 CD, 일련번호 312, 1999.

[6] 장대근, 황찬식, "이미지 필터와 제한조건을 이용한 문서영상 구조분석", 한국정보처리학회 논문지 B, Vol. 9-B, No.3, pp.311-318, 2002.

[7] 장명욱, 천대녕, 양현승, "연결화소를 이용한 문서 영상의 분할 및 인식" 한국정보과학회 논문지, Vol. 20, No. 12, pp.1741-1751, 1993.

[8] 전병태, 배영래, 양영규, 오길록, "다단계 특징 추출에 의한 일반화된 자막 영역 추출 방법", 제12회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp.429-434, 2000.

[9] 정창부, 김수형, "문서 영상 내 테이블 영역에서의 단어 추출", 한국정보처리학회 논문지 B, Vol. 12-B, No.4, pp. 369-378, 2005.

[10] A. Belaid, L. Pierron, and N. Valverde, "Part-of-Speech Tagging for Table of Contents Recognition", Proceedings of the International Conference on Pattern Recognition, pp451-454, 2000.

[11] S. Bow and R. Kasturi, "A Graphics-Recognition System for Interpretation of Line Drawing", in Image Analysis Applications, Marcel Dekker, pp.37-72, 1990.

[12] R. Crane, A simplified approach to Image Processing, Prentice Hall, 1997.

[13] L.A. Fletcher and R. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.10, No.6, pp.910-918, 1988.

[14] Gonzalez & Woods, Digital Image Processing, Addsion Wesley Longman, 1992.

[15] Y. Hirayama, "A Block Segmentation Method for Document Image with Complicated Column Structures", Proceedings of the 2nd International Comference on Document Analysis and Recognition, pp.91-94, 1993

[16] X. Lin and Y. Xiong, Detection and Analysis of Table of Contents Based on Content Association, Hewlett -Packard Technical Report, HPL-2005-105, May 31, 2005.

[17] S. Mandal, S.P. Chowdhury, A.K. Das, and B. Chanda, "Automated Detection and Segmentation of Table of Contents Page from Document Images", Proceedings of the 7th International Conference on Document Analysis and Recognition, pp.398-402, 2003.

[18] L. O'Gorman, "The Document Spectrum for Page Layout Analysis", IEEE Trans. on PAMI, Vol. 15, No. 11, pp1162-1173, 1993.

[19] L. O'Gorman and R. Kasturi, Document Image Analysis, IEEE, 1996.

[20] S. Tsujimoto and H. Asada, "Major Components of A Complete Text Reading System", Proceedings of IEEE. Vol. 80, No.7, pp.1133-1149, 1992.

[21] S. Tsuruoka and C. Hirano, "Image-based Structure Analysis for a Table of Contents and Conversion to XML Documents", Proc. DLIA Workshop, 2001.

[22] F.M. Wahl, K.Y. Wong and R.G. Gasey, "Block Segmentation and Text Extraction in Mixed Text/Image Document", Computer Graphics and Image Processing, Academic Press, pp.375-390, 20. 1982.

[23] D. Wang and S.N. Srihari, "Classification of Newspaper Image Block Using Texture Analysis", Computer Vision, Graphics and Image Processing, Vol.47, pp327-352, 1989.

권 영 빈

e-mail : ybkwon@cau.ac.kr

1978년 아주대학교 전교수석(공학사)

1981년 한국과학기술원(공학석사)

1986년 프랑스 파리 ENST(공학박사)

1986년~현재 중앙대학교 컴퓨터공학부 교수

2007년~현재 (사)대학정보화협의회 회장

관심분야 : 패턴인식, 문자인식, 생체인식, 자동인식 등