테이블 내의 호목단 구조 판별 자동화에 대한 연구

조성수* , 김명호*
*숭실대학교 컴퓨터공학과
e-mail:xxsoo@naver.com

A Study on Automated HoMokDan Structure Determination in Table

Sung-Soo Cho* , Myung Ho Kim*
*Department of Computing, SoongSil University

요 약

현재 법률과 관련된 문서들은 변경 사항에 대한 공표와 기록의 중요성을 가지고 있다. 따라서 변경 사항을 자동으로 인지하고 공표할 수 있는 자동화 시스템에 대한 관심과 연구가 진행되고 있다. 그러나 대부분의 문서들은 복잡한 구조이기 때문에 자동화에 어려움이 많다. 이로 인해 복잡한 구조의 문서를 자동으로 판별할 수 있는 방법에 관한 관심이 증대되고 있다. 현재 국내외에서는 전자 문서 파일의 텍스트 및 테이블을 판별해서 분류하는 자동화에 대한 연구가 진행되고 있다. 하지만 이전 연구에서는 호목단 구조를 갖는 계층적인 테이블을 판별하지 않는다. 그래서 본 논문에서는 호목단을 정의하고, 테이블의 호목단 구조를 패턴 별로 분류하며, 테이블의 호목단 구조 판별 방법을 제시한다.

1. 서론

현재 법률과 관련된 문서들은 변경 사항에 관한 내용을 알지 못하면 불이익을 당할 수 있다. 따라서 이러한 문서의 변경사항을 자동으로 인지하고 공표할 수 있는 자동화 시스템에 대한 관심과 연구가 진행 되고 있다[1]. 그러나 이런 대부분의 문서는 중첩 테이블, 호목단 등의 복잡한 구조를 포함하고 있다. 이로 인해 복잡한 구조의 문서를 자동으로 판별할 수 있는 방법에 관한 관심이 증대되고 있다.

국내외에서는 전자 문서 파일의 텍스트 및 테이블의 메타 데이터 추출 자동화에 대한 연구가 진행되고 있다[2]. 이전 연구에서는 테이블 내의 계층적인 호목단 구조를 판별하지 않는다. 그 이유는 호목단 구조의 테이블은 각 셀이 다른 셀들과의 관계를 갖는데 이 구조를 파악하기 힘들기 때문이다. 본 논문의 테이블의 호목단 구조 판별을 사용하면, 셀의 구조적 위치를 알 수 있다. 이것을 이용하여 수정사항을 표현하면, 단순히 “몇 행 몇 열이 수정 되었다.”가 아니라 “제 1호의 나목이 수정 되었다.”처럼 직관적으로 표현할 수 있다.

본 논문 2장에서 기존 호목단 관해 알아보고, 3장에서 테이블의 호목단 구조를 패턴 별로 분류하며, 4장에서 테이블의 호목단 구조 판별 방법을 제시한다.

2. 호,목,단의 정의

일반 호목단은 각종 문서에서 쓰이고 있고, 특히 법률 문서에서 많이 쓰인다.

2.1 호,목,단 표시

호 : 숫자 “1. , 2.” 로 표시하고, 점과 내용 사이에 한 칸을 띄우며, “제 1호, 제 2호” 로 지칭한다.

목 : 문자 “가. , 나.” 로 표시하고, 점과 내용 사이에 한 칸을 띄우며, “가목, 나목” 으로 지칭한다.

단 : 숫자 “(1) , (2)” 로 표시하고, 중괄호와 내용 사이에 한 칸을 띄운다. “1단 , 2단” 으로 지칭한다. 단은 호, 목과 달리 쓸 때는 “(1)”라고 만 쓴다.

2.2 호,목,단 규칙

호목단은 계층적인 구조를 가진다. 즉, <표 1> 호목단의 사용 예제에 나타낸 것과 같이 호는 다수의 목을 포함하고, 목은 다수의 단을 포함한다. 한 라인의 시작은 호와 목과 단의 표시로 시작되고, 한 라인 중간에 이어서 호목단 작성이 안 된다. 제 1호에 가목이 있고, 제 2호에도 가목이 있는 것처럼 상위 호나 목이 다음 차례로 넘어가면, 목이나 단은 순서를 처음부터 다시 시작 한다. 하지만 같은 레벨의 호, 목, 단의 중첩은 허용하지 않는다. 또한, 호, 목, 단이 차례로 나와야된다. 호 다음 단이 나올 수 없다.

<표 1> 호목단의 사용 예제

1. 제 1호 이다.
  가. 제 1호의 가목이다.
  나. 제 1호의 나목이다.
    (1) 제 1호 나목(1)이다.
2. 제 2호 이다.
  가. 제 2호의 가목이다.

3. 테이블 내 호목단 패턴 분류

현재 호목단 테이블의 생성 방법에 관한 명확한 정의나 표준은 없다. 하지만 현대 사회에서 많은 전자 문서들이 호목단 테이블을 포함하고 있다. 그 많은 문서 중 호목단을 가장 많이 포함하고 있는 것은 법률이다. 호목단의 구조와 패턴을 분석하기 위해서 법률 정보를 제공하는 국내 법체처의 현행 별표에서 테이블 정보를 수집 했다.

3.1 테이블 분류

현행 별표의 수집된 테이블 정보 중 4분의 1을 호목단 구조의 테이블로 분류를 했다. 호목단 구조의 테이블의 분류 규칙은 단순하다. 테이블에 호목단이 들어가 있어야 한다. 그리고 호목단이 테이블의 다른 셀과 연관 관계가 있어야 한다. 예로, 테이블의 한 셀에 호목단이 있으나 테이블 내의 다른 셀과 관련이 없고, 해당 셀 내에서 호목단 구조가 끝나면 호목단 형태의 테이블이 아니다.

호목단 테이블로 분류된 문서들의 테이블 구조를 분석했다. 그 결과 5개의 호목단 테이블의 구조를 찾았다. 테이블의 구조는 <표 2>에서 <표 6> 과 같다.

<표 2> 호목단 테이블 패턴1(a)

(*호가 테이블 바깥에 있을 수도 있다.*)

제목	제목	제목
1. 호 들어감	내용이 들어감	내용이 들어감
가. 목이 들어감	...	...
...	...	...

<표 3> 호목단 테이블 패턴 1(b)

제목	제목	제목
1. 호가 들어감 가. 목이 들어감 ... 2. 호가 들어감 ...	내용이 들어감	내용이 들어감

<표 4> 호목단 테이블 패턴 1(c)

제목	제목	제목
1. 호가 들어감 가. 목이 들어감 ...	내용이 들어감	내용이 들어감
2. 호가 들어감 ...	내용이 들어감	내용이 들어감

<표 5> 호목단 테이블 패턴 2(a)

제목	제목	제목
1. 호가 들어감	가. 목 또는 내용이 들어감	단 또는 내용이 들어감
	나. 목 또는 내용이 들어감	...
	다. 목 또는 내용이 들어감	...
2. 호가 들어감	가. 목 또는 내용이 들어감	...

<표 6> 호목단 테이블 패턴 2(b)

제목	제목	제목
1. 호가 들어감	가. 목 또는 내용이 들어감 나. 목 또는 내용이 들어감	내용이 들어감
2. 호가 들어감	가. 목 또는 내용이 들어감 나. 목 또는 내용이 들어감	내용이 들어감

3.2 호목단 패턴 분류

호목단 구조의 테이블 5개를 호목단의 진행 패턴으로 분류하면 2개의 패턴이 나온다. 첫째, 호목단의 진행 방향이 아래쪽인 패턴이다. <표 2>, <표 3>, <표 4>가 이에 해당되며, 이것은 <표 2>로 묶을 수가 있다. 이것을 “패턴 1”로 지칭하겠다. 둘째, 호목단의 진행 방향이 우측인 패턴이다. <표 5>, <표 6>이 이에 해당되며, 이것은 <표 5>로 묶을 수가 있다. 이것을 “패턴 2”로 지칭 하겠다.

“패턴 1”은 1열에 모든 호, 목, 단을 모두 나타낸다. 즉, 1열의 각 행마다 호, 목, 단이 들어간다. 그리고 1열을 뺀 다른 열들은 모두 내용들이 들어가거나, 공백이다. 2열 부터 하나의 열에 같은 깊이에 해당하는 호, 목, 단 내용만을 표시한다. “패턴 2”는 1열은 호만 들어간다. 그리고 2열은 목이나 내용만 들어가며, 3열은 단이나 내용만 들어간다.

4. 테이블 내 호목단 판별

호목단 테이블은 “패턴 1”과 “패턴 2”의 형태를 가진다. 그래서 “패턴 1”과 “패턴 2”를 가지고 호목단 테이블을 판별해 내야한다.

4.1 필드의 타입

[2]에서 테이블의 필드들을 박스(Box)라 부르고, 박스를 4타입 BLK, INS, EXP, IND로 분류 했다. 본 논문에서는 IND_HMD를 추가 한다. 내용은 <표 7>과 같다.

<표 7> 박스 타입과 라벨

Label	Description
BLK	타입은 박스가 비어 있는 빈칸(blank)
INS	미리 인쇄 된 문자에 붙이거나 채워 넣을 수 있는 추가(insertion)
EXP	박스에 일반적인 설명을 포함하는 설명(explanation)
IND	추가나 빈칸의 항목을 표시하는 표시(indication)
IND_HMD	IND의 확장으로 호목단 구조임을 나타냄

우리는 박스의 타입을 사용하여 테이블을 라벨로 분류하고, 이 작업을 “라벨링”이라 할 것이다. 하지만 [1]은 오직 4가지 타입만을 사용한다. 그러나 호목단 구조를 가지는 테이블을 판별하기 힘들다. 그래서 본 논문은 IND를 나눠 호목단을 표시하는 IND_HMD 라벨을 추가한다. 즉, IND_HMD는 본 논문 2장 에서 정의한 호목단의 패턴을 가지는 IND이다. 이렇게 본 논문은 5개 타입의 라벨을 사용한다. <표 2>와 <표 5>를 각각 라벨링 하면, <표 8>, <표 9>이다.

<표 8> 패턴 1의 라벨링

EXP	EXP	EXP
① IND_HMD	② EXP	BLK
③ IND_HMD	④ BLK	EXP
⑤ IND_HMD	⑥ ...	...

<표 9> 패턴 2의 라벨링

EXP	EXP	EXP
IND_HMD	IND_HMD	EXP
	IND_HMD	EXP
	IND_HMD	EXP
IND_HMD	IND_HMD	EXP

4.2 호목단 테이블 판별

테이블내 호목단 구조 판별은 라벨링 완료 후 진행 된다. 라벨링된 호목단 구조는 1열은 제목이고, 다음 행은 무조건 IND_HMD 가 들어가야 한다. 만약 이 조건을 만족 하지 않으면 해당 테이블 라벨링은 IND_HMD을 사용하지 않는다. 이 경우 해당 테이블은 오직 IND 만을 사용한다. 그래서 <표 8>에서는 ①로 호목단 테이블의 유무를 확인 한다. 즉, IND_HMD이면 호목단 테이블이고, IND이면 호목단 테이블이 아니다.

4.3 호목단 패턴 판별

라벨링된 테이블의 “패턴 1”과 “패턴 2”의 분류는 쉽다. 2열에 IND_HMD 라벨이 있으면 “패턴2”이고, 없으면 “패턴 1”이다.

다음은 호목단을 구분해야 한다. “패턴 2”는 호목단의 진행 방향이 우측이다. 그래서 열 증가를 보고 호목단을 쉽게 판단할 수 있다. <표 9>의 1열은 호이고, 2열은 목이다. “패턴 1”은 호목단의 진행 방향이 아래쪽이다. 그러나 행의 증가로는 호목단을 구분할 수 없다. <표 8>에서 호목단을 구분하기 위해서는 1개의 열에는 하나의 호목단 내용만 들어가는 “패턴 1”의 특성을 사용해야 한다. <표 8>의 ①은 시작되는 호(목 일수도 있고,)이고, ②는 EXP로 해당 열이 EXP인 것은 호다. 다른 셀은 BLK가 들어간다. ④가 BLK이므로 ③은 ①의 호에 속한 목이다. ⑤는 아직 호목단이 정해져 있지 않다. ⑤는 ⑥에 따라서 호목단이 정해진다. ⑥이 EXP라면 ⑤는 ① 다음 호가 되고, BLK이라면 ⑤는 ①에 속한 ③다음의 목이 된다.

5. 결론 및 향후 과제

본 논문은 현재 존재하는 문서들의 테이블을 분석하여 호목단 테이블의 패턴을 찾고, 그 패턴을 적용했다. 그럼으로 문서에 테이블이 호목단 구조라는 표시 없이도 자동화되어 인식된다.

그러나 많은 문서들 중에는 호목단 구조의 테이블뿐 아니라 다른 계층적인 구조나 중첩된 구조를 갖는 복잡한 테이블들이 존재 한다. 계속해서 이러한 테이블들을 분석하고 적용시키면 다른 구조의 복잡한 테이블 또한 자동화할 수 있다.

참고문헌

[1] T. Watanabe, Q. Luo, N. Sugie "Layout Recognition of Multi-Kinds of Table-Form Documents," IEEE PAMI, pp.432-445,1995

[2] A Amano, N Asada, and M Mukunoki "Modification Table Form Generation System based on the Form Recognition," IEEE 17th ICPR, pp.659-664, 2004

[3] 법제전문교육훈련기관 법제교육포탈 법령 조항과 별표 서식. http://edu.klaw.go.kr/StdInfInfoR.do?astSeq=96