패턴인식기법을 이용한 편목전문가시스템 설계

A Study on Design Of Cataloging Expert System

김현희(명지대학교 문헌정보학과)

Hyun-Hee Kim(Myong Ji University)

본 연구에서는 표제면과 판권지의 서지요소의 레이아웃 특성과 구문적 특성을 이용하여 서지요소의 종류를 패턴인식 지식베이스와 전거화일들을 이용하여 자동 인지하고 인지된 서지요소를 한국문헌자동화목록형식(KORMARC)과 한국문헌자동화목록법(KORMARC) 기술규칙에 기초하여 KORMARC 형태로 출력해 주는 편목전문가시스템을 구축하였다.

1. 서 론

본 연구에서 설계한 시스템은 문헌 표제면과 판권지의 서지요소의 종류를 패턴인식기법에 의해 판독하고 판독된 서지요소를 KORMARC 포맷에 맞춰 출력하는 시스템이다. 먼저 서지요소를 자동판독하기 위해서 18개의 서지요소의 종류를 구분할 수 있는 패턴인식 지식베이스과 전거화일들을 구성하였다.

패턴인식 지식베이스는 크게 레이아웃 규칙(물리적)과 구문적 규칙으로 구분할 수 있다. 레이아웃 규칙이란 문자의 유형(한글 명조체, 한글 고딕체)과 문자의 크기, 데이타의 위치를 규정하여 서지 요소를 판독하는 규칙이다. 한편 구문적이라는 것은 표제면 상에서 저자명 다음에 오는 저작 역할어(저, 공저, 역) 등과 같은 어휘 단위를 이용하여 서지요소를 판독하는 규칙이다.

다음은 서지요소의 종류가 확인되면 이 데이타를 기초로 하여 KORMARC 포맷에 맞춰 출력해 줄 수 있는 편목 지식베이스를 구성하였다. 이때 이용된 지식베이스는 한국문헌자동화목록형식(KORMARC)과 KORMARC 기술규칙에 기초하여 구성하였다.

본 시스템에서 구축한 패턴인식 지식베이스와 편목 지식베이스는 생성규칙을 이용하여 지식을 표현했으며 추론 방법은 전진추론을 이용하였다. 시스템을 운용하는 모든 프로그램은 Turbo-C언어를 이용하여 작성하였다.

2. 편목전문가시스템 설계

2.1 시스템의 개관

본 연구를 통해 설계한 시스템은 문헌 표제면/판권지의 서지요소의 종류를 자동인지하고 인지된 서지요소를 KORMARC 형태로 출력해 주는 편목 전문가시스템으로 이 시스템은 크게 지식베이스, 지식획득시스템, 서지요소 자동인지 알고리듬, 편목 알고리듬, 전거화일들로 구성된다.

2.2 지식베이스 설계

본 시스템에서 구축한 지식베이스에는 패턴인식 지식베이스와 편목 지식베이스가 있다. 패턴인식 지식베이스는 표제면의 서지 요소의 종류를 판독하기 위하여 구축한 생성규칙이고 편목 지식베이스는 판독된 서지요소들을 KORMARC 포맷으로 출력할 수 있도록 해 주는 생성규칙이다. 패턴인식 지식베이스는 12개의 서지요소의 종류를 자동 인지하기 위해서 구축한 22개의 생성규칙으로 구성되며 편목 지식베이스는 표시기호 생성규칙, 지시기호 생성규칙, 식별기호 생성규칙의 세가지 종류의 생성규칙으로 이루어진다.

2.3 지식획득시스템

본 시스템에서 구축한 지식베이스에는 패턴인식 지식베이스와 편목 지식베이스가 있다. 패턴인식 지식베이스를 구축하기 위해서 먼저 155개의 실험 문헌집단의 표제면과 판권지의 서지 데이타를 분석하였고 이 실험 문헌집단에 포함되지 않는 도서관의 문헌들도 분석하였다. 참조한 문헌으로는 단행본용 한국문헌자동화목록법기술규칙(1985년) 등이 있다. 편목 지식베이스는 단행본용 한국문헌자동화목록형식(1993년), 단행본용 한국문헌자동화목록법기술규칙(1985

년) 등의 문헌들과 편목전문가들의 의견을 참조하여 구성하였다.

2.4 알고리듬 설계

다음은 지식베이스들과 전거화일들을 이용하여 표제면과 판권지의 서지요소을 인지해 주는 서지요소 자동인지 알고리듬과 인지된 서지요소를 KORMARC 형태로 출력해 주는 편목 알고리듬을 설계하였다

2.4.1 서지요소자동인지 알고리듬

1) 실험 데이타를 선택 : 155권의 실험문헌집단 선정.

2) 단일어화일(표제면) 생성 : 표제면의 서지 데이타를 문자인식시스템을 이용하여 기계가독형화일인 표제면화일로 변환한다. 표제면화일에서 공란과 구두점을 단어 구분자로 하여 단어를 추출하고 각 단어에 속성(문자유형, 문자크기, 위치)을 수록하여 표제면 단일어화일을 구성한다.

3) 事前 서지요소체크(표제면) : 표제면의 단일어화일을 입력화일로 하여 저작역할어, 발행사, 발행지의 서지요소를 체크한다.

4) 복합어화일 생성 : 사전 서지요소체크에서 서지요소가 구분이 안될 경우는 글자유형과 글자크기가 같고 사이간격이 50㎜이내에 있는 문자열을 결합하여 복합어화일을 만든다.

5) 事後 서지요소 체크

(1) 제 1 단계 : 표제면의 복합어화일을 분석하여 불용어를 제거하고 발행사, 발행지, 저자명을 인지한다.

(2) 제 2 단계 : 판권지 데이타화일을 이용하여 표제면에서 누락된 정보를 보완한다.

(3) 제 3 단계 : 패턴인식 지식베이스를 이용하여 12 종류의 서지요소를 인지한다.

2.4.2 편목 알고리듬

앞의 표제면과 판권지의 자동인지 절차를 통해서 서지요소들의 종류를 확인한 후 편목 지식베이스와 전거화일들을 이용하여 KORMARC 포맷에 데이타를 출력한다.

(1) 제 1 단계 : 개인저자군을 세

분하고 세분된 서지 코드를 할당해 준다.

(2) 제2단계 : 편목지식베이스에 의해 KORMARC 형태로 출력해 준다.

3. 편목전문가시스템 평가

이 시스템의 성능을 평가하기 위해서 패턴인식 지식베이스의 생성을 위해 분석한 155권의 실험문헌집단과 86권의 검증문헌집단을 이용하여 적중률을 조사해 보니 실험문헌집단의 경우는 94%, 검증문헌집단의 경우는 93%의 적중률을 나타냈다.

4. 결 론

본 연구에서 설계한 시스템은 패턴인식기법을 이용하여 서지 요소를 자동인지하고 인지된 서지요소를 KORMARC포맷에 따라 출력해 주는 시스템이다. 이 시스템은 문헌의 표제면과 판권지를 자동으로 판독하여 편목자들이 표제면과 판권지를 읽고 서지요소들을 확인하는 과정을 생략하여 신속하게 편목작업을 할 수 있도록 도와 줄 뿐 아니라 KORMARC에 익숙하지 않은 편목자들이 쉽게 KORMARC포맷을 배울 수 있도록 도와준다. 이 시스템의 효과로는 다음과 같은 점을 들 수 있다.

1) 편목에 소요되는 경비와 시간을 절약할 수 있다.

2) 편목업무의 표준화작업을 수행할 수 있다.

3) 국가 표준포맷인 KORMARC포맷을 활용한다.

4) 초보자들의 교육프로그램으로서의 역할을 한다.

5) 1)에 의해서 사서들이 정보서비스(참고서비스 등)에 좀 더 많은 시간을 할애할 수 있다.

본 시스템이 실제 업무에서 좀 더 효율적으로 활용되기 위해서는 먼저 문자유형과 크기를 그대로 아스키 화일로 재생시켜 주는 문자인식률이 향상된 문자인식 시스템이 개발되어야 한다. 또한 서지요소 자동인지의 모든 절차가 자동화되기 위해서는 본 연구에서 수작업으로 수행한 문자열 속성의 추출 과정을 수행해 주는 프로그램이 개발되어야 한다.

표제면/판권지 외에 패턴인식 기법에 의하여 서지요소를 인식할 수 있는 대상으로 CIP(Cataloging In Publication) 데이타가 있다.