이미지 화일과 텍스트 화일의 검색효율성 비교

A Comparison of Retrieval Effectiveness
between Image File and Text File

임 영선, 이두영
(중앙대학교 대학원 문헌정보학과)

Yim Young-sun, Lee Doo-young.
Dept. of Library & Information Science, Chung Ang University

본 논문은 본문 전체가 기계가독형 화일로 구성된 텍스트 전문데이터베이스와 이미지화일로 구성된 이미지 전문데이터베이스와의 검색효율성을 비교함으로써 도서관과 최종이용자의 입장에서 바람직한 전문데이터베이스가 어떤 것인지를 제안하고자 한다.

1. 서 론

오늘날 전문데이터베이스(Fulltext database)는 서지사항이나 초록 등 2차정보만을 수록하지 않고 원정보인 전문 전체를 수록하고 있는 데이터베이스를 의미한다. 따라서 전문데이터베이스는 서지데이터베이스에서 얻을 수 있는 정보외에 추가정보로서 전문이나 요약문과 인용문까지 출력되고, 원문을 문장별로 나누어서 전체원문에서 필요한 부분정보를 제공할 수 있어야 하며, 인용문의 전부를 원문 그대로 제공할 뿐만 아니라 삽도정보의 검색도 가능해야 한다. 또한 전문데이터베이스는 불용어를 제외한 알파벳의 2문자이상의 모든 명사들이 전부 키워드가 되기 때문에 최종이용자는 자연어와 전문탐색에 주로 사용하는 연산자인 인접연산자나 불리안 연산자들을 이용해 포괄적 탐색을 할 수 있어야 한다.

이와같은 전문데이터베이스를 조직하고 검색하기 위해서는 전문이 기계가독형 형태의 텍스트 화일로 만들어진 상태에서 검색프로그램을 실행할 수 있는 데이터베이스를 만들어야 한다. 그러나 현재 국내에서 사용하고 있거나 구축중인 전문데이터베이스의 경우 제목과 저자명등의 서지사항에서 추출된 색인어와 전문을 스캐닝(Scanning)해서 만든 이미지화일을 색인화일로 연결해서 구성한 전문데이터베이스를 사용하거나 구축하고 있는 실정이다.

본 연구에서는 이미지화일로 구성된 전문데이터베이스와 기계가독형 텍스트화일로 구성된 전문데이터베이스의 검색효율성을 측정함으로써 시스템의 유용성을 고찰하고자 한다.

2. 이론적 배경

2.1 전문데이터베이스의 특성

기계가독형 텍스트화일로 구성된 기존의 전문데이터베이스를 보면 첫째, 출력형태에 있어 서지데이터베이스의 기본출력형태와 같이 서명, 저자명, 기관명, 언어 및 출판사항에 관한 정보를 얻을 수 있으며, 추가정

보로서 전문이나 요약문과 인용문까지 출력된다. 또한 데이터베이스화일에 따라 약간의 차이는 있으나 원문을 문장별로 나누어서 전체원문에서 필요한 부분정보를 제공할 수 있어서 이용자의 시간 및 비용절감에 도움을 주고 있다. 또한 인용문의 전부를 원문 그대로 제공할 뿐 아니라 색상표, 삽도 및 사진과 같은 삽도정보를 화일에 입력시켜 이에 대한 탐색도 가능하도록 하고 있다.

둘째, 전문데이터베이스의 내용중 불용어와 알파벳의 2문자이상의 모든 명사들이 전부 키워드화되기 때문에 화일에 신규정보를 입력할 경우 전문색인자가 필요없게 된다. 따라서 전문색인자에 대한 인건비를 줄일 수 있다.

셋째, 이용자입장에서는 기초적인 터미날 조작방법만을 인지하면 어떤 이용자라도 즉시 탐색을 시작할 수 있는 장점이 있다. 또한 전문데이터베이스는 원정보에 대한 접근이 가능하고 독특한 정보의 접근이 용이하다. 전문데이터베이스의 장단점으로는 원문의 모든 단어가 대부분 디스크립터가 될 수 있기 때문에 접근점이 많다는 장점이 있지만, 접근점이 많을수록 그만큼 출력건수도 많아져서 전문데이터베이스는 재현율이 높은 반면에 정도율이 떨어지게 된다. 또한 부적합 정보의 출력건수가 많으면 많을수록 이용자는 추가의 비용을 부담해야 할 뿐만 아니라, 적합정보선택에 있어서도 혼란을 일으키게 된다는 단점도 지니고 있다.

2.2 전문데이터베이스의 구축원리

2.2.1 문헌의 구조화

문헌을 표준적인 방법으로 일관성있게 작성하고, 출력시 문헌요소마다 처리방법을 달리해 원하는 모양의 문헌을 얻기 위해서는 문헌을 구조적으로 보고 문헌을 이루는 문헌요소를 객체 단위로 보는 것이 필요하다. 1986년 ISO는 실제텍스트에서 각 문헌요소를 구별하기 위한 표준 마크업 언어(SGML : Standard Generalized Markup Language)를 제정한바 있다.

SGML은 개괄적으로 마크업 언어의 구문만을 정의한 메타언어(metalanguage)로서 마크업 언어에 대한 표준방식(standard mechanism)을 제공하고 있다. SGML문헌은 문헌선언부(Document Declaration), 문헌유형정의부(DTD ; Document Type Definition), 실제문헌부(DI ; Document Instance)로 이루어져 있으며, SGML문헌을 작성하기 위해서는 문헌유형을 정의한 후 실제 문헌을 작성하며 이렇게 작성된 문헌이 SGML 형태로 제대로 작성되었나를 검사하기 위해서 구문분석기를 사용한다.

전문데이터베이스 구축시 여러 마크업언어들중에서 표준으로 제정된 마크업언어인 SGML을 사용하게 되면 DTD를 통해 문서의 구조를 파악하여 문헌의 구성요소들이 나타내고 있는 의미까지 파악해 주므로 전문데이타베이스 구축 및 검색이 용이하고, 이미 국제표준으로 채택되어 있어 정보의 공유와 전달이 가능하다. 또한 미국출판자협회가 제공하는 수식/도표에 관한 DTD나 TEI가 제공하는 문서종류에 따른 DTD를 사용할 수 있는 장점도 가지게 된다.

2.2.2 문자인식시스템

문자인식시스템이란 인쇄매체로 작성된 문서를 영상형태로 입력하고 이를 분석하여, 문서상의 문자를 컴퓨터가 사용하는 내부코드로 바꾸어 텍스트화일(text file)을 만들어 내는 것을 말한다. 문자인식시스템은 문자영상 입력, 전처리, 문자인식, 후처리의 4단계로 구성된다.

1) 문자 영상 입력

문자영상을 입력받는데는 카메라, 스캐너, 팩시밀리 등이 이용되며, 신문기사 정도의 작은 글자도 인식될 수 있기 위해서는 300 DPI 이상의 해상도와 흑백의 상태로 입력받아야 한다. 문자영상 입력방법으로는 첫째, 영상의 처리결과를 카메라의 Zooming기능에 피드백시킬수 있도록 구성함으로서 적은 메모리와 짧은 시간내에 문

서영상을 처리할 수 있지만, 아날로그 신호를 디지탈 신호로 변환할 수 있는 고속의 A/D변환기가 부착되어 있는 카메라로 입력한다.
둘째, 문서영상을 레스터 스캔(raster scan) 방식으로 주사하면서 스캐너의 CCD(Charge Coupled Device)로 입력되는 영상신호를 저장하는 스캐너로 입력하는방법이 있다.
셋째, 팩시밀리로 입력받는 방법은 문서영상을 레스터 스캔방식으로 주사하여 팩시밀리의
CCD로 입력되는 영상신호를 MH부호화(Modified Huffman Coding)하여 저장된다.
팩시밀리는 원거리인식시스템, 원거리학습시스템을 구현하는데 유용하게 이용할 수 있다.

2) 전처리

인식대상 문서영상으로부터 문자영상을 분리 추출하여 문자인식 과정에서 인식할 수 있도록 하는 것으로서 블록화, 블록추출, 유형분류, 문자분리, 자모분리의 5단계로 이루어진다.
첫째, 일정거리 이내의 흑화소끼리 묶어줌으로써 영상을 형태별로 분리할 수 있도록 하는 블록화 방법으로 RLSA(Run Length Smoothing Algorithm)와 RXYC(Recurisive X-Y Cuts)방법이 있다.
둘째, 접속된 흑화소를 순번화(labeling)하는 방법과 접속된 흑화소 열을 추적하는 방법으로 나뉘는 블록추출방법이 있다.
셋째, 추출된 블록영상의 유형을 분류하는 것으로 블록의 흑화소의 평균 주행길이를 비교하는 방법과 블록의 속성을 이용하는 유형분류방법이 있다.
넷째, 문서에서 문자를 추출하는 방법에는 그림속의 문자를 효과적으로 추출할 수 있는 Hough Transform을 이용하는 방법과 그림속의 문자를 추출할 수는 없으나 문서를 한번 주사함으로써 비교적 적은 메모리와 시간에 그림과 문자를 분리하여 추출할 수 있는 Histogram을 이용하는 문자분리방법이 있다.
다섯째, 분리된 문자는 결합하고, 접촉된 문자는 분리하며, 문자의 자모를 분리함으로써 보다 쉽게 인식할 수 있는 자모분리 방법이 있다.

3) 문자인식

문자인식은 영상형태로 입력받은 문서상의 문자를 컴퓨터가 사용하는 내부코드로 바꾸어 텍스트화일을 만드는 과정으로 첫째, 인식대상이 되는 모든 문자의 영상을 2차원배열에 저장하고, 인식하고자 하는 입력문자의 영상을 저장된 각각의 문자영상과 대응되는 화소단위로 비교하여 불일치된 화소의 갯수가 가장 작은 문자로 입력문자를 판단하는 원형비교방법이 있다. 한글인식에 이 방법을 그대로 적용하는 것은 불가능하다.
둘째, 통계적방법은 인식대상인 각 문자에 대한 충분히 많은 문자영상을 추출하고, 각각의 문자영상에 대해 정해진 방법에 따라 N개의 특성값을 추출하여 N차원 공간의 백타로 표현하여, 이 백타들의 평균값과 가장 가까운 거리에 있는 특성 백타에 해당하는 문자로 인식하는 방법이다.
셋째, 문자의 구성원리에 입각하여 자획등과 같은 문자를 구성하는 기본요소와 그들간의 연관성을 추출하여 문자를 인식하는 구조적방법이 있다.
넷째, 인간의 신경망조직을 모델로 하여 많은 수의 단순한 프로세서들을 망으로 연결한 시스템을 사용해서 패턴인식을 하는 인공신경망을 이용하는 방법이 있다.

4) 후처리

어절 내 문자의 연결관계에 관한 정보만을 사전에 저장하고 그것을 이용하여 어절단위의 오인식 수정을 하는 방법과 인식된 어절을 찾아내 사용자가 대화식으로 고칠 수 있게 하는 문자인식의 후처리방법이 있다. 한글문서인식 시스템에 대한 오인식 수정알고리즘은 1988년 처음 개발되었는데 기존의 영문 오인식 수정알고리즘들을 한글 문장의 띄어쓰기 규칙에 맞게 혼합한 방식을 사용하였다.

3. 전문데이터베이스의 효율성 측정

본 연구는 이미지화일로 구성된 전문데이터베이스와 텍스트화일로 구성된 전문데이터베이스의 검색효율성을 측정하고 비교하기 위해 다음과 같은 두가지의 실험환경을 구성한다.

3.1 텍스트화일로 구성된 전문데이터베이스

- 「도서관」 잡지 2년분의 본문 전체를 워드프로세서를 사용해 텍스트화일로 만든다.
- 워드로 작성한 텍스트화일을 RTF (Rich Formated Text)형식으로 변환한다.
- Microsoft MediaView Authoring Tool의 한글 Compiler 프로그램을 사용해 전문데이터베이스 검색이 가능한 화일로 변환한다.
- 특수문자나 도표는 스캐닝해서 이미지화일로 만들어 MediaView 프로그램을 사용해 하이퍼텍스트 방식으로 텍스트화일과 연결한다.
- 전체본문에서 불용어를 제외한 2자 이상 의 모든 명사를 키워드화한다.
- 키워드의 불리안연산과 인접연산이 가능

3.2 이미지화일로 구성된 전문데이터베이스

- 기사의 제목과 목차는 워드프로세서를 사용해 텍스트화일로 작성한다.
- 워드로 작성한 텍스트화일을 RTF형식으로 변환한다.
- 본문 전체를 스캐닝해서 BMP형식의 이미지화일로 만든다.
- Microsoft MediaView Authoring Tool의 한글 Compiler 프로그램을 사용해 기사의 제목과 목차에 대해서 전문데이터베이스 검색이 가능한 화일로 변환한다.
- 기사의 목차항목마다 해당되는 이미지화일을 MediaView 프로그램을 사용해 하이퍼텍스트 방식으로 연결한다.
- 기사의 제목과 목차에서 불용어를 제외한 2자 이상의 모든 명사를 키워드화한다.
- 키워드의 불리안연산과 인접연산이 가능

3.3 검색효율성 비교

다섯개의 질문식을 만들어 질문식에 포함된 탐색어로 검색해서 두 데이터베이스의 재현율과 정도율을 측정하고 비교한다.

두개의 독립적인 표본에 사용하는 검증법으로서 두 독립표본에서의 순위합을 이용하는 Wilcoxon의 두 표본 순위검정법을 사용해서 측정치를 검증한다.

4. 예상되는 결론

이미지화일과 텍스트화일의 검색효율성을 측정하고 비교함으로써 두 시스템의 유용성이 검증될 것이다.

참고문헌

강한배, 문자인식에 의한 장서관리시스템. 숭실대(석사), 1990.

남영준, 전문데이터베이스의 검색효율성 분석. 중앙대(석사), 1989.

노정순, "본문데이터베이스 연구에 관한 고찰과 그 전망", 도서관학 1989,v.17, 49-83.

오민경, SGML을 이용한 문헌의 구조화 및 텍스트검색에 관한 연구. 연대(석사), 1995.

최원태, "전자도서관에 관한 연구", 도서관 1995.겨울, v.50,no.4, 94-124.