서적에서의 자동색인

조성래*, 황도삼*, 최기선**
*영남대학교 컴퓨터공학과, **한국과학기술원 전산학과

Automatic Production of Book Indices

Sungrae Cho*, Dosam Hwang*, Keysun Choi**
*Dept. of Computer Engineering, Yeungnam University, **KAIST

요 약

현재의 자동색인 시스템들은 주로 정보검색을 하기 위한 목적으로 개발되고 있으며 단일 서적(저술 분야)색인에 관한 연구는 아직 미진한 상태이다. 또한 워드프로세서의 발달로 인해 단일서적에서 다양한 문서 특징이 나타나게 되었다. 본 논문에서는 관련 서적들의 색인어를 이용한 유사도 기반의 방법과 단일 서적의 문서 특징을 이용한 자동색인 기법을 제안한다. 제안한 기법의 평가를 위해 이미 출판된 서적을 대상으로 한 자동색인 시스템을 개발하고 자동추출 색인어와 수작업 색인어를 비교하였다. 관련 서적내의 색인어와 새로운 대상 문서간의 유사도 비교를 통해 색인어를 추출함으로써 통계적 빈도에 의존하는 색인 기법에서 발생하는 색인어 오추출과 과다한 색인어 추출을 줄일 수 있었다.

I. 서론

자연언어 처리의 응용분야인 자동색인은 정보화 물결 속에서 급속한 문헌의 증가에 의해 많은 연구가 진행되어 왔으며, 현재 여러 응용시스템이 개발되어 있다. 국내의 자동색인에 관한 연구는 1980년대부터 시작되었으며, 초기에는 도서관의 문헌 검색을 위한 시스템 개발이 주류를 이루었으나, 인터넷의 출현으로 정보검색용 자동색인 시스템이 개발되고 있다.

단일 서적(저술 분야) 자동색인을 위한 연구는 1983년 김영환의 연구[2]가 있었으나, 이 연구에서는 색인어로 선택되는 단어의 특징을 분석한 후 몇 가지 규칙을 세워 단어나 구를 색인어로 선정한다. 그리고 조사 사전만을 이용해서 어휘분석을 하게 되며 언어처리의 기본 단계인 형태소 해석을 하지 않는다. 그러므로 잘못된 조사분석으로 인한 명사 오추출로 부적합한 색인어를 추출하게 된다. 또한, 현재 저술분야의 문서는 워드프로세서의 발전으로 인해 다양한 문서 특징이 나타나게 되며 이러한 특성을 반영할 수 있는 자동색인 시스템이 요구되게 되었다.

◎ 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

본 논문에서는 이러한 문제점을 해결하기 위해 단일 서적 문서의 특징을 활용하고 관련 서적의 색인어와의 유사도(Similarity)에 기반한 자동색인 시스템을 제안한다. 본 시스템은 기본적으로 관련 서적의 색인어에서 추출한 단어를 기반으로 대상 문서를 색인 하는 방법을 사용한다. 그리고, 기존 색인어와 대상 문서내의 후보어의 비교 과정에서 간단한 외래어 처리 및 시소러스를 이용한 단어 확장을 하게 된다. 또한, 단일 서적 문서에서 빈번히 나타나는 문서 특징인 폰트정보, 외국어 병기정보, 특정어구 사용정보, 문장의 위치정보를 추가로 이용한다.

II장에서는 단일 서적 색인을 위한 방법론에 대해서 설명하고, III장에서는 시스템 설계 및 구현에 대해서 기술하고, IV장에서 실험 결과 및 고찰을 하고, V장에서는 결론 및 향후 계획에 대해서 기술하였다.

II. 단일서적 색인 방법론

2.1 유사도 기반 색인

본 시스템은 서적의 색인을 위한 자료로 관련 서적의 색인어를 이용한다. 대체로 전문 서적의 경우 문서의 주제가 동일한 경우 내용 면에서 유사한 부분이 많을 뿐 아니라 색인어에 있어서도 유사한 단어가 주로 사용되기

때문이다. 먼저, 관련 서적의 색인어를 형태소 해석 후, 품사패턴 및 단어를 추출한다. 그리고 주어진 문장에서 관련 색인어 품사패턴과 일치하는 구문을 색인어 후보로 선정한 후 해당 구문내의 단어들과 관련 색인어의 단어들을 비교함으로써 색인어를 선정하게 된다.

2.2 단일서적내의 문서 특징

워드프로세서로 작성된 전문 분야의 단일 서적(저술 분야)의 경우 문서 내에서 다양한 특징을 가지게 된다. 그 중에서 자동색인에 이용될 수 있는 정보는 다음과 같다.

첫째, 저자는 문서에서 중요한 의미를 지니는 단어에 대해 특정한 폰트(강조, 기울임 등)로 강조를 하게 된다.

둘째, 전문 서적에서는 독자의 용어 이해를 돕기 위해서 해당 용어에 대한 원어(영어 혹은 한자)를 함께 기재하는 경우가 많다.

셋째, 일반적으로 전문 서적의 경우, 문서가 체계적으로 구조화가 되어 있다. 예를 들면, "Chapter", "Section", "Subsection", "Paragraph"등으로 구분이 되어 있으며 대체로 이러한 구분의 앞에는 제목이 명기되어 있다. 여기에 기재된 단어는 문서 내에서 중요한 의미를 가지고 있으며 색인어를 포함하는 경우가 많다. 이러한 정보는 색인어로서 효용성이 높기 때문에 추출할 때에 이용할 수 있다. 또한, 용어를 설명하기 위한 어구(~이란, ~라(고) 하다 등)와 결합하는 명사구를 색인 추출에 이용한다.

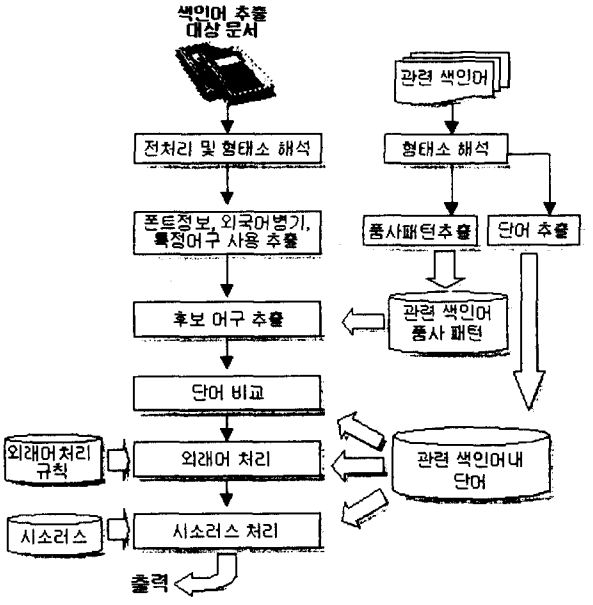
2.3 외래어 및 시소러스 처리

유사도 기반 색인에서 기존 색인어와 대상 단어를 비교할 때, 외래어 표기의 문제가 고려되어야 한다. 저자에 따라서 같은 영어 단어를 상이하게 표기하는 경우가 있다. 예를 들면, "불리안", "불리언"과 "마스터", "마스타" 등은 각각 "boolean"과 "master"에 대한 상이한 표기이다. 본 연구에서는 코퍼스를 통해 동일한 단어에 대해 다른 표기를 찾아 낸 후, 몇 개의 규칙을 도출하여 이를 외래어 처리를 위한 방법으로 사용했다.

다음으로 시소러스를 이용하여 의미적으로 유사한 단어에 대해서도 처리를 한다. 추출 대상 단어의 상위 개념의 단어가 기존의 색인어 목록에 들어 있으면 이 단어도 색인어로 될 가능성이 높다. 예를 들면, 주어진 단어가 "기호논리학"일 때 기존의 색인어 목록에 상위개념의 단어인 "논리학"이 있으면 "기호논리학"을 색인어로 추출한다.

III. 시스템 설계 및 구현

본 논문에서 제안하는 시스템 구조는 [그림 1]과 같다. 먼저, 전처리 및 형태소 해석 단계를 거친 후, 폰트 정보, 외국어 병기 정보, 특정 어구 등을 이용하여 색인어를 추출한다. 그리고 관련 서적내의 색인어들을 형태소 분석한 후 생성한 관련 색인어 품사패턴을 이용하여 대상 문장 내에서 동일한 품사패턴의 명사구를 후보로 추출한다. 다음으로, 관련 색인어의 단위명사와 후보 명사구의 단어를 비교함으로써 색인어를 선정하게 된다.



[그림 1] 시스템 구성도

3.1 관련 색인어 품사패턴 및 단어 추출

자동색인 대상 서적과 같은 분야의 서적의 관련 색인어들을 수집하여 이를 형태소 해석한 후, 관련 색인어에 출현하는 품사 패턴을 추출하고, 단어들을 단위명사 단위로 분리하여 각 단위명사별로 관련 색인어의 최소 결합 단어수를 부가하였다. 예를 들면, 관련 색인어에 "자연언어 처리"와 "자연 법칙"이 있다고 가정하면 "자연"의 최소 결합 단어수는 2가된다.

이 값은 단어 비교 단계에서 관련 색인어 품사패턴의 단어수와 비교되며, 값이 작거나 같으면 해당 단어가 일치하는 것으로 처리한다. 예를 들면, "자연"이란 단어와 최소 결합 단어수 2가 관련 단어 목록에 저장되어 있을 때, 기존 색인어 품사 패턴 "Noun+Noun Noun"에 의해 "자연언어 처리"의 명사구가 후보로 추출되었다고 가정한다. 단어 비교 과정에서 "자연"이라는 단어의 최소 결합 단어수가 관련 색인어 품사패턴의 단어수인 3보다 작거나 같으므로 "자연"은 일치하는 것으로 처리한다. 만일, "자연"의 최소 결합 단어수가 4인 경우에는 일치하지 않는 것으로 처리한다. "자연"이라는 단어는 최소 4개이상의 단어패턴에서 색인어로서 가치가 있기 때문이다.

3.2 외래어 및 시소러스 처리

명사구 구성 단어의 비교 과정에서 일치하지 않을 경우에 두 가지 방법으로 처리한다. 첫째, 단어가 외래어(영어)일 경우 외래어 표기상의 차이로 인한 불일치로 보고 외래어 처리 규칙에 의해 단어를 변환해서 비교를 시도한다. 외래어 처리 규칙은 KT-SET[7]에서 동일한 영어에 대한 상이한 외래어 표기들을 이용하여 추출하였으며 <표 1>과 같다.

<표 1> 외래어 처리 규칙

형태	규칙	예
모음 변화	'ㅏ' ⇔ 'ㅣ' 'ㅣ' ⇔ 'ㅗ'	마스타 ⇔ 마스터 컴팩트 ⇔ 콤팩트
자음 변화	'ㅍ' ⇔ 'ㅎ' 'ㅈ' ⇔ 'ㄷ'	파일 ⇔ 화일 스케쥴링⇔스케듈러
음절 축약	중성으로 끝나는 음절+'트' ⇔ 중성+종성'ㅅ' 중성으로 끝나는 음절+'크' ⇔ 중성+종성'ㄱ'	네트 ⇔ 넷 스파크 ⇔ 스팍
음절 확장	'ㅗ' ⇔ 'ㅗ'+오 'ㅓ' ⇔ 'ㅓ'+어	소스 ⇔ 소오스 서비스 ⇔ 서어비스

둘째, 시소러스를 통해 주어진 단어의 상위어를 추출한 후, 그 단어가 색인어의 목록에 있는가를 비교하게 된다. 상위개념의 단어가 색인어로 존재할 경우 해당 단어는 일치하는 것으로 처리한다.

IV. 실험 결과 및 고찰

본 시스템의 성능평가를 위해 이미 출판되어 있는 "자연언어처리(홍릉과학출판사)"[8]를 대상으로 실험을 하였다. 총 8,576어절을 실험 대상으로 하였으며, 관련 색인어를 추출하기 위해 같은 분야의 서적인 "자연언어처리(교학사)", "자연언어처리입문(대광서림)"에서 색인어 696개를 추출하여 형태소 해석을 한 후 품사패턴 50개와 단위명사 571개를 분리 저장하였다.

<표 2>는 구현된 시스템에 의해 자동으로 추출한 색인어와 수작업으로 추출한 서적에 실려 있는 색인어를 비교한 결과이다. 여기서 자동색인어수는 시스템이 추출한 색인어의 수이며, 적합색인어수는 자동 추출된 결과에서 서적에 실려 있는 색인어와 일치하는 색인어의 수이다.

<표 2> 자동 추출 색인어와 수작업 색인어 비교

입력어절수	자동색인어수	수작업색인어수	적합색인어수
8,576	172	134	117

실험 결과는 색인어의 재현율(적합색인어수/수작업색인어수)이 87.3%, 정확율(적합색인어수/자동색인어수)이 68%로 나타났다. <표 3>은 적합색인어의 추출 정보에 따른 분류이다. 하나의 색인어가 두개 이상의 정보를 가질 수 있으므로 합계가 117보다 크다. 자동 추출된 색인어 중 적합색인어가 아닌 것은 형태소 해석의 오류와, 관련 색인어의 부족 등으로 분석되었다.

<표 3> 적합색인어의 분류

폰트정보	외국어병기	특정어구	관련색인어
59	41	32	64

V. 결론 및 향후 계획

본 연구에서는 관련 서적의 색인어를 이용해서 단일 서적의 색인어를 자동 추출하는 시스템을 제안하고 개발하였다. 색인어 추출 과정에서 외래어 표기상의 불일치를 몇 가지 규칙을 도출한 후 이를 통해 해결하고, 시소러스를 이용해서 비교 대상을 확장하는 방법을 제안하였다. 또한, 단일 서적 문서의 특징인 폰트 정보, 외국어 병기 정보를 색인 추출에 이용하였으며 실험 결과 87.3%의 재현율과 68%의 정확율을 나타냈다. 시스템의 색인어 오추출 원인은 형태소 해석의 오류, 관련색인어내의 단어 부족 등으로 분석되었다.

관련 서적을 이용한 색인기법은 관련 색인어 자료가 체계적으로 구축이 되어 있어야만 좋은 성능을 낼 수 있으므로, 전문서적을 주제별로 분류하고 색인자료를 구축하는 일이 중요하다. 향후, 형태소 해석 오류를 보정하고 관련 색인어 자료의 자동생성에 관한 연구를 진행할 계획이다.

참고문헌

[1] Keysun Choi, Yung S. Han, "Syntactic Analysis Based Automatic Indexing for Korean Texts" Proc. of the Korean-US Bilateral Work-shop on Computers, Artificial Intelligence and Cognitive Science, pp.199-206, 1991.

[2] 김영환, "한글 한자 혼용문의 자동색인 시스템", 한국과학기술원 석사학위 논문, 1983.

[3] 정진성, "단일문서내에서의 언어 및 통계정보를 이용한 자동색인", 한국과학기술원 석사학위논문, 1992.

[4] 유춘식, 우선미, 유철중, 이종득, 권오봉, 김용성, "자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동색인 시스템에 관한 연구", 한국정보처리학회지 논문지, 제 5권, 제 6호, pp.1552-1561, 1998.

[5] 임형묵, 정상철, 신동욱, 김형근, 최기선, "시소러스를 기반으로 하는 자동색인 시스템에 관한 연구", 한국정보과학회 봄 학술발표논문집, 제 21권, 제 1호, pp. 173-176, 1994.

[6] 정길순, "정보검색을 위한 외래어 자동 추출 및 영어 단어로의 자동 음역", 충남대학교 석사학위논문, 1998.

[7] 김성혁 외, "자동색인기 성능 시험을 위한 Test Set 개발", 정보관리학회지 제 11권 1호, pp.929-932, 1994.

[8] 황도삼 외, "자연언어처리", 홍릉과학출판사, 1998.

[9] 정영미, 정보검색론, 구미무역(주), 1993.