지능형 정보검색을 위한 자동색인 기법

강 승 식

한성대학교 정보전산학부
136-792 서울특별시 성북구 삼선동2가 389
sskang@hansung.ac.kr

Automatic Indexing Techniques for Intelligent Information Retrieval

Kang, Seung-Shik
School of Information and Computer Engineering
Hansung University

요약

한국어 자동색인은 정보자료의 유형 및 특성에 따라 그 성능에 차이가 있으며, 검색결과에 많은 영향을 미치기도 한다. 따라서 지능형 정보검색을 위해서는 정보자료의 유형과 특성에 따라 색인 기법의 전문화 및 다양성이 요구되고 있다. 정보검색 시스템이 사용자의 요구사항에 적합한 정보자료를 제공할 수 있도록 자동색인의 관점에서 검색 성능을 향상시키기 방법으로 문서 유형에 따른 차별화된 색인 기법, 불용어 처리 기법, 색인어 관리 기법, 색인어의 유형 정보와 위치 정보를 활용하는 방법 등을 제안한다.

1. 서론

지능형 정보검색(intelligent information retrieval)은 "사용자의 요구조건에 적합한 검색결과를 제시한다"는 기본적인 원칙에 의해 질의응답(question-answering) 형식의 지식처리를 요하는 정보검색과 다양한 문서에서 질의어에 적합한 문서를 제시해 주는 유형으로 구분된다. 질의응답 형식의 지능형 정보검색은 정보자료를 지식베이스(knowledge-base)로 가공하는데 기술적인 어려움이 있다. 본 논문에서는 기존의 정보검색 시스템의 성능을 개선하여 사용자의 검색요구에 적합한 검색결과를 제시하는 두번째 유형을 중심으로 논의하고자 한다.

지능형 정보검색의 성능에 가장 큰 영향을 미치는 요소 중의 하나는 자동색인 시스템의 성능이다. 특히, 한글 문서에는 복합명사와 외래어, 고유명사 등 미등록어가 다수 포함되고 있으며 이러한 용어들은 색인어로서 가치가 매우 크다. 따라서 복합명사와 미등록어가 포함된 문서에서 색인어를 추출할 때는 다른 색인어에 비해 복합명사와 미등록어의 색인어로서 가치를 높여주는 방법이 가능하다.

자동색인 시스템의 성능은 정보자료의 유형에 따라 정확도와 재현율이 달라질 수 있다. 일반적인 정보자료는 맞춤법 오류가 거의 없고 복합명사를 띄어쓰기 때문에 색인 정확도가 높은 편이다. 그러나 신문기사에서는 '-ㄹ수있다', '-ㄹ것이다', '-ㄹ때' 등 1음절 어휘가 포함된 어절과 복합명사를 붙여쓴 경우가 많다. 따라서 신문기사의 자동색인에서는 색인어가 누락되거나 불용어가 포함될 가능성이 높다. 그 이유는 자동색인 시스템이 책이나 잡지, 신문기사, 보고서 등 띄어쓰기나 맞춤법 오류어가 없는 일반적인 문서들을 대상으로 하고 있기 때문이다.

최근에는 인터넷과 전자우편이 보편화됨에 따라 홈페이지 문서나 전자우편 문서 등 맞춤법 오류어와 약어, 은어가 다수 포함된 문서에 대한 정보검색에서 오류어 처리 기법이 요구되고 있다. 특히, HTML과 XML(eXtended Markup Language), SGML(Standard Generalized Markup Language) 등 구조적인 문서의 정보검색에서는 색인어가 출현된 위치에 따라 색인어의 중요도를 추정할 수 있으며 색인어의 출현 위치 및 색인어 유형 정보가 검색결과에 미치는 영향이 커지게 된다.

또한, 이미지와 음성이 포함된 멀티미디어 정보자료에 대한 자동색인 요구가 증가하고 있다. 동영상과 정지영상의 검색에서는 색상이나 움직임 등 일반적인 정보자료에서 불용어로 간주되는 수식어와 술어가 검색어로 사용되는 특징이 있다. 특히, 음성 자료는 문어체가 아니라 구어체이므로 자동색인 시스템이 구어체를 인식할 수 있도록 그 기능이 확장되어야 한다. 이와 같이 멀티미디어 정보검색에서는 검색 기능뿐만 아니라 색인어를 추출하고 관리하는 방법에서 기존의 정보검색 시스템과는 많은 차이가 있다.

본 논문에서는 지능형 정보검색 시스템에서 사용자의 요구사항에 가장 적합한 검색결과를 제시하는데 요구되는 한글 문서의 자동색인과 관련된 제반 문제점 및 정보자료 유형에 따른 색인어 처리기법을 고찰하고, 멀티미디어 정보자료 및 구조화된 문서의 색인 문제, 띄어쓰기 오류어가 포함된 문서의 색인 등 지능형 정보검색을 위한 자동색인 기법들을 제안한다.

2. 정보자료의 유형 및 특성

일반적으로 정보자료의 유형은 수치정보, 사실정보, 문서정보, 서지정보, 그림정보 등으로 분류된다[1]. 그

런데 자동색인의 관점에서 정보자료들은 그 유형 및 특성에 따라 색인어 추출결과에 차이가 있다. 자동색인 기법을 적용하는데 차별화가 요구되는 문서의 유형 및 특성은 매우 다양하지만 대표적인 문서 유형들을 살펴보면 다음과 같다.

2.1 일반 문서

일반 문서는 주로 국어사전에 수록된 어휘들로 구성된 문서이다. 교과서를 비롯하여 일반 서적 등 맞춤법 오류가 거의 없고, 복합명사의 경우 띄어쓰기 원칙을 따르는 문서들이 이에 속한다. 고유명사를 제외하면 사전 미등록어가 드물기 때문에 색인어로 추출되는 명사들의 대부분이 사전에 수록된 용어이다.

이 유형의 문서는 미등록어가 드물게 사용되기 때문에 형태소 분석이나 자동색인 정확도가 높다. 반면에 색인어로 추출된 명사들에 대해 색인어의 중요도를 차별화하기가 쉽지 않다.

2.2 전문분야 문서

특정 분야의 논문이나 전공서적 등의 문서에서 전문용어는 색인어로서 가치가 매우 크지만, 보편적인 어휘들의 중요도는 매우 낮다. 즉, 추출된 색인어의 유형이 복합명사 형태의 전문용어, 단일명사 형태의 전문용어(주로 미등록어), 보편적인 용어 등으로 명확하게 구분된다.

문서에 따라 복합명사를 띄어쓰기도 하고 붙여쓰기도 하기 때문에 복합명사 분해 문제가 자동색인 기능에 미치는 역할이 크다. 또한, 문서 제목과 항목별 제목(장 제목, 절 제목 등)에 출현하는 용어의 중요도가 높다. 특히, 논문의 경우에는 요약과 결론에서 추출된 용어의 중요도가 높아지는 특징이 있다.

컴퓨터 분야, 의학 분야 등 전문분야에 대한 정보검색 시스템은 색인어휘집(전문용어 사전)에 속한 용어들만 추출하더라도 검색 성능이 어느 정도 보장된다. 그러나 각 분야마다 색인어휘집을 구축하기가 쉽지 않을 뿐만 아니라 새로운 용어(신조어)들이 계속해서 추가되기 때문에 색인어휘집을 관리하기가 어렵다. 따라서 분야별 색인어휘집을 구축하는 것보다 범용 자동색인 시스템의 복합명사 및 미등록어 추정에 의해 전문용어를 인식하는 것이 효율적이다.

전문분야 문서 중에서 법률문서는 맞춤법의 띄어쓰기 원칙을 무시하고 독자적인 띄어쓰기 규칙에 따라 문서를 작성하는데 그 예는 아래와 같다.

> 벤처기업육성에관한특별조치법시행령안
>
> 제1조(목적) 이 영은 벤처기업육성에관한특별조치법에서 위임된 사항과 그 시행에 관하여 필요한 사항을 규정함을 목적으로 한다.
>
> 제2조(벤처기업의 범위) ①벤처기업육성에관한특별조치법(이하 "법"이라 한다) 제2조제1항제1호의 규정에 의한 벤처기업은 ... (이하 생략)

법률 관련 문서에서 법률 이름은 붙여쓰기를 하고 있으며, 복합명사 또한 붙여쓰는 것이 일반적이다. 그런데 예제문서의 밑줄친 부분과 같이 법률 이름은 논문이나 문서 제목과 같이 수식어가 포함된 명사구이므로 술어와 문법형태소가 포함되어 있다. 따라서 법률 문서에 대한 자동색인은 미등록어 추정이나 복합명사 분해 기능만으로 색인어를 추출하는 것이 불가능하다. 이러한 문제는 법률뿐만 아니라 법률을 심의하는 국회, 지방의회 등 자치단체 위원회의 회의록 검색 시스템에도 동일하게 적용된다.

2.3 구조화된 문서

HTML로 작성되는 인터넷 문서 등 전자문서를 XML이나 SGML을 표준 문서 형식으로 채택하게 됨으로써 구조화된 문서(structured document)에 대한 새로운 유형의 검색 요구가 제기되고 있다. 구조화된 문서는 구성요소별로 태그가 부여되므로 동일한 색인어라 하더라도 문서내 출현 위치와 중요도 등 색인어 관련 정보를 파악하기가 용이하다. 구조화된 문서에서는 사용자의 검색 요구에 대해 문서의 구성요소별 검색이 가능하도록 구조화된 자동색인 기법(structured automatic indexing method)과 구조화된 색인어 저장 시스템(structured indexing system)이 필요하다.

비구조화된 정보자료에서도 제목, 본문, 작성자, 작성 일자 등의 형식을 갖춘 경우가 있으며, 이 경우에 정보자료에서 추출된 색인어는 어떤 항목에서 추출되었는지에 따라 중요도가 달라진다. 도서정보 검색에서는 제목과 저자명, 출판사 등 참조정보 검색 기능만으로 대부분의 검색요구를 충족시킬 수 있다. 그러나 학술논문 검색 시스템은 사용자들이 저자명이나 논문제목에 사용된 용어를 모른 상태에서 주제어로 검색하는 것이 일반적이다. 따라서 학술논문 검색 시스템은 각 논문마다 주제어를 부여함으로써 주제어 검색 방법을 이용하고 있다.

2.4 대화체 문서와 PC통신 오류어

자동색인의 대상이 되는 정보자료는 대부분 문어체 자료이고, 자동색인 시스템은 문어체 자료에서 색인어를 추출하는 기능을 중심으로 구현되어 있다. 그런데 소설이나 시나리오 등 정보자료의 특성에 따라 구어체 어휘들이 사용되기도 한다. 특히, 방송자료와 같이 드라마 대본에 사용되는 구어체 어휘들은 문어체 정보자료와는 특성이 많이 다르기 때문에 문어체 위주의 자동색인 시스템을 이용했을 때 색인어 추출 오류가 빈번하게 발생한다.

대화체 문서는 구어체 문법형태소뿐만 아니라 축약현상이 빈번하게 발생한다. 대화체에서 주로 사용되는 문법형태소는 조사/어미 사전을 보완함으로써 해결할 수도 있다. 그러나 어휘형태소와 문법형태소의 경계에서 발생하는 축약 현상은 대부분 각 어휘마다 별도의 처리과정이 필요하기 때문에 구어체 어휘들을 수집하고 이로부터 색인어를 추출해야 하는 문제가 있다.

전화나 회의 등 실생활(real world)에서 녹음된 음성을 기록한 대화체 음성언어에서는 '에', '음'과 같은

간투어가 삽입되기도 하고, 표준어가 아닌 어휘들을 사용하는 등 맞춤법 규정을 무시하는 경향이 많기 때문에 대화체 문서의 자동색인은 많은 문제를 야기하고 있다.

대화체 음성언어의 예
만나구요(만나고요), 같애(같아), 갈려고(가려고),
준비허고(준비하고), 그거는(그것은), 그러믄(그러면),
힘들더래도(힘들더라도), 어떠까요(어떨까요),
맨날(매일), 날으는(나는), 땜에(때문에)

인터넷이나 PC통신에서 사용되고 있는 어휘(은어, 축약어 등)들도 대화체 어휘와 특성이 비슷하며 범용 자동색인 시스템에서 이러한 어휘들은 고려되지 않고 있다. 대화체 어휘에 비해 인터넷-PC통신 어휘는 띄어쓰기 오류가 매우 심각하기 때문에 별도의 처리과정이 필요하다. 따라서 이러한 유형의 문서들에 대해서는 색인어 추출 오류가 매우 많이 발생한다.

인터넷 뉴스그룹 어휘의 예
갈켜줘(가르쳐 줘), 가르쳐줘잉, 그때가서,
그동안모아둔, 그외다수, 그중한사람,
글코보면(그러고 보면)

현재 정보검색 시스템에서 제공하고 있는 자료들은 서적이나 문서 등 구어체 어휘가 드물게 사용되는 자료들이며, 대화체 어휘 및 인터넷-PC통신 오류어에 대한 색인 문제가 무시되어 왔다. 그러나 방송이나 연극 대본과 같이 구어체 어휘들이 빈번하게 사용되는 정보자료에 대한 정보검색 시스템은 구어체 어휘의 처리 방법에 따라 그 성능에 차이가 많다.

2.5 멀티미디어 문서

일반적인 정보검색 시스템에서 형용사와 동사는 '품사 불용어'로 간주하여 색인어로 추출되지 않는다. 그러나 디자인이나 예술과 관련된 정보자료에서는 색상에 관한 형용사가 질의어로 사용된다. 이 경우에 '빨강', '노랑' 등 명사 질의어보다는 '노란 꽃'과 같이 형용사 질의어를 사용하는 경우가 더 많다. 이와 같이 형용사를 질의어로 사용하려면 형용사를 색인어로 추출해야 한다.

최근에는 방송자료를 비롯하여 정지영상과 동영상이 포함된 멀티미디어 정보자료에 대한 정보검색 요구가 증가하고 있다. 멀티미디어 정보자료의 특성을 기술하는 정보자료에 대한 질의어는 형용사뿐만 아니라 동사나 관형사 등 일반적으로 품사 불용어로 간주되는 용어가 사용되기도 한다. 이러한 경우는 주로 멀티미디어 정보자료의 특성에 관한 정보검색에서 발생하지만 예1), 예2)와 같이 일반적인 정보검색에서 필요한 경우도 있다.

예1) "빨간 장미"
예2) "새빨간 거짓말"
예3) "노랗게 물들인 머리"
예4) "해가 떠오르는 장면"
예5) "해가 반쯤 떠 있는 장면"
예6) "해가 하늘 높이 떠 있는 장면"

또한, 예8)~예10)처럼 때로는 부사에 의하여 검색되는 정보자료가 구분되어야 하는 경우도 있다.

예7) "주인공이 뛰어가는 장면"
예8) "주인공이 빨리 뛰어가는 장면"
예9) "주인공이 매우 빨리 뛰어가는 장면"
예10) "주인공이 조금 뛰어가다가 멈추는 장면"

때로는 '안녕!'이나 '야호!'에 관한 자료를 검색할 때와 같이 감탄사를 질의어로 사용될 수도 있기 때문에 동사와 형용사, 부사뿐만 아니라 관형사와 감탄사 등 모든 어휘들을 색인어로 추출해야 한다. 다만, 색인어의 유형에 따라 중요도가 다르기 때문에 색인어 저장 시스템에서는 색인어 유형에 대한 정보가 함께 저장되어야 한다.

3. 자동색인 방법론과 문제점

영어를 비롯한 굴절어(inflectional language)에서는 문서로부터 색인어(index term)를 추출하기 위하여 명사의 복수형이나 동사의 3인칭 단수, 형용사의 비교급, 접두사와 접미사 분리 등 어간추출 알고리즘(stemming algorithm)을 이용한다. 이와 더불어 복합어를 인식하고 색인하기 위하여 구문분석 기법을 활용하기도 한다.

굴절어는 하나의 형태소가 하나의 단어를 이루기 때문에 자동색인 방법이 비교적 단순한 편이다. 이에 비해 한국어는 교착어(agglutinative language)로서 한 어절(word phrase)이 여러 개의 형태소로 구성되기 때문에 어간이나 형태소들을 인식하고 색인어를 추출하는 방법이 굴절어와는 다르며, 굴절어에서 사용되는 방법을 그대로 적용하기가 어렵다. 한글 문서에서 색인어를 추출하는 자동색인 기법은 아래와 같이 크게 2가지 유형으로 구분된다[2].

(1) 색인어휘집, 기능어휘집을 이용하는 방법
(2) 형태소 분석, 구문 분석을 이용하는 방법

색인어휘집이나 기능어휘집을 이용하는 방법은 아래와 같이 세 가지 유형으로 구분된다.

- 색인어휘집을 이용한 자동색인
- 기능어휘집을 이용한 자동색인
- 색인어휘집과 기능어휘집을 이용한 자동색인

색인어휘집(controlled vocabulary)을 이용하는 방법은 구현이 용이할 뿐만 아니라 색인어 집합이 결정된 폐쇄적인 문서집합에 매우 적합하다. 그러나 여러 가지 유형의 문서집합 혹은 색인어 집합이 가변적인 경

우에 색인어휘집을 유지-관리하는데 많은 문제가 있다. 이러한 단점을 보완하기 위해 두 가지 어휘집을 함께 사용하기도 하지만 근본적인 문제점은 해결되지 않는다.

한국어 정보처리 기법을 이용한 자동색인은 형태소 분석과 구문 분석 기법의 활용 정도에 따라 세 가지 유형으로 구분된다.

- 부분적인 형태소 분석 기법을 이용한 자동색인
- 형태소 분석기를 이용한 자동색인
- 형태소 분석 및 구문분석을 이용한 자동색인

'부분적인 형태소 분석' 기법을 이용한 자동색인은 문서로부터 명사들만 추출하는 시스템을 개발하는 것으로 체언 이외의 용언, 독립언 등은 불용어로 간주한다. 따라서 품사 불용어의 인식, 미등록어 인식, 형태론적 모호성 해결 등이 필수적이다. 이러한 기능들이 보완된 자동색인은 결국 완전한 형태소 분석기의 기능과 동일하다. 따라서 '부분적인 형태소 분석'을 이용하는 방법은 자동색인의 성능이 낮아도 문제가 되지 않는 제한적인 용도로 사용된다.

형태소 분석 기법을 이용하는 자동색인은 현재 가장 보편적으로 사용되고 있는 방법이다. 이 기법에서는 형태소 분석기의 복합명사 처리 기능과 미등록어 인식 기능이 자동색인의 성능에 미치는 영향이 매우 크다. 복합명사와 미등록어가 색인어로서 가치가 매우 높기 때문이다. 그런데 복합명사와 미등록어가 포함된 어절에서 문법형태소를 분리할 때 모호성이 발생하는 문제가 있기 때문이다.

'보고서', '참고서'와 같이 용언과 체언이라는 두 가지 분석결과('보-'$_{동사}$+'-고서'$_{어미}$, '보고서'$_{명사}$)로 분석이 가능한 어절에서 명사가 미등록어이면 이를 추출하지 못할 수도 있다. 또한, '홍부가', '제주도', '컴퓨터공학과', '교수회의' 등 '가/과/도/의' 등 조사 음절로 끝나는 미등록어나 복합명사에서 "색인어 추출 모호성"이 발생하여 색인어 추출 오류가 발생한다.

'홍부가' ① '홍부'$_{명사}$ + '가'$_{조사}$
② '홍부가'$_{명사}$
'제주도' ① '제주'$_{명사}$ + '도'$_{조사}$
② '제주도'$_{명사}$

색인어 추출 모호성으로 인한 색인어 추출 오류에 의해 재현율이 낮아지는 문제를 예방하기 위하여 조사가 결합된 유형도 색인어로 추출할 수도 있다. 그러나 이 경우에는 모든 미등록어, 복합명사에 대해 조사가 결합된 유형이 색인어로 추출되는 문제가 있다.

형태소 분석기에 형태소 태깅 시스템을 결합하여 색인어를 추출하는 방법도 있으나 태깅 과정에서 색인어가 누락되면 재현율이 낮아지는 문제가 있다. 또한, 구문분석 기법을 이용하는 경우에도 구문분석기의 정확도 문제, 구문론적 중의성(syntactic ambiguity) 문제, 자동색인 속도 문제 등이 해결되어야 한다.

4. 자동색인의 주요 기능

정보검색 시스템은 재현율과 정확률을 모두 만족시키기가 어려우며 일반적으로 재현율이 더 중요시된다. 불필요한 많은 문서가 검색되더라도 문서들의 우선순위를 부여할 수 있기 때문이다. 그런데 우선순위 부여 정확도가 낮은 부작용이 심각하며, 이를 극복하기 위해 자동색인의 기능이 확장-개선되어야 한다.

재현율과 정확률을 높이기 위해 자동색인 시스템에 추가되거나 보완되어야 하는 기능으로는 복합명사의 분해-결합, 주제어-비주제어 등 색인어 차별화, 자동 띄어쓰기 기법 등이 있다.

4.1 복합명사의 분해-결합

자동색인에서 문서의 내용을 대표하는 주제어는 전문용어나 신조어 등 형태소 분석사전에 수록되지 않은 미등록어가 대부분이다. 전문용어나 주제어는 복합명사인 경우가 많으며 복합명사는 띄어쓰기와 붙여쓰기가 모두 허용되기 때문에 복합명사를 색인하고 검색하는 방법이 매우 중요하다.

복합명사 색인 기법은 복합명사를 분해하거나 결합하는 방식에 따라 차이가 있으며, 분해와 결합을 모두 허용하는 것도 가능하다. 현재, 대부분의 자동색인 시스템은 "붙여쓴 복합명사를 분해"하는 기법을 취하고 있다[3,4,5]. 띄어쓴 복합명사와 구(phrase)나 절(clause)로부터 복합명사를 인식하여 결합할 때는 복합명사가 아닌 것을 복합명사로 인식하는 오류가 많기 때문이다[6].

<u>복합명사의 색인 방법</u>

① 문서에 출현한 용어를 그대로 색인
② 붙여쓴 복합명사를 분해하여 색인
③ 띄어쓴 복합명사를 결합하여 색인
④ 복합명사를 분해 또는 결합하여 색인

복합명사의 분해-결합 문제는 붙여쓰기-띄어쓰기로 인한 복합명사의 표준화(normalization) 문제이다. 예1), 예2)는 복합명사 분해 기법에 의해 복합명사를 구성하는 단일명사들을 추출한 것이다. 그러나 예3), 예4), 예5)는 복합명사 결합 기법을 사용하지 않았기 때문에 단일명사들만 추출되고 붙여쓴 유형은 추출되지 않는다. 따라서 질의어 '정보검색'에 의하여 예2)만 검색되고, '정보검색시스템'에 의해서는 예1)만 검색되며, '검색시스템'에 대해서는 검색되는 문서가 없다.

예1) 정보검색시스템
⇒ 정보검색시스템, 정보, 검색, 시스템
예2) 정보검색 시스템
⇒ 정보검색, 정보, 검색, 시스템
예3) 정보 검색 시스템
⇒ 정보, 검색, 시스템
예4) 정보를 검색하는 시스템
⇒ 정보, 검색, 시스템

예5) 정보의 검색이 가능한 시스템
⇒ 정보, 검색, 시스템

동일한 복합명사에 대한 색인어의 표준화 문제를 해결하기 위해서는 질의어에 대해 복합명사 분해과정을 거친 후에 검색하는 방법이 가능하다. 이 경우에는 단일명사로만 검색했을 때 붙여쓴 복합명사가 출현한 문서와 단일명사가 개별적으로 출현한 문서를 차별화해야 하는 질의어의 출현위치 문제가 발생한다.

4.2 주제어와 비주제어 색인

일반적으로 정보검색 시스템은 정보자료에 출현하는 색인어에 대해 출현빈도와 위치정보(문서내 몇 번째 어절)를 저장하지만 문서내에서 중요도를 부여하지 않고 있다. 색인어들에 대해 주제어-비주제어, 색인어 유형 정보, 출현위치(제목, 요약, 본문 등)를 색인어와 함께 저장한다면 검색된 문서들의 우선 순위를 쉽게 계산할 수 있을 것이다.

주제어-비주제어 색인 방법
① 주제어만 색인하는 방법
② 주제어-비주제어를 구별하지 않고 색인
③ 주제어-비주제어를 구별하여 색인
④ 주제어-비주제어를 유형별로 세분화

검색된 문서의 우선순위는 대용량 정보자료에서 검색된 문헌수가 많아질수록 매우 심각하다. 한국어의 경우에 색인어의 유형이 보통명사, 복합명사, 미등록어 등으로 구별되므로 이러한 특성과 구문분석 기법, 빈도수 정보를 활용하면 주제어 추출 알고리즘을 개발할 수 있다. 주제어 추출 알고리즘은 문서 유형에 따라 정확도의 차이가 커질 수 있다.

따라서 주제어만 색인하는 기법은 보편적으로 활용하기 어려우며, 추출 알고리즘의 정확도가 낮은 경우에는 보조적인 색인기법으로 활용될 수 있다. 또한, 주제어 추출 정확도가 높다 하더라도 질의어에 비주제어가 사용되는 경우가 많기 때문에 비주제어도 함께 색인하는 것이 바람직하다.

4.3 자동 띄어쓰기

자동 띄어쓰기 기능이 필요한 경우는 법률문서에서 법률 이름을 붙여쓰는 경우가 대표적이다. 이외에도 데이터베이스에 저장된 기존 자료에서 특정 항목은 공백을 제거한 상태로 저장된 경우가 있다. 이러한 자료들을 검색하려면 색인할 때 자동 띄어쓰기(automatic word spacing) 기능이 필수적이다[7,8].

자동 띄어쓰기의 필요성
① 띄어쓰기를 무시한 정보자료의 색인
② 문자인식기에서 줄바꿈 문제 해결
③ 전자출판 문서에서 줄바꿈 문제 해결
④ 연속어절 음성인식에서 띄어쓰기 문제

최근에는 문자인식에 의해 대량의 정보자료를 입력할 때 줄바꿈 위치에서 공백을 삽입해야 하는지를 판단하는 문제가 발생하고 있다. 줄바꿈 위치의 띄어쓰기 문제는 전자출판(desktop publishing) 시스템에서 편집된 자료의 정보검색이나 문서편집기로 작성된 문서에서 줄바꿈 위치에서 띄어쓰기 문제를 고려하지 않은 경우에도 발생하고 있다.

자동 띄어쓰기는 정보검색뿐만 아니라 한글을 입력할 때 띄어쓰기 문제를 자동으로 처리하거나 맞춤법 검사기에서 띄어쓰기 오류는 교정하는데 활용되기도 한다. 또한, 차세대 사용자 인터페이스로서 연속어절 음성인식 기능이 상용화되면 어절경계를 인식할 때 중요한 역할을 하게 될 것이다.

5. 불용어 처리 및 색인어 관리 기법

불용어는 "색인어로서 가치가 없는 단어"로서 an, by, of, the, to와 같이 발생빈도가 높은 단어의 대부분이 이에 속한다. 전형적인 영어문서에서는 10번 이상 출현하는 단어 중 20~30%가 불용어로 인식되고 있다. 이러한 불용어들을 제거할 경우에 색인 및 검색 속도를 향상시킬 수 있고, 색인어 저장 공간의 크기가 작아지기 때문에 검색효과를 높일 수 있다[9].

5.1 영어의 불용어 처리

Rijsbergen(1975)은 250개의 불용어 목록을 제시하고 있으며, 100만 어절 Brown corpus로부터 추출한 불용어의 개수는 425개로 그 예는 아래와 같다.

- a, about, above, across, after, against, all, ...
- b, back, backed, backing, backs, be, because, ...
- c, came, can, cannot, case, cases, certain, ...

빈도가 높은 단어 중에는 time, war, home, life, water 등 색인어로서 중요하게 사용되는 것도 있다. 또한, 컴퓨터 분야에서 computer, program, language, windows 등은 색인어로서 가치가 낮기 때문에 동일한 용어에 대해서도 분야별로 색인어의 중요도를 계산하는 방법이 달라질 수 있다.

특히, 자동색인에서 불용어로 간주하여 제거된 색인어는 검색이 되지 않기 때문에 불용서 처리는 매우 신중해야 하며, 상업적인 정보검색 시스템에서는 불용어가 거의 없다. 예를 들어, ORBIT search service는 an, and, by, from, of, the, with 등 8개만을 불용어로 간주하고 있다.

5.2 한국어의 불용어 처리

한국어에서 불용어로 간주되는 용어는 품사 불용어와 명사 불용어, 그리고 숫자 불용어 등이 있다. 일반적으로 체언을 제외한 모든 품사(조사, 어미, 동사, 형용사, 관형사, 부사, 감탄사)는 품사 불용어로 간주된다. 체언 중에서도 대명사, 수사, 1음절 명사는 색인어로서 가치가 매우 낮으므로 불용어로 간주할 수 있다. 또한, 일반적으로 '1999년', '1,200원', '제1절' 등 한글과 숫자가 혼합된 것과 'x값', 'y축' 등 한글과 숫자가 혼

합된 것도 불용어인 경우가 대부분이다.

한글, 영문자, 숫자, 문자 등이 혼합된 경우에도 'LG 전자', '비타민 A', 'C++', 'B+ 트리' 등은 색인어로서 매우 중요하다. '3.1절', '한글97', '윈도 3.1', '비주얼 베이직 6' 등 '3.1'이나 '6'과 같은 숫자도 색인어의 일부로서 가치가 매우 높은 경우가 있다.

수사 어절은 아라비아 숫자와 한글의 혼용 및 띄어쓰기에 따라 "230000원", "230,000원", "이십삼만원", "이십 삼만원", "2십3만원", "2십 3만원" 등과 같이 다양하게 표현이 가능하다. 따라서 수사 어절은 표준화 기법에 의해 동일한 표현으로 정규화함으로써 불용어로 처리할 수 있다.

품사 불용어로 간주되는 용어들도 문서 유형에 따라 혹은 특정 문서에서는 색인어로서 매우 중요한 경우도 발생하기 때문에 형태소 분석결과만으로 불용어를 제거하는 방법은 바람직하지 않다. 이는 형태소 분석결과에 의한 불용어 제거 기법보다는 구문분석이나 복합명사 결합 기법에 의해 색인어의 중요도를 계산하는 방법에 의해 해결되어야 할 문제이다.

5.3 색인어 관리 기법

색인 시스템을 세분하면 형태소 분석 등을 이용하여 색인어를 추출하는 '색인어 추출'과, 색인어를 저장하는 '색인어 저장', 그리고 저장된 색인어를 관리하는 '색인어 관리'로 구분된다. 이 중에서 상대적으로 색인어 관리 문제는 소홀하게 다루어져 왔다.

그러나 검색 정확도와 검색 속도를 개선하기 위해 색인어를 저장하고 관리하는 기법이 정보검색 시스템의 성능에 미치는 영향이 커지고 있다. 즉, 색인어 저장과 색인어 관리 시스템을 잘 활용함으로써 자동색인의 효율 및 검색 효율을 높일 수 있다.

자동색인에서는 맞춤법 오류어, 미등록어 추정 오류 등으로 인하여 불용어가 색인되는 경우가 많다. 이러한 자동색인의 단점을 보완하는 방안으로 자동색인 결과에 대해 색인전문가의 검증을 거치는 방법과 색인전문가가 주기적으로 색인어를 관리하는 방법이 있다.

그 예로는 일정기간 동안 전혀 검색되지 않은 색인어들을 추출하여 전문가가 확인하여 제거하는 방법이 있다. 또한, 다수의 문서에 출현한 색인어의 관리 기법으로 문헌빈도가 임계치를 넘는 색인어들을 추출하여 전문가가 확인하고 불용어로 간주되는 용어들을 제거함으로써 효율적으로 색인어를 관리할 수 있다.

색인어 관리

① 일정기간 검색되지 않은 색인어 제거
② 문헌빈도가 임계치를 넘는 색인어 제거
③ 유형별로 색인어 저장 시스템 분할

색인어 추출과정에서 주제어 선별 기법이 도입되면 검색 효율을 높이기 위하여 주제어와 비주제어를 구분하여 저장하는 기법을 활용하는 것이 가능하다. 이 기법은 정보자료의 양이 매우 많을 때 비주제어가 출현한 정보자료는 검색할 필요가 없게 되는 경우이다. 즉, 주제어를 먼저 검색하여 검색 문헌수가 임계치보다 작은 경우에만 비주제어를 검색하는 방법을 사용할 수 있다. 특히, 주제어에 비해 비주제어가 훨씬 많을 것으로 예상되기 때문에 정보검색 시스템의 유형에 따라 색인어 저장 시스템을 색인어의 유형별로 구분함으로써 색인 및 검색 효율을 향상시킬 수 있다.

6. 결론

정보자료의 양이 증가하고 정보자료의 유형이 다양화됨에 따라 적합한 문서를 검색하는데 검색된 문서의 우선순위를 결정하는 방법은 한계가 있다. 이를 극복하기 위해 자동색인의 역할이 더욱 중요시되고 있으며, 정보자료의 유형 및 특성에 따라 사용자의 검색 요구를 충족시킬 수 있도록 자동색인 기법이 보완되거나 발전 방향을 모색하였다.

구체적인 예로서 법률 문서의 띄어쓰기 문제, 구조화된 문서의 색인 문제, 멀티미디어 자료의 색인 문제, 불용어 처리 문제 등을 살펴보았다. 이러한 문제점을 해결하는 방안으로 문서 유형에 따라 자동색인 기법의 차별화, 주제어-비주제어 및 색인어 유형에 따라 색인어의 차별화 기법, 복합명사 분해-결합 기법, 불용어 처리 기법, 색인어 관리 기법 등을 제안하였다.

제안된 기법 중에서 법률분야 문서에 대한 색인 기법으로 자동 띄어쓰기 기능을 적용하여 색인어를 추출하였으며 문서 유형 및 특성에 따라 전문적인 자동색인 기법이 필수적임을 확인하였다.

참고문헌

[1] 정영미, 정보검색론, 구미무역 출판부, 1993.

[2] 강승식, 권혁일, 김동렬, "한국어 자동색인을 위한 형태소 분석의 기능", 정보과학회 춘계 학술발표 논문집, 22권, 1호, pp.929-932, 1995.

[3] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회 논문지(B), 25권, 1호, pp.172-182, 1998.

[4] 윤보현, 임희석, 임해창, "통계정보를 이용한 한국어 복합명사의 분석 방법", 한국정보과학회 봄 학술발표 논문집, pp.925-928, 1995.

[5] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.262-267, 1996.

[6] 서은경, "구문·통계적 기법을 이용한 한국어 자동색인에 관한 연구", 정보관리학회지, 10권, 1호, pp. 97-124, 1993.

[7] 심광섭, "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지(B), 23권, 9호, pp.991-1000, 1996.

[8] 강승식, "한글 문장의 자동 띄어쓰기", 제10회 한글 및 한국어 정보처리 학술발표 논문집, pp.137-142, 1998.

[9] 류근호, 김진호, 정보검색, 시그마프레스, 1995.

[10] Rijsbergen, *Information Retrieval*, Butterworths, 1975.