한글 문서의 색인어와 색인 기법†

국민대학교 강승식*

1. 서 론

정보검색 시스템의 성능을 평가하는 요소는 재현율(recall)과 정확률(precision)이고, 재현율과 정확률을 결정하는데 가장 큰 영향을 미치는 것은 문서에 대한 색인어와 색인어 가중치이다[1]. '질의어'에 적합한 문서를 검색할 수 있는지를 결정하는 것은 "적합 문서에 대해 색인이 되어 있는가?"하는 문제이며, 이는 재현율에 직접적인 영향을 미치게 된다. 즉, 적합 문서를 색인할 때 '질의어'에 대한 색인이 되어 있지 않은 문서는 검색이 되지 않으며, 또한 부적합 문서에 색인이 되어 있으면 부적합 문서들이 다수 검색되기 때문에 정확률이 낮아지게 된다.

검색된 적합 문서들을 검색 결과의 앞부분(상위에 랭크)에 제시하기 위한 방법(검색 모델)은 색인어 가중치를 기반으로 한다. 벡터 모델, 확률 모델, 언어 모델 등 검색 모델에서 질의어와 각 문서간의 유사도는 색인어와 색인어 가중치에 의해 계산되기 때문이다. ***색인어 가중치는 색인어가 각 문서에 대해 중요도를 가늠하는 척도가 되며, 색인어 가중치에 의해 질의어와 문서간의 유사도를 계산하여 검색된 문서들의 순위를 결정한다.*** 따라서 정보 자료에 대한 색인어 추출 방법과 색인어 가중치 부여 방법은 검색 엔진의 성능을 좌우하는 매우 중요한 역할을 한다.

이처럼 색인어와 색인어 가중치가 검색 엔진의 성능을 결정하는 가장 기본적인 요소임에도 불구하고 정보검색 연구 분야에서는 정보 자료의 색인 기법보다는 좀 더 정교한 검색 모델에 의해 성능 향상을 추구하는 방향으로 연구가 진행되어 왔다. 즉, 색인어와 색인 기법은 검색 모델에 비해 검색 결과에 미치는 영향이 매우 크기 때문에 검색 모델 못지 않게 매우 중요함에도 불구하고 상대적으로 검색 모델에 비해 중요시되지 않는 경향이 있다. 그 이유는 정보검색 학문 분야가 영어자료를 대상으로 발전해 왔으며, 영어 문서의 색인 기법은 한국어, 중국어, 일본어 등 타 언어들에 비해 단순하기 때문에 색인어와 색인어 가중치 기법의 차별화에 의해 성능을 향상시킬 수 있는 여지가 매우 적기 때문이라고 판단된다.

본 논문에서는 한국어 정보검색 시스템의 성능을 향상시키는데 가장 근본적인 문제인 한글 문서의 색인 기법과 관련하여 (1) 각 문서에 대해 어떤 색인어를 부착(또는 추출)할 것인지, (2) 색인어 추출 기법과 색인 방식에는 어떤 것들이 있으며, 각각의 장단점은 무엇인지에 관하여 고찰하고자 한다.

2. 색인어와 불용어 선별 기법

문서의 내용을 지시하기 위한 목적으로 문서에 부착되는 '색인어(index term)'는 검색의 관점에서는 '질의어(query term)'가 된다. 즉, 색인어로서 가치가 있는 용어는 해당 문서를 검색할 때 질의어로 사용되는 용어들이며, 질의어로 사용될 가능성이 없는 용어는 불용어가 된다. 따라서 어떤 문서에 대한 가장 이상적인 색인어 집합은 그 문서를 검색할 때 질의어로 사용될 용어들의 집합이다. 수동 색인(manual indexing)에 의한 용어 부여 색인에서는 색인 전문가가 해당 정보자료에 대하여 질의 예상 용어들을 색인어로 부여한다. 그러나 자동 색인에서는 용어 추출 색인 방식을 취하기 때문에 기본적으로 문서에 출현한 용어들을 색인어로 선택하고, 문서에 출현하더라도 문서 내용과 무관하거나 색인 가치가 없는 용어는 제외한다.

2.1 색인어 선별 기법

문서에 출현하는 단어들의 출현 빈도는 Zipf의 법칙에 의해 각 단어들을 출현 빈도순으로 나열했을 때 출현 빈도와 단어의 순위의 곱이 상수로 계산된다[2,3,4]. Luhn은 이처럼 출현 빈도에 따라 색인어로서 가치가 있는 용어(significant word)와 가치가 없는 용어(non-significant word)로 구분하고, 출현 빈도가 매우 높은 것과 매우 낮은 것을 절단하는 방법을 사용하였

† 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

*종신회원

다. 문서에 출현한 용어들은 문서 내용을 특징짓는 용어(specialty word)와 통상적으로 문장 또는 문서를 구성하는데 사용되는 용어(function word)로 구분된다. 영어 문서에서 출현 빈도가 매우 높은 용어들을 절단하는 이유는 관사, 전치사, 접속사 등 문서 내용과 무관하게 사용되는 기능어들의 빈도수가 매우 높다는데 기인한다. 그리고 빈도가 매우 낮은 용어는 통상적으로 해당 문서의 주제와 관련성이 매우 낮다고 가정하여 색인어에서 제외한 것이다.

Luhn의 색인어 선별 기법은 영어 문서의 색인과 관련하여 "모든 단어는 색인어 후보가 된다"는 것과, "관사, 전치사, 접속사, be 동사 등 출현 빈도가 매우 높고, 대부분의 문서에서 출현하므로 문헌 빈도가 높은 단어들은 색인어로서 가치가 없다"는 영어의 언어적 특성을 이용한 것이다[5]. 이에 따라 영어에서는 색인어로 부적합한 불용어를 선택하는 기준으로 출현 빈도가 사용된다. 한국어의 경우에도 조사와 어미, 선어말 어미, 그리고 '-하다', '-이다' 등을 독립적인 토큰으로 간주한다면, 출현 빈도에 의한 색인어 선별 기법이 적용될 수 있다. 그러나 한국어와 영어에서 색인어와 불용어의 선별 방법은 교착어와 굴절어라는 언어의 특성, 조사/어미/접미사 등 문법 형태소를 색인이 가능한 용어로 간주할 것인지 등 많은 차이가 있다[6]. 따라서 한글 문서에서 색인어와 불용어를 선별하는 방법 또한 한국어의 형태론적 특성이 충분히 고려되어야 한다.

2.2 한글 문서의 색인어와 불용어

고빈어 어휘를 불용어로 간주하는 방법은 색인 DB의 크기를 줄임으로써 검색 속도를 개선하는데 매우 효율적이다. 그러나 고빈도 어휘 중에서도 의미 중의성 혹은 문맥에 따라 색인 가치가 있는 용어들이 있으므로 고빈도 어휘를 항상 불용어로 간주할 수는 없다. 예를 들어, 한국어의 1천만 어절 말뭉치에서 최상위 고빈도 어휘 30개는 다음과 같다.

있다, 그, 수, 있는, 이, 것이다, 한, 한다, 및, 대한, 것이, 것은, 하는, 할, 그러나, 우리,
때, 등, 또, 같은, 것으로, 것을, 하고, 있었다, 그리고, 없는, 위한, 했다, 따라, 것

이들은 대부분 통상적으로 모든 문서에 사용되는 어휘들이지만, '대한'의 경우에 '소한/대한'과 같이 절기를 의미하는 경우가 있다. 또한, '우리'는 '가축의 우리'라는 의미로 사용되었다면 불용어로 간주되지 않아야 한다. 영어의 경우에도 'vitamin A'의 'A'는 불용어로 간주되었을 때 검색이 되지 않으므로 상용 검색 엔진에서는 일반적으로 불용어 처리를 하지 않는 추세이다.

기본적으로 "색인 대상이 되는 어휘는 검색할 때 질의어로 사용되는 어휘"들이다. 영어에서는 단어의 품사와 무관하게 거의 모든 단어들이 '질의어'로 사용될 수 있다. 특히, 질의어로 가장 많이 사용되는 복합명사의 경우에 'automatic indexing', 'named entity' 등과 같이 명사뿐만 아니라 형용사와 분사형도 복합명사의 구성요소가 된다. 또한, 영어의 어휘들은 명사/동사/형용사 등 한 단어에 대한 품사 중의성이 매우 보편적이다. 즉, 대부분의 어휘들이 명사구를 구성하는 요소가 될 수 있으므로 단어의 품사에 의한 색인어-불용어 선별이 쉽지 않다.

반면에, 한국어의 복합명사는 명사들로 구성되는 특징이 있고, 체언과 용언은 조사/어미에 의해 품사가 명확하게 구별된다. 특히, 품사에 따라 형용사의 관형형, 책이나 영화 제목 등 일부 예외인 경우를 제외하면, 검색어는 명사들로 한정된다고 가정할 수 있다. 따라서 한국어 정보검색 시스템에서는 일반적으로 명사만을 색인어로 추출하고 있다.

3. 색인어와 색인 기법

문서에 대해 부여되는 색인어의 단위로는 형태소 단위 색인, 어절 단위 색인, 그리고 n-gram 방식의 색인 기법이 있다. 형태소 단위 색인은 단어의 형태소를 색인하는 방식이고, 어절 단위 색인은 문서에 출현한 어절 자체를 색인하는 방식이며, n-gram 방식은 단어 혹은 형태소를 구성하는 부분 문자열을 색인의 단위로 한다. 색인어 추출 기법으로는 형태소 분석기를 이용하는 방법과 n-gram 기법이 있으며, n-gram 기법의 단점을 보완하기 위하여 형태소 분석기를 이용하여 추출된 색인어에 대해서만 n-gram 기법으로 색인하는 혼합 기법이 사용되기도 한다.

3.1 형태소 색인 기법

형태소 단위 색인 기법은 문서의 내용을 대표하는 용어로부터 어근(stem)을 추출하여 추출된 어근들을 색인하는 방식이다. 이 기법은 통상적으로 (1) 불용어 제거, (2) 접미사 절단, (3) 동일 어근 검출이라는 세 단계 과정으로 구성된다. 영어에서는 고빈도 어절을 불용어로 간주하여 색인어에서 제외함으로써 색인어의 개수가 30%에서 50%까지 감소되는 효과가 있다. 접미사 절단에서는 "동일한 어근(stem)으로부터 파생된 용어들은 동일한 개념을 내포하고 있으므로 동일한 용어로 색인한다"는 원칙에 따라 파생 접미사를 제거한다.

영어는 nature, natures, natural, naturally, naturalness, naturalize 등과 같이 하나의 어근으로부터 다양한 품사로 파생되는 경우가 보편적이다. 이러

한 용어들에 대해 대표어인 nature로 색인을 한다. '동일 어근 검출' 문제는 영어에서 stemming 알고리즘에 의한 접미사 절단할 때 'absorb', 'absorpt'와 같이 어근이 동일한 형태가 아닌 경우에 동일한 어근으로 색인하기 위해 '-pt'로 끝나는 문자열을 '-b'로 교체하여 동일한 어근으로 색인되도록 하는데 필요한 과정이다.

3.2 어절 색인 기법

어절 색인은 형태소 색인 기법의 단점을 보완하기 위한 방안이다. 동일 개념으로부터 파생된 어휘들을 동일한 용어로 색인했을 때, 사용자는 'nature'가 출현한 문서를 검색하기를 바라는데, 'naturally'가 출현한 문서들이 검색되는 문제가 발생한다. 형태소 단위 색인에서는 이처럼 각 파생어들에 대해 각각 검색하고자 하는 사용자의 검색 의도를 반영할 수가 없다. 따라서 정보 자료의 수가 너무 많은 웹 검색 엔진에서는 어절 자체를 색인하는 것이 적합 문서를 검색하는데 효율적이다. 이러한 비효율성을 극복하기 위한 방안으로 어절 색인 기법에서는 형태소 색인 기법의 '접미사 절단'과 '동일 어근 검출' 과정을 하지 않고 어절 자체를 색인하는 방법을 취한다.

구글의 영어 검색 엔진의 경우, stemming을 하지 않기 때문에 명사의 단수형과 복수형도 각각 별개의 색인어로 색인하고 있다. 예를 들어, 'hotel'과 'hotels'는 별개의 색인어로 간주되기 때문에 'hotel'을 검색하면 'hotels'가 검색되지 않는다. 그러나 구글 검색 엔진에서 한글 문서의 색인은 한글의 교착어 특성을 감안하여 영어와 달리 형태소 분석을 통해 색인을 한다. 즉, 영어의 접미사에 대응되는 한국어의 문법 형태소 '조사/어미/접미사'를 형태소 분석기에 의해 분리한 후에 각 형태소를 색인어로 추출한다. 그런데 탈락과 불규칙 활용 등 변형된 어간이나 변형된 어미는 원형을 복원하지 않는다. 변형된 어간의 경우에 입력 어절에서 어간의 길이만큼 음절 단위로 절단하여 색인어로 추출한다. 예를 들어, '아름다운'은 '아름답'+'ㄴ'이므로 '아름다'와 '운'으로 색인을 한다. 구글의 어절 색인 방식은 원문 일치(exact match) 검색의 효과를 얻을 수 있는 반면에, 한글 문서의 검색에서는 질의어가 '게'일 때 '어떻게/빠르게/짧게' 등이 검색되는 단점이 있다.

한국어에서는 '바람과 함께 사라지다', '샤갈의 눈 내리는 마을' 등과 같이 제목으로 검색할 때 명사 색인어만으로는 적합 문서를 검색하기 어려운 문제를 해결하는 방안으로 어절 자체를 색인하는 방법을 사용하기도 한다. 이 방법에서는 한 어절에 대해 2개 이상의 색인어를 추출하는 다중 색인 기법을 취하여 형태소 분석에 의한 어휘 형태소와 더불어 어절 자체를 색인어로 추출한다. 다중 색인에 의한 어절 색인 기법은 소수의 질의어에 대한 검색 만족도를 높이기 위해 모든 어절에 대한 색인을 해야 하므로 영어 문서에 대한 어절 색인과는 달리 색인어의 개수가 매우 많아지는 단점이 있다.

3.3 n-gram 색인 기법

n-gram 색인 기법은 한글, 한자 등 2 바이트 문자 체계(DBCS: double-byte code system)를 사용하는 언어에서 stemming과 접미사 절단 등 형태소 분석에 의한 어근 추출 오류를 해결하는 방안으로서 형태소 분석을 통하지 않고 모든 부분 문자열들을 각각 색인하는 방법이다. n이 2인 bigram 기법이 가장 많이 사용되는데, 연속된 2개 문자들을 모두 색인하므로 '컴퓨터 시스템을'의 경우 '컴퓨', '퓨터', '터시', '시스', '스템', '템을'이 색인어로 추출된다.[1] 따라서 n-gram 색인 기법에서는 모든 어휘들에 대한 부분 문자열을 검색할 수 있는 장점이 있는 반면에, 다른 유형의 검색 오류로써 '핑클'에 대해 '서핑 클럽'이 검색되는 문제가 발생한다. 이 오류는 질의어로 시작되거나 혹은 끝나는 것만 검색되도록 제한하는 방법에 의해 방지할 수 있다. 그러나 이러한 오류를 방지하려면 모든 bigram 색인어마다 어휘의 시작-끝 위치 정보를 기록해야 하기 때문에 색인 DB의 크기가 더욱 커지게 된다. 또한, n-gram 색인 기법은 검색 시에도 질의어를 각각 2 문자씩 분할하여 검색한 후에 AND 연산을 해야 하므로 검색 속도가 느리다. 즉, '컴퓨터 시스템'을 검색할 때 bigram 방식에서는 최소한 5번의 bigram 문자열에 대한 검색과 각 검색 결과에 대한 AND 연산이 필요하다.

n-gram 색인 기법은 형태소 분리 문제를 쉽게 해결하는 장점이 있는 반면에, (1) 색인 DB의 크기가 매우 크고, (2) 검색 속도가 낮으며, (3) 사용자에게 색인어를 제시하지 못하는 단점이 있다. n-gram 색인 기법의 단점을 개선하는 방안으로 "형태소 분석기를 이용하여 색인 대상이 되는 어휘들을 추출하고 추출된 색인어에 대해서만 n-gram 기법을 적용"하는 방법과, "형태소 분석기에 의해 추출된 용어 중 복합명사에 대해서만 n-gram 기법을 적용"하는 방법이 있다. 이 개선 방안들은 결국, 형태소 분석에 의한 색인 기법에서 복합명사를 분해하지 않고 n-gram 방식으로 복합어를 분해한 것으로 볼 수 있다.

영어와 같이 1 바이트 문자를 사용하는 언어에서는 복합어의 띄어쓰기가 정확하기 때문에 이 방법을 적용할 필요가 없으나, 한글과 한자어에서 색인 및 검색 효율성

1) 용어 앞뒤의 공백 문자도 색인하는 방식을 취하면 색인어에 '_컴'과 '을_'이 추가된다.

이 낮아지는 것을 감수하고 형태소 분리와 관련된 색인 문제를 쉽게 해결하는 방안으로 사용되기도 한다. n-gram 기법은 초기 시스템이나 연구개발 단계에서 실험용으로 사용되기도 하였으나 상용 시스템은 형태소 분석기를 이용한 색인 기법을 취하고 있다.

4. 한글 문서의 색인 특성

4.1 색인어 추출 기법

색인어 추출 기법에는 2가지 방법이 있다. 첫 번째 방법은 일반적으로 색인어 추출 기법으로 널리 알려져 있는 "stemming과 접미사 절단"[2] 기법이고, 두 번째 방법은 일본어나 중국어 등과 같이 띄어쓰기를 하지 않기 때문에 단어 분할(word segmentation) 문제가 형태소 분석의 핵심이 되는 언어에서 사용하는 "형태소 분리와 변형"[3] 기법이다. "stemming과 접미사 절단" 기법은 기본적으로 단어가 공백 문자로 구분되어 있으므로 입력 어절에 대해 불용어를 제거하고 stemming과 접미사 절단하는 과정으로 구성된다. 이에 비해, "형태소 분리와 변형" 기법에서는 형태소들이 공백에 의한 구분이 없이 접속되어 있으므로 형태소(또는 단어) 경계를 인식하는 문제가 가장 중요한 요소이다.

위 두 가지 기법은 언어의 분류 체계에서 굴절어(inflectional language)와 교착어(agglutinative language)의 특성 차이에 의한 것으로 한국어는 교착어에 속하기 때문에 "형태소 분리와 변형" 기법이 적합하다. 그러나 띄어쓰기를 전혀 하지 않는 일본어나 중국어와 비교할 때, 독립된 어절로부터 조사/어미를 분리하는 과정을 굴절어의 '접미사 절단'에 대응하는 것으로 간주할 수 있기 때문에 "stemming과 접미사 절단" 기법이 더 적합하다고 볼 수도 있다. 즉, 한국어의 경우 두 가지 색인어 추출 기법의 관점에 볼 때 경계 언어에 속한다. 즉, 한국어는 어절 경계가 분명하고 어근을 중심으로 어절이 구성되어 있으므로 "stemming과 조사/어미 절단" 기법을 적용하고, 복합명사 분해 문제와 통상적으로 자주 범하는 띄어쓰기 오류에 대해서는 "형태소 분리와 변형" 기법을 적용하는 두 단계 과정으로 구분하여 색인어를 추출하는 것이 바람직한 방법이다.

4.2 한글 문서의 색인 방식

한글 문서의 색인 방식은 색인 대상이 되는 색인어의 기준을 무엇으로 결정할 것인지, 그리고 어떤 유형의 어휘들을 불용어로 간주할 것인지에 따라 차이가 있다. 한국어의 색인어와 색인 방식은 색인어 선택 기준에 따라 아래와 같이 여러 가지로 구분된다.

- 어휘 형태소 색인 기법

'어휘 형태소 색인' : 모든 어휘 형태소들을 색인어로 추출

'명사 추출 색인' : 어휘 형태소 중에서 명사만을 색인어로 추출

- 기능어 색인 기법

'형태소 원형 색인' : 조사/어미를 포함한 모든 형태소들의 기본형을 색인

'형태소 분할 색인' : 조사/어미를 포함한 모든 형태소들의 표층형을 색인

- 복합어 색인 기법 - 복합명사 분해-조합 및 구성 요소 색인 여부에 따라

'복합어와 구성 요소 색인' : 복합어와 더불어 해당 구성 요소를 색인어로 추출

'복합어 조합 색인' : 복합어 구성 요소간의 조합형들을 색인어로 추출

'복합어 분해 색인' : 복합어의 구성 요소만 색인어로 추출

색인어 선택 기준에 따른 각 색인 기법별로 아래 예문에 대한 색인 예는 다음과 같다.

〈색인 예문〉

국민대학교는 북한산 국립공원 아래의 수려한 캠퍼스와 첨단 정보화된 교육시설 및 국내 어느 대학보다 뛰어난 교수진을 갖추고 있는 종합대학으로서 "다른 생각이 다른 내일을 만들어간다"는 정신으로 학생들의 개성과 창의를 개발하고 있습니다.

4.2.1 어휘 형태소 색인 결과

- 국민대학교 북한산 국립공원 아래 수려하다 캠퍼스 첨단 정보화 교육시설 및 국내
- 어느 대학 뛰어나다 교수진 갖추다 있다 종합대학 다르다 생각 다르다 내일
- 만들다가다 정신 학생 개성 창의 개발하다 있다

4.2.2 명사 추출 색인 결과

- 국민대학교 북한산 국립공원 아래 캠퍼스 첨단 정보화 교육시설 국내 대학 교수진
- 종합대학 생각 내일 정신 학생 개성 창의

4.2.3 형태소 원형 색인 결과[4]

- 국민대학교+는 북한산 국립공원 아래+의 수려하

2) 영어의 접미사 절단(suffix stripping) 기능은 한국어에서 기능어(조사, 어미, 접미사)를 분리하는 것으로 볼 수 있다.

3) 일본어, 특히 중국어의 형태소 분리는 어휘 형태소와 기능 형태소 간의 분리가 아니라, 기본적으로 어휘 형태소 간의 분리를 의미한다.

4) 이 색인 결과에서 '+' 기호는 색인어 구분자로써 일관성 있게 표현하려면 공백으로 해야 하지만, 이해를 돕기 위해 한 어절에서 분리된 색인어의 구분자는 '+'로 표현하였다.

+ㄴ 캠퍼스+와 첨단 정보화+되+ㄴ
- 교육시설 및 국내 어느 대학+보다 뛰어나+ㄴ 교수진+을 갖추+고 있+는
- 종합대학+으로서 다르+ㄴ 생각+이 다르+ㄴ 내일+을 만들어가+ㄴ다 는 정신+으로
- 학생+들+의 개성+과 창의+를 개발하+고 있+습니다

4.2.4 형태소 분할 색인 결과

- 국민대학교+는 북한산 국립공원 아래+의 수려한 캠퍼스+와 첨단 정보화+된
- 교육시설 및 국내 어느 대학+보다 뛰어난 교수진+을 갖추+고 있+는
- 종합대학+으로서 다른 생각+이 다른 내일+을 만들어간+다 는 정신+으로 학생+들+의
- 개성+과 창의+를 개발하+고 있+습니다

위 4가지 색인어 추출 예는 복합명사에 대해 구성 요소 분해를 적용하지 않은 것으로써 복합명사의 색인 방식에 따라 위 예문의 복합명사 '국민대학교, 국립공원, 교육시설, 종합대학'에 대한 색인어 추출 예는 아래와 같다.

- 복합어와 구성 요소 색인

국민대학교, 국민, 대학교, 국립공원, 국립, 공원, 교육시설, 교육, 시설, 종합대학, 종합, 대학

- 복합어 분해 색인

국민+대학교, 국립+공원, 교육+시설, 종합+대학

"복합어와 구성 요소 색인" 기법은 복합어 자체와 그 복합어를 구성하는 요소들을 모두 색인하는 방법이다. 따라서 이 기법은 복합어와 구성 요소가 모두 색인어로 추출되므로 한 어절에 대해 동일 문자열이 중복해서 색인이 되는 다중 색인 기법이다. 이와 유사한 다중 색인 기법으로 "복합명사 조합 색인"은 복합어의 구성 요소가 3개 이상인 경우에 각 구성 요소를 조합하여 만들어지는 복합명사들을 색인하는 방법이다. 예를 들어, '정보검색 시스템'에 대해 '정보', '검색', '시스템'과 더불어 '정보검색', '검색 시스템'을 조합하여 색인어에 추가한다. 또한, 이웃한 구성 요소가 아니더라도 '정보 시스템'과 같이 색인어로 사용될 수 있는 색인어를 조합하여 색인어에 추가할 수 있다.

"복합어 조합 색인" 기법은 검색 가능한 복합어 조합들을 미리 색인하는 전조합 방식을 취하는 반면에, "복합어 분해 색인"은 복합어를 구성하는 구성 요소만을 색인하기 때문에 질의어 처리 과정에서 입력된 문자열이 단위 형태소로 분해하는 과정이 선행되어야 한다.

4.3 복합어와 다중 색인

한글 문서에서 복합어는 단위명사(혹은 접미사)로 구성되고 있으며, 질의어가 단위명사일 때 복합어를 검색해야 할 필요가 있다. 따라서 검색의 편의성을 위해 복합어 자체와 복합어의 구성 요소, 그리고 경우에 따라서는 구성 요소들의 조합을 색인하기도 한다. 이 때, 한 어절에 대해 2개 이상의 색인어가 부여되는 다중 색인 현상이 발생한다. 다중 색인 기법은 영어와 같이 한 단어에 어근이 1개인 언어에서는 불필요하다고 판단된다. 그러나 한국어의 같이 복합명사처럼 구성 요소를 색인하거나 복합어에 대한 2개 이상의 지시어를 부여할 경우에 적합한 방법이다. 예를 들어, '국민대학교'는 아래와 같이 '국민대', '국민대학' 등의 질의어에 대해 검색이 되어야 하므로 아래와 같이 다중 색인을 허용할 필요가 있다.

'국민대학교' → '국민대학교', '국민대', '국민대학', '국민', '대학', '대학교', '학교'

또한, 한국어의 복합명사는 구성 요소의 경계가 명확하지 않고 모호하여 복합명사 분해 결과가 중의적인 경우가 있다. 예를 들어, '칼국수집'은 '칼국수'+'집', '칼'+'국수집', '칼'+'국수'+'집' 등 3가지가 가능하다. 이 경우에 '칼'+'국수'+'집'와 같이 (1) 구성 요소의 단위를 세분화하고, (2) 질의어 또한 동일한 방식으로 분해하여 후조합 색인 방식을 취하는 방법으로 '칼국수'와 '국수집'에 대한 검색 요구를 해결할 수 있다. 그러나 세분화된 복합어 분해 방식으로 모든 문제가 해결되는 것은 아니다. 다른 예로써, '국민대학교'는 '국민'+'대학교', '국민대'+'학교', '국민'+'대'+'학교', '국민'+'대학'+'교' 등이 가능하다. 그런데 '국민대', '국민대학', '국민', '대학', '학교' 등 모든 검색 요구를 만족시키는 분해 방법은 존재하지 않는다.

이러한 문제점을 해결하는 방안으로 "복합명사에 대해서만 n-gram 색인 기법을 적용하는 방법"을 사용할 수 있다. 그러나 n-gram 색인 기법에서는 '칼국수집'에서 '국수', '국민대학교'에서 '대학교'를 검색할 수 있으나, '서핑클럽'에서 '핑클', '맞춤정보'에서 '춤정보', '일반문서'에서 '반문'과 '반문서'가 검색되는 또 다른 문제를 발생시킨다. 따라서 이 문제에 대한 근본적인 해결 방법은 없으며, 그 중에서 색인어의 중요도에 따라 1가지를 취하는 방법이 옳을 것이다. 예를 들어, '칼국수집'에서는 '국수집'이나 '국수'보다는 '칼국수'를 취하여 '칼국수'+'집'으로 분해한다. 웹 검색 엔진은 특히 검색되는 문서들의 수가 너무 많아서 구글과 같이 hotel과 hotels와 같은 복수형도 별개의 색인어로 간주하여 색인하는 것이 사용자 만족도가 높게 나타난다. 따라서 한국어의 복합명사 분해 문제도 세분화를 적게 하는 것이 더 효율적일 것으로 판단된다.

5. 결 론

정보검색을 위한 한국어의 색인어와 색인 기법에 대해 고찰하였다. 한국어는 2바이트 문자를 사용하는 언어의 특성과 복합명사 분해 문제가 색인어 추출과 색인 기법에 가장 큰 영향을 미친다. 이러한 한국어의 특성은 초기 시스템에서 중국어, 일본어와 같이 n-gram 색인 기법을 사용하는 원인을 제공하였다. 그러나 n-gram 기법은 여러 가지 비효율적인 문제로 인해 상용 시스템은 대부분 형태소 분석기를 이용한 색인 기법을 사용하고 있다.

한국어의 질의어는 거의 명사 유형이다. 그러나 '아름다운 세상'과 같은 명사구, 패션-색상 정보에서 '빨간 모자'와 같은 형용사 질의어, 그리고 '바람과 함께 사라지다' 유형의 제목에 대한 원문 일치 검색과 관련하여 명사 이외의 어휘들도 색인할 필요가 있다. 즉, 명사만을 색인하면 이러한 검색 요구를 충족시키지 못하므로 이를 보완하기 위해 '어절 색인 기법'을 사용하는 방법과 구글과 같이 형태소 분석 결과로 추출되는 조사/어미/접미사 등 모든 형태소를 색인하는 방법이 있다. '어절 자체를 색인하는 방법'과 기능어의 '형태소 원형 색인'(혹은 '형태소 분할 색인')은 장단점이 있으며, 어떤 방법이 더 효율적인지 검증할 필요가 있다.

한글 문서의 색인 기법에서 가장 중요한 문제의 하나는 복합어, 특히 복합명사의 색인과 관련된 것이다. 복합어는 그 자체뿐 아니라 구성 요소를 색인하는 방법과 복합어 자체는 제외하는 방법이 있는데, 이 또한 어떤 방법이 더 효율적인지 연구되어야 할 문제이다.

참고문헌

[1] Baeza-Yates, R. and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[2] Zipf, H. P., Human Behaviour and the Principle of Least Effort, Addison-Wesley, 1949.

[3] Miller, G. A. and E.B. Newman, "Tests of a statistical explanation of the rank-frequency relation for words in written English," American Journal of Psychology, 71, pp.209-218, 1958.

[4] Miller, G. A., E. B. Newman, and E. A. Friedman, "Length-frequency statistics for written English," Information and Control, 1, pp.370-389, 1958.

[5] Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, 2, pp.159-165, 1958.

[6] 강승식, 한국어 형태소 분석과 정보 검색, 홍릉출판사, 2002.

강 승 식

1982~1986 서울대학교 컴퓨터공학과(학사)
1986~1988 서울대학교 컴퓨터공학과(석사)
1988~1993 서울대학교 컴퓨터공학과(박사)
1994~2001 한성대학교 정보전산학부 부교수
2001~현재 국민대학교 컴퓨터학부 부교수
E-mail : sskang@kookmin.ac.kr