한국어 정보검색을 위한 색인어 추출방법에 관한 연구

고려대학교 컴퓨터학과
최순우† · 김상범 · 임해창

A Study on Keyword Extraction for Korean Information Retrieval System

Soon-Woo Choi†, Sang-Bum Kim, Hae-Chang Rim
Department of Computer Science, Korea University, Seoul, Korea

요 약

본 논문에서는 색인 방법에 따른 한국어 정보검색시스템의 성능차이를 살펴보고 이를 분석하여 보다 검색성능을 높이기 위한 색인어 추출방법을 제안한다. 이를 위해 기존의 대표적인 색인법이라 할 수 있는 명사단위 색인법, 형태소 단위 색인법, 바이그램 단위 색인법, 어절단위 색인법에 대하여 실험을 통한 비교분석을 하였고, 질의별 분석을 통해 검색성능에 영향을 주는 요소들을 찾아내었다. 그 결과 빈칸, 명사분해, 명사, 동사, 형용사, 숫자등을 포함한 실질 형태소, 형식형태소의 제거, 외래어 등 추정명사의 분해 및 발음확장, 후방 단음절 명사로 구성된 복합명사의 분해, 의미를 변절 시키는 바이그램 제거, 분해된 명사 수에 따른 복합명사 첨가 및 제거 등이 그 요소임을 확인할 수 있었다. 이를 토대로 각 색인법의 장점을 살려 색인 및 검색을 수행하여 보았다. 제안하는 방법은 동일한 실험집합에서 일관성 있는 성능향상을 가져다 줌을 알 수 있었다.

서 론

대량으로 저장되어 있는 문서의 내용을 분석 한 뒤 찾기 쉬운 형태로 조직하여 보관하고 있다가 정보에 대한 요구가 발생하였을 때 해당 문서를 찾아 제공하는 시스템을 정보검색시스템(Information Retrieval System)이라 말한다. 특히 인터넷의 발달로 인해 대량의 정보가 전자 문서화되어 공유되고 있으며, 수 많은 문서들 중에서 적합한 문서를 빠른 시간 내에 찾아낼 수 있는 정보검색시스템에 대한 요구가 증가되고 있는데, 이를 위해 문서들을 자동으로 분석하기 위한 방법이 꾸준히 연구되고 있고, 다양한 방법이 제시되고 있다.

문서를 분석하는 방법이란 일반적으로 문서를 대표하는 색인어(단어)를 추출하는 것을 말한다. 일반적으로 영어는 빈칸을 구분자로 하여 단어를 추출한 후 원형복원(Stemming)하는 방법을 사용하나 한국어는 여러 개의 단어가 조합되어 하나를 이루기 때문에 색인어 추출에 많은 어려움이 있다. 한국어는 여러 개, 여러 종류의 단어가 합하여 하나의 단어를 이루기 때문에 영어와는 달리 색인어를 얼마나 잘 추출하느냐에 따라 정보검색시스템 성능에 많은 차이를 보일 수 있다.

기존의 한국어 정보검색시스템에서 사용된 색인방법들을 살펴보면, 명사단위 색인,[1] n-gram단위 색인(바이그램 단위 색인),[2] 형태소 단위 색인[3] 등이 대표적이며 이 외에도 위치 관계를 고려한 구 단위 색인, 유의어 및 발음확장, 접사처리, 구분할을 이용한 명사구 기반 색인 등 성능향상을 위해서 다양한 방법이 시도되었다.

특히 [2]는 어절단위 색인법, 형태소단위 색인법, n-그램(바이그램)단위 색인법을 소개하고, 이를 비교하였다. 또한 각 색인법을 조합하여 성능향상을 보였다. 우선 어절단위 색인법은 빈칸 등을 구분자로 하여 문서 내에서 어절을 인식한 후 최장일치원칙에 의한 비색인분절 및 불용어를 제거한 색인어를 추출하였다. 형태소 단위 색인법은 최소 의미의 명사들을 색인어로 추출을 하였는데, 절차는 형태소 해석, 애매성 제거, 단일명사 추출, 불용어 제거의 네 단계를 거치고 있다. 이와 달리 n-그램 색인법은 빈칸 등의

†E-mail : chsw@nlp.korea.ac.kr
E-mail : sbkim@nlp.korea.ac.kr
E-mail : rim@nlp.korea.ac.kr

구분자로 어절을 구분한 뒤 불용어 제거, 최장 일치법에 의한 비색인 분절어 절단, 추출된 색인어에 대해 음절단위 n-그램을 뽑아 색인어로 사용하였다. 또한 형태소 단위 색인법으로 검색한 결과와 음절단위 n-그램으로 검색한 결과를 조합하여 성능의 개선을 보여주었다.

그러나 [2]에서는 상세한 분석을 통하여 각 색인방법의 장단점을 분석하지는 않아, 한국어 정보검색시스템을 위해서 어떠한 전략으로 색인방법들을 조합해야 하는지에 대해서는 언급하지 않고 있다. 본 논문에서는 기존의 여러가지 색인방법을 사용하여 검색성능을 평가한 후, 이를 질의별로 분석하여 검색성능에 영향을 주는 요소를 찾아내고, 이러한 요소들을 반영하여 적절한 한국어 색인어 추출방법을 제안하고자 한다.

본 논문에서는 색인방법으로는 명사단위 색인방법, 형태소 단위 색인방법, 바이그램 단위 색인방법, 어절단위 색인방법을 사용하였으며, 가중치 할당방법은 TREC에서 우수한 성능을 나타내는 2-포아송모델을 사용하였다. 또한 실험결과를 통하여 각 색인방법의 장단점을 분석하고 이를 기반으로 보완된 한국어 자동색인방법을 제안한다.

한국어 색인어 추출 방법

1. 명사단위 색인어 추출방법

[1]은 명사단위 색인을 위해 명사를 추출하는 방법을 소개하고 있다. 이 논문에서는 다른 색인방법과 비교설명을 하고 있지는 않고, 질의어의 조합에 따른 성능을 비교하였으며 적합성 피드백에 따른 영향을 분석하고 있다. 명사 추출하는 방법을 간단히 살펴보면 분석배제 정보를 이용하여 형태소 분석을 배제하고, 후절어를 이용하여 명사를 추출하였고, 명사로 추출되지 않은 어절에 대해서는 형태소 분석과정을 거쳐 음운현상 복원을 통하여 명사를 추출하였다.

본 논문에서는 비교실험을 위하여 명사단위 색인으로 [1]이 제안하고 있는 명사추출방법을 사용하였는데, 그 과정은 다음과 같다.

1) 분석배제 정보를 이용한 형태소 분석 배제 : 명사가 없는 어절에 대해서는 분석과정을 생략한다.

2) 후절어을 이용한 명사추출 : 후절어 앞에 위치한 체언을 검사하여 명사추출한다. 복잡한 형태소 분석과정을 거칠 필요가 없다.

3) 형태소 분석과 음운 현상 복원 : 후절어 분석을 통하여 명사로 분석되지 않은 경우에 해당한다.

예를 들면 '정보검색에서' 이란 어절은 '정보, 검색, 정보검색' 으로 색인어가 추출된다.

이와 같이 명사추출기는 완전한 형태소분석을 수행하지 않고 분석배제정보나 후절어정보등을 사양하여 최소한의 언어지식(문법 및 사전 등)을 사용하면서도 효율성을 극대화한 색인어 추출 방법이다. 따라서 정확도에 있어서는 형태소 분석기에 비하여 떨어진다.

2. 형태소 단위 색인어 추출방법

명사추출방법과 달리 [3]에서는 품사부착기를 사용하여 색인어를 추출하였다. 먼저 형태소를 분석하고, 품사를 태깅한 후 그 결과를 사용하여 색인어를 추출하였는데, 이와 동시에 여러 가지 규칙들을 이용하여 합성명사를 생성하고, 합성명사로 추출된 색인어의 경우 합성명사분할을 통하여 색인어를 추출하였다. 또한 접사를 분리 제거 하였으며, 유의어 및 발음확장 사전을 사용하여 유의어 및 발음확장을 하였다.

본 논문에서는 자연언어처리기법에 의한 색인어 추출의 성능평가를 위하여 [6]에서 제안된 품사부착기를 사용하였는데, [6]은 수동으로 품사부착된 말뭉치로부터 HMM기반 학습을 통하여 자동으로 품사를 부착하는 품사부착기이다. 본 논문에서는 [6]의 분석결과에서 실질 형태소 만을 채택하였으며 또한 접사 중 접두사는 중요한 키워드가 될 수 있다고 판단하여 함께 색인하였다. 그 구체적인 순서는 다음과 같다.

1) 어절 또는 문장에 대한 품사 부착기 결과를 얻는다.

2) 얻어진 결과에서 부착된 품사를 보고 실질 형태소를 추출한다.

3) 실질 형태소에서 '/', 부착된 품사를 제거하여 최종 색인어를 얻는다.

예를 들어 '자율이동 로봇에 관해서' 를 분석하여 '자율/NNCG+이동/NNCV+로봇/NNCG+에/PA+관하/VV+아서/EFG' 의 결과를 얻으면 품사와 '/' 을 제거하여 '자율, 이동, 로봇, 에, 관하, 아서' 의 색인어를 얻게 된다.

형태소 분석을 하기 위해서는 다양한 언어지식(문법 및 사전 등)이 필요하고, 품사를 부착하기 위해서 정밀한 검사를 요하기 때문에 속도가 떨어지나, 결과가 명사추출기에 비하여 비교적 정확하다는 장점을 가지고 있다.

3. 바이그램 색인어 추출방법

바이그램 색인어 추출방법은 흔히 영어권에서 사용되는 n-그램 방식의 색인어 추출 방법으로, 일반적으로 한국어는 두개의 음절이 모여 하나의 의미를 나타내는 단어를 형성하기 때문에 본 논문에서는 [2]와 마찬가지로 바이그램 색인어 추출방법을 사용하였다. 그 구체적인 방법은 다음과 같다.

1) 빈칸, 특수문자 등을 구분자로 하여 어절을 추출한다.

2) 어절 내에서 영문자, 숫자, 특수문자 등을 구분하여 추출한다. 한편, 영어단어가 추출되었을 경우에는 어근화(Stemming)를 수행한다.

3) 추출된 순수 한국어 어절을 첫 음절로부터 인접한 2음절씩 색인어를 추출한다.

예를 들면 '정보검색'이란 어절은 '정보, 보검, 검색'으로 색인어가 추출된다.

바이그램 색인어 추출방법은 별도의 언어지식을 필요로 하지 않는 방법으로, 손쉬운 방법으로 정보검색시스템을 구축할 수 있도록 해준다. 언어지식을 사용하지 않으므로 분석속도가 빠르나 의미 없는 분석결과가 상당수 발생하며 분석되어 나오는 색인어의 수가 비교적 많아 역파일의 크기가 커진다.

아울러 본 논문에서는 비 색인분절 절단시 오류를 발생시킬 문제를 줄이고, 자연어 처리 기술이 반영되지 않는 색인법임을 강조하기 위하여 비 색인분절어를 절단을 하지 않았다.

4. 어절단위 색인어 추출방법

일반적으로 한국어는 의미를 가지는 말이 앞에 존재하고 형식 형태소는 뒤에 존재하여 어절의 앞부분에 실질 형태소 및 명사를 가지게 된다. 어절을 그대로 색인하는 방법도 있으나, 어절 앞 부분에 존재하는 실질 형태소 및 명사를 살리기 위해서 다양한 방법들이 제시되고 있다. 다음은 본 논문에서 사용한 방법이다.

1) 빈칸, 특수문자 등을 구분자로 하여 어절을 추출한다.

2) 어절 내에서 영문자, 숫자, 특수문자 등을 구분하여 추출한다. 추출된 영문자는 Stemming처리를 한다.

3) 추출된 어절을 첫 2음절부터 한 음절씩 더하여 색인어를 추출한다.

예를 들면 '정보검색'이란 어절은 '정보, 정보검, 정보검색'으로 색인어가 추출된다.

어절단위 색인법은 바이그램과 같이 언어지식이 필요없는 색인법으로 바이그램보다 추출되는 색인어의 의미변질이 작으나 어절중간에 존재하는 명사등을 추출하지 못하는 단점이 있다. 언어지식을 사용하지 않으므로 처리속도는 빠르다. Table 1은 위에서 제시한 색인어 추출방법별로 "자율이동 로봇에 관해서"라는 어구를 분석한 결과이다.

Table 1. 색인어 추출방법별 예제

색인어	추출된 색인어
명 사	자율, 이동, 자율이동, 로봇,관해
형 태 소	자율, 이동, 로봇, 관하
바이그램	자율, 율이, 이동, 로봇, 봇에, 관해, 해서
어 절	자율, 자율이, 자율이동, 로봇, 로봇에, 관해, 관해서

색인어 추출방법에 따른 검색성능의 비교분석

1. 문서집합

연구개발정보 센터에서 배포한 HANTEC 2.0 질의집합 및 문서집합을 사용하였다. 질의집합은 전체질의 50개와 과학질의 30개로 구성되어 있으며, 본 논문에서는 전체질의 50개를 사용하였으며, 평가를 위한 정답집합으로 G2를 사용하였다. Table 2는 문서집합을 이루는 문서들에 대한 내용이고, Table 3은 사용한 질의어의 예이다.

2. 검색모델

본 논문은 색인어 추출 방법에 따른 검색성능을 비교하는 것이 목적이므로 한 종류의 가중치 할당방법을 사용하였다. 2-포아송(Poisson) 확률분포에 기반한 TREC-8의 Okapi 시스템[5]이 사용하였던 BM25방법을 사용하여 가중치 할당을 하였다. 수식은 다음과 같다.

$$sim(d_j,q)=\sum_{t_i \in d_j \cap q}\left(\frac{tf_i}{k1\bullet((1-b)+b\bullet\frac{dl_j}{avdl})+tf_i}\bullet\log\frac{N-df_i+0.5}{df_i+0.5}\right)\bullet qtf_i$$

위 수식에서 K1과 b값은 실험에 의해서 적절한 값을 선택해야 하는데, [1]에서 같은 실험집합(HANTEC 2.0)에 대

Table 2. 문서집합

구 분	문서집합
일반종합분야	1994년에 발행된 한국일보 기사 : 22,000 건 gov 확장자를 갖는 웹 페이지 : 9,000 건 com 확장자를 갖는 웹 페이지 : 9,000 건
사회과학분야	1994년에발행된 한국경제신문기사 : 39,480 건 한국여성개발원이 발행한 정기간행물 여성연구에 게재된 논문 : 110 건 경북도의회 회의록 : 410 건
과학기술분야	과기처지원 연구보고서 : 10,000 건 해외과학기술 동향 : 18,000 건 논문 서지 사항 : 12,000 건
일반종합분야	1994년에 발행된 한국일보 기사 : 22,000 건 gov 확장자를 갖는 웹 페이지 : 9,000 건 com 확장자를 갖는 웹 페이지 : 9,000 건
계	120,000 건

Table 3. 질의어 예시

종 류	질 의 어
TITLE	로봇
DESC	자율이동 로봇에 관해서
NARR	자율이동 로봇 자체의 설계와 개발 및 평가 등이 종합적으로 쓰여진 문헌이나, 혹은 자율이동 로봇의 부분적인 시스템 (경로제어, 물체인식 등) 설계에 ....

해 실험하여 1.5, 0.5값을 가장 좋은 값으로 설정하였으므로 별도의 실험없이 해당 값을 설정하였다(tf : 색인어 빈도수, dl : 문서길이, avdl : 문서의 평균길이, df : 문서빈도, qtf : 질의어내의 색인어 빈도수, N : 전체 문서수).

3. 색인결과

다음 Table 4는 4가지 색인어 추출 방법으로 색인한 결과이다.

Table 4에서 문서의 길이는 총 색인어의 수로 하였다. 명사단위 색인이 한 문서 내에서 추출되는 색인어의 수가 다른 색인법의 1/2정도이고, 총 색인어의 수는 형태소 단위 색인과 바이그램 단위 색인보다 크고, 색인어 당 DF 값이 작다.

이는 검색되는 문서의 수가 가장 작아 검색속도가 빠르다는 것을 의미한다. 어절단위 색인 또한 색인어당 DF가 작으므로 다른 색인법에 비하여 검색속도가 빠름을 짐작할 수 있다.

Table 5는 TITLE필드를 사용하여 검색되는 총 문서의 수를 표로 나타내었다. 정답수는 Top1000에 존재하는 정답 수 이다.

결과를 보면 명사단위 색인법과 어절단위 색인법의 경우 검색된 문서 중 정답문서의 수의 비율로 평가할 때 가장 효율적인 것으로 보인다. 형태소 단위 색인의 경우는 가장 많은 문서의 수를 SELECT하는 것으로 보이나 특정 몇 개의 질의에 대하여 많은 문서를 SELECT하여 평균치가 높은 것으로 나타났다. 모든 질의에 대해 상위 1000개의 문서에 포함된 정답 수는 거의 비슷했으나 바이그램 색인 방법이 다소 많은 정답을 상위에 랭크시켰음을 알 수 있다.

4. 검색정확도의 비교

4가지 방법으로 색인했을 경우의 검색정확도를 비교하여 보았다. 평균 정확도(Average Precision), 상위 10문서에 대한 정확도(10-Precision)와 상위 1000개 문서에 대한 재현률(Recall)에 대하여 평가를 하였다.

대체로 어절 단위 색인법을 제외하고는 비슷한 결과를 보이고 있다. 어절단위 색인법의 경우 다른 색인법과는 달리 단어 중간에 있는 명사 등을 색인하지 못하여 성능이 저하되는 것으로 판단이 된다.

또한 질의어가 길어질수록 명사단위 색인법이 우수해지는데, 이는 의미없는 바이그램이 많이 생성되는 바이그램 방법이나 형태소분석 과정에서 나타나는 복합명사의 단일 색인어 추출이나 형태소분석 자체의 오류등에 긴 질의가 민감하게 반응하는데 비해 명사추출방법의 경우 이러한 문제가 상대적으로 덜하기 때문이라 짐작된다.

가장 짧은 질의인 TITLE에서는 형태소 단위의 색인법의 성능이 가장 좋은 것으로 나타났다.

Table 4. 색인결과

색인어	문서 평균길이	총 색인어 수	평균 문서빈도
명 사	235	1,543,790	9
형 태 소	538	852,544	25
바이그램	548	390,890	87
어 절	548	5,015,940	8

Table 5. 검색되는 문서의 총 수

색인어	평 균	최 대	최 소	정답수
명 사	9,105	33,304	122	2,067
형 태 소	16,398	92,069	121	2,048
바이그램	15,163	49,124	579	2,128
어 절	7,628	43,402	296	2,023

5. 빈칸의 중요성

빈칸의 중요성을 알기 위해서 명사단위 색인법(명사추출A)과 명사단위 색인법+이은단어 색인법(명사추출B), 그리고 바이그램 색인법(바이그램A)과 빈칸없는 바이그램 색인법(바이그램B)을 비교하여 보았다.

여기서 이은단어 색인법이란 명사추출기로 명사를 추출한 후 빈칸을 사이에 둔 명사들을 조합하여 함께 색인하는 방법을 말하며, 빈칸없는 바이그램 색인법은 주어진 문서에 빈칸을 모두 없애고 하나의 어절로 간주하여 바이그램

Table 6. 색인어 추출방법에 따른 검색결과

색인어	질의어	Avg-p	p10	Recall
명 사	TITLE	0.2135	0.4580	0.5345
형 태 소		**0.2230**	**0.4980**	0.5296
바이그램		0.2176	0.4560	**0.5502**
어 절		0.1895	0.4300	0.5231
명 사	DESC	**0.2598**	0.4920	0.6380
형 태 소		0.2463	0.4920	0.6201
바이그램		0.2566	**0.5240**	**0.6387**
어 절		0.2210	0.4780	0.5888
명 사	NARR	**0.2690**	0.5300	**0.6811**
형 태 소		0.2553	0.4980	0.6620
바이그램		0.2496	0.5280	0.6695
어 절		0.2475	**0.5340**	0.6576

Table 7. 이은단어, 빈칸없는 바이그램 분석 예

색인어	평 균
명사추출B	자율, 이동, 자율이동, 자율이동 로봇, 로봇, 관해, 로봇관해
바이그램B	자율, 율이, 이동, 동로, 로봇, 봇에, 에관, 관해, 해서

Table 8. 이은단어, 빈칸없는 바이그램 실험결과

색인어	질의어	Avg-p	p10	Recall
명사추출A	TITLE	0.2135	0.4580	0.5345
명사추출B		0.2085	0.4420	0.5343
바이그램A		0.2176	0.4560	0.5502
바이그램B		0.1971	0.4240	0.5299
명사추출A	DESC	0.2598	0.4920	0.6380
명사추출B		0.2462	0.4700	0.6369
바이그램A		0.2566	0.5240	0.6387
바이그램B		0.2402	0.4800	0.6181
명사추출A	NARR	0.2690	0.5300	0.6811
명사추출B		0.2682	0.5280	0.6814
바이그램A		0.2496	0.5280	0.6695
바이그램B		0.2279	0.4690	0.6364

Table 9. 복합명사 분해제거 실험결과

색인어	질의어	Avg-p	p10	recall
명 사	TITLE	0.2135	0.4580	0.5345
명사분해X		0.1817	0.4531	0.4373
형 태 소		0.2230	0.4980	0.5296
명 사	QUER	0.2864	0.5400	0.7321
명사분해X		0.2315	0.4780	0.6638
형 태 소		0.2813	0.5660	0.7267

Table 10.

구분	질의어	명사	바이그램	형태소
TITLE	초고정밀 화상의 의료기술로의 응용	0.1487	0.1026	**0.2117**

을 수행하는 방법이다.

Table 8에서 볼 수 있듯, 명사추출 A와 바이그램 A가 각각의 B방법에 비해 더 나은 성능을 보여준다. 기본적으로 이 두 방법은 모두 붙여 쓰여있어야 하는 어구들이 띄어져 쓰여있는 경우 발생할 수 있는 문제를 완화시키는데 그 목표가 있다. 그러나 실험결과 이러한 전략은 성능향상을 가져다 주지 못하는 것으로 나타났다. 이는 문제를 완화시키는 정도에 비해 의미없는 색인어가 생기는 정도가 훨씬 심각하다는 것을 말해준다. 따라서 문서에서 색인어를 추출할 때 빈칸으로 구분자로 하는 어절을 색인어추출의 기본단위로 보아도 크게 무리가 없다는 결론을 내릴 수 있다.

6. 복합명사 분해의 중요성

복합명사의 경우 대개 문서나 질의에서 중요한 의미로 사용될 뿐 아니라 중의적 의미를 갖게 될 가능성은 매우 희박하다. 따라서 복합명사들을 구성하는 단일명사들은 색인에서 제외하고 복합명사만을 하나의 단위로 취급하여 색인할 경우, 재현율을 다소 희생하더라도 정확도는 향상시킬 수 있을 것이라는 가설이 존재할 수 있다.

본 논문에서는 이러한 가설을 검증하기 위하여 5.와는 반대로 어절내 인접해 있는 복합명사에 대해 단일명사로의 분해를 수행하지 않을 경우의 성능변화를 알아보았다. 예를 들면 '정보검색'은 명사단위 색인방법에서는 '정보, 검색, 정보검색' 모두가 색인되지만 복합명사 분해하지 않는 방법에서는 '정보검색'만 색인이 된다.

Table 9는 이 실험에 대한 결과를 보여주고 있는데 예상과는 달리 결과를 보면 분해하지 않는 경우 재현율 뿐 아니라 상위 10문서에서의 정확도도 상당히 떨어지는 것을 알 수 있다. 특히 대부분 단일명사만을 색인하고 복합명사를 거의 색인하지 않는 형태소분석방법이 상위 10문서에서의 정확도가 가장 높았다는 점은, 복합명사를 색인하는 전략에 매우 세심한 주의가 기울여져야 한다는 사실을 알려준다. 이는 [7]에서와 같이 복합명사와 분해된 명사가 함께 색인될 경우 2중의 가중치를 부여받기 때문에 유사도에 왜곡현상이 생긴다는 주장을 뒷받침해주는 결과이기도 하다.

실험결과 고찰

3절에서의 실험결과는 명사추출방법, 형태소분석방법, 바이그램방법이 평균정확도에 있어서 모두 비슷한 결과를 보여준다. 그러나 실제적으로 이 세 방법은 질의에 따라 다른 양상을 보임을 알 수 있었다. 본 논문에서는 이 세가지 색인방법이 각각 어떠한 특성을 갖고 있는지 좀 더 자세히 알아보기 위하여 질의별 분석을 수행하였다. 질의분석의 효과를 높이기 위해서 NARR 필드의 질의는 사용하지 않고 TITLE, DESC의 질의어를 사용하였다.

질의별 검색결과는 색인방법에 따라 크게 다섯 가지 유형으로 나뉘어졌는데, 우선 형태소 분석방법, 바이그램 색인방법의 각 두 방법에 의해 좋은 성능을 보여주는 경우들과 명사추출, 형태소분석, 바이그램 색인방법에 의해 성능이 저하되는 세 가지 경우들로 나눠볼 수 있었다. 명사단위 색인은 나머지 두 색인방법에 비하여 유의미하게 높은 성능을 보여주는 질의가 존재하지는 않았는데, 대부분의 경우 명사단위 색인의 성능이 좋으면 바이그램 단위 색인과 형태소 단위 색인 중 어느 하나와 비슷한 성능을 보였다.

1. 형태소 단위 색인에서 좋은 성능을 보이는 질의

'초고정밀 화상의 의료기술로의 응용'에서는 이라는 질의의 경우 명사 및 바이그램 단위 색인법은 '초고'와 같이 낮은 빈도로 나타나지만 실제로 존재하는 명사를 인정하여

관련성없는 문서들을 검색해낸다. 또한 '정밀'과 같이 일반적으로 사용되는 명사까지 질의에 추가되어 검색성능은 더욱 낮아지는 결과를 낳았다. 반면 형태소분석결과 초고정밀이 "초+고정밀"로 분석이 됨으로서 좋은 성능을 보여줄 수 있었다.

'초'를 접두사로 판단하고, '초고'를 명사로 판단하지 않는 것은 상당히 어려운 문제인데, 이와 같이 언어지식 없이 분석하기 어려운 경우, 예를 들면 수식언, 접두사 등이 많은 경우에는 형태소 단위 색인법이 좋은 성능을 보일 수 있을 것이다.

2. 바이그램 단위 색인에서 좋은 성능을 보이는 질의들

"β-아미로이드 단백질"이라는 질의의 경우 바이그램에 의한 색인어추출법이 다른 방법에 의해 월등히 우수했다. 이는 미등록어, 특히 외국어/외래어의 경우 여러 가능한 우리말 표기(ex : 데이터/데이타, 아미로이드/아밀로이드)가 존재하기 때문인 것으로 추측된다. 이를 해결하기 위한 방법으로는 발음확장 사전을 이용한 발음확장방법[3]이 있다.

따라서 명사추출기, 형태소 분석기와 같이 언어지식을 사용하는 방법의 경우, 최대한 정확하고 높은 적용율을 보여주는 사전을 만들고 이를 유지보수할 필요가 있다. 그러나 지속적으로 생기는 미등록어들을 다룰 수 있는 사전시스템을 구축하는데는 상당한 비용을 요구한다는 것이 어려운 점이다.

이민법에서 이민, 장애인에서 장애, 성인병에서 성인은 검색성능에 영향을 주는 중요한 색인어이다. 명사, 형태소 단위 색인법에서는 이러한 색인어를 색인하지 못하여 성능이 낮아졌다.

일음절 명사가 명사 후방에 존재하는 경우 명사분해가 이루어지 않아 성능이 낮아졌지만 바이그램은 이를 보완하고 있다.

3. 명사추출에 의한 색인법에서 좋지 않은 성능을 보이는 질의들

"월드컵축구", "정치참여", "산업폐기물"은 모두 복합명사이고 각각 월드컵 : 축구 : 월드컵축구, 정치 : 참여 : 정치참여, 산업 : 폐기물 : 산업폐기물로 색인된다. 명사단위 색인법에서는 복합명사를 존재할 경우 복합명사+분해된 명사 모두를 색인하게 되어 복합명사는 질의어의 다른 색인어에 비하여 이중의 가중치를 부여받게 됨을 이미 3절에서 언급했는데, 특히 "월드컵축구"와 같이 빈번하게 발생하는 복합명사가 질의에 포함되어 있을 경우 유사도의 왜곡현상은 더욱 심각해진다고 할 수 있다. 이는 복합명사를 색인하지 않는 경우인 형태소 단위 색인법이 가장 좋은 성능을 보이는 사실로도 쉽게 알 수 있다.

Table 11.

구 분	질 의 어	명 사	바이그램	형태소
TITLE	데이터 구동화상처리	0.3663	**0.4193**	0.3855
	β-아미로이드 단백질	0.0450	**0.2696**	0.0582

Teble 12.

구 분	질 의 어	명 사	바이그램	형태소
TITLE	미국 이민법	0.2315	**0.3997**	0.2881
	장애인의 복지	0.2954	**0.3358**	0.2643
DESC	성인병의 예방방법	0.0775	**0.1379**	0.0568

4. 바이그램 단위 색인법에서 좋지않은 성능을 보이는 질의들

바이그램 색인법에 의해 추출된 바이그램들 중에는 본래 추출되어야 할 색인어의 의미와는 무관한 색인어들을 많이 발생시키는데, 이러한 색인어들 중 우연히 실제로 종종 사용되는 명사가 포함되어 있을 경우 적합하지 않은 문서들을 많이 검색하게 된다. 예를 들어 주사파의 주사, 청소년의 청소, 동점자의 점자 등은 그것이 추출된 어절의 의미와는 전혀 무관한 바이그램이지만 자주 사용되는 명사들이다.

분석결과 바이그램 색인법은 성능의 편차가 다른 색인방법에 비하여 매우 컸는데, 이는 바이그램 색인법이 질의어의 특성에 크게 좌우된다는 사실을 말해준다.

5. 형태소 단위 색인법에서 좋지않은 성능을 보이는 질의들

유통시장, 탈영사건, 석유탐사, 지방자치단체, 국제통사협력, 행정사무감사, 세계무역기구 등의 특징은 각 복합명사를 구성하고 있는 명사들이 매우 일반적인 의미를 갖는 명사일 뿐 아니라, 복합명사 자체도 상당히 자주 사용된다는 점이다. 이러한 질의의 경우 복합명사를 추출하여 색인한 명사추출기 기반 색인과 바이그램 색인법이 좋은 성능을 보이고 있는데 이와 같이, 경우에 따라서는 복합명사를 색인어에 더하여 줌으로서 좋은 성능을 보일 수 있으며, 이

Table 13.

구 분	질 의 어	명 사	바이그램	형태소
TITLE	월드컵 축구 유치	**0.6633**	0.7223	0.7365
	여성의 정치참여	**0.5470**	0.5847	0.6139
	산업폐기물 처리	**0.1153**	0.1511	0.1927

Table 14.

구 분	질의어	명사	바이그램	형태소
TITLE	주사파 파동	0.1669	**0.0838**	0.1201
	청소년 상담	0.5101	**0.4453**	0.5096
	점자 번역	0.5352	**0.4732**	0.5152
	기계번역의 평가	0.6424	**0.5846**	0.6532

Table 15.

구 분	질 의 어	명 사	바이그램	형태소
	유통시장	0.1034	0.1153	**0.0644**
	장교 탈영사건	0.5577	0.6149	**0.4411**
	석유탐사	0.3263	0.3669	**0.2883**

는 복합명사를 색인어로 인정하여 성능이 저하되었던 3.에서의 경우와는 상반된다.

결국 질의어에 따라서 복합명사의 색인 자체가 도움이 되기도 하고 그렇지 않기도 하는데, 이를 예측하는 일은 매우 어려운 문제이다. 그러나 위의 실험으로 미루어 분해된 명사가 매우 일반적인 의미를 갖는다던가 3개 이상의 명사로 구성된 복합명사는 단일 색인어로서의 가치가 있는 것으로 판단된다.

색인방법의 결합 및 검색성능의 평가

3장과 4장에 분석한 결과를 다음과 같이 정리하여 보았다.

1. 복합 명사분해

한국어는 많은 복합명사가 존재하는데, 복합명사만 색인하는 것보다, 분해된 명사를 색인하는 것이 좋은 성능을 보인다.

2. 외래어 등 추정명사의 분해 및 발음확장

외래어 등의 추정명사는 다른명사와 복합명사를 구성할 경우, 추정명사를 분해하지 못하는 경우가 발생하고, 이것이 성능을 저하시킨다. 또한 외래어는 우리말 표기시 여러 철자로 표기될 수 있으며, 같은 의미를 지니게 된다. 외래어등의 추정명사는 발음확장과 같은 조치를 해야 좋은 성능을 보인다.

3. 명사분해 후 복합명사의 첨가 또는 제거

복합명사를 명사분해한 후 복합명사를 제거할 때 성능 좋아지는 경우와 첨가할 때 성능이 좋아지는 경우가 있다.

질의별 분석결과 복합명사가 3개 이상(또는 복합명사의 길이가 6자 이상)으로 분해가 될 경우에는 복합명사가 첨가되는 것이 좋고, 그렇지 않을 경우에는 제거되는 것이 좋은 것으로 나타났다.

분석된 결과를 종합한 색인방법을 다음과 같이 제안하고 실험을 하였다.

1) 빈칸, 특수문자 등을 구분자로 하여 어절을 추출한다.

2) 어절 내에서 영문자, 숫자, 특수문자 등을 구분하여 추출한다. 추출된 영문자는 Stemming처리를 한다.

3) 추출된 어절을 품사부착기에 의해서 명사분해 하여 명사 등 실질형태소를 추출/색인한다.

4) 3)에서 추출된 실질형태소가 추정명사나 동사이면 바이그램하여 색인한다.

5) 또한 3)에서 추출된 실질형태소가 3음절이고, 명사분해가 발생하지 않은 경우 앞 2음절을 색인한다.

6) 추출된 어절을 명사추출기에 의해서 명사를 추출하고, 추출된 명사가 3)에서 추출된 실질형태소에 존재하지 않고, 길이가 6자(12바이트)이상인 경우 색인어로 추가한다.

Table 16. 제안한 색인방법에 의한 실험결과

색인어	질의어	Avg-p	p10	Recall
명 사	TITLE	0.2135	0.4580	0.5345
형 태 소		0.2230	**0.4980**	0.5296
바이그램		0.2176	0.4560	0.5502
제안방법		**0.2365**	0.4880	**0.5648**
명 사	DESC	0.2598	0.4920	0.6380
형 태 소		0.2463	0.4920	0.6201
바이그램		0.2566	**0.5240**	0.6387
제안방법		**0.2652**	0.5160	**0.6509**
명 사	NARR	0.2690	0.5300	0.6811
형 태 소		0.2553	0.4980	0.6620
바이그램		0.2496	0.5280	0.6695
제안방법		**0.2854**	**0.5560**	**0.6930**

Table 16의 결과에서 보면 제안한 방법이 다른 방법에 비하여 좋은 성능을 보이는 것을 알 수 있다.

형태소단위 색인법과 비교하여 질의별로 살펴보면, 낮은 성능을 보이는 질의는 많은 성능향상을 보였으며, 높은 성능을 보이는 질의는 비슷한 성능을 보였다. 이는 제안한 방법이 품사부착에 의한 실질 형태소 단위 색인의 문제점을 보완하여 성능향상을 가져왔음을 보여준다.

결론 및 향후 연구

본 논문에서는 명사단위 색인, 형태소 단위 색인, 바이그램 단위 색인, 어절단위 색인을 비교분석하였으며, 다양한 실험 및 질의별 분석을 통하여 검색성능에 영향을 미치는 요소을 확인하였다. 우선 복합명사는 반드시 분해하여 색인해야 하며, 외래어 등 미등록어를 포함하는 질의어의 경우 바이그램에 의한 색인전략이 효과적이었다. 또한 형태소분석기를 반드시 사용해야 단순분석이 어려운 어절에서 명사와 상당어구를 추출해 낼 수 있음을 알 수 있었다. 이러한 사실에 기반하여 각 방법을 조합하는 알고리즘을 고안하고 이를 사용하여 색인어 추출을 한 결과 성능향상에

도움이 됨을 알 수 있었다.

그럼에도 불구하고 가장 난해한 부분은 복합명사를 색인어로 추가할 것인지, 추가하지 않을 것인지를 결정하는 문제인데, 본 논문에서는 복합명사의 길이정보를 사용하여 색인어 추가여부를 결정했으나, 향후 연구로는 복합명사 처리를 위한 좀 더 개선된 방법을 고안할 것이다.

REFERENCES

1) 김상범, 한경수, 이도길, 임재수, 고명숙, 임해창(2000) : "고려대학교 정보검색엔진 HUIR의 구조 및 특징", 제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집, pp164-174
2) 이준호, 김광현, 김지승(2000) : "다양한 한글 문서 색인 방법들에 대한 평가", 제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집
3) 이승우, 조봉현, 이근배, 서정연(2000) : "P-Norm 모델에 기반한 통계적 자연언어 검색 시스템 THE IR ; 한글 TREC-1", 제5회 한국 과학기술 정보인프라(KOSTI) 워크샵 학술발표 논문집, pp189-202, 12
4) 이도길, 이상주, 임해창(2003) : "명사 출현 특성을 이용한 효율적인 한국어 명사 추출 방법", 한국정보과학회논문지 : 소프트웨어 및 응용, 제30권, 2호, pp173-183
5) Robertson, S.E et al(2000) : *"Okapi at TREC-8", In The Eighth Text Retrieval Conference (TREC -8). Gaithersburg, MD : NIST*
6) 김진동, 이상주, 임해창(1998) : "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델", 제10회 한글 및 한국어정보처리 학술발표 논문집, pp3-8
7) K. Sparck Jones, S. Walker and S.E(1998) : *Robertson, "A probabilistic model of information retrieval : development and status." University of Cambridge Computer Laboratory Technical Report no. 446*