한국산학기술학회논문지
Vol. 5, No. 1, pp. 44-48, 2004

한국어 정보검색 시스템을 위한 구 단위 색인

윤 성 희*

Phrase-based Indexing for Korean Information Retrieval System

Sung-Hee Yoon*

요 약 본 논문에서는 자연언어 처리 기술인 구문 분석 모듈을 도입해 단어 이상의 단위인 구 단위를 색인과 검색의 단위로 삼는 구 단위 색인 및 검색 기법의 사용을 제안한다. 초기의 정보검색의 방법으로 단일 주제어를 키워드로 색인하여 검색하는 방식이 널리 사용되어 왔으나 문서의 내용을 정확히 표현하기 어렵고 검색 결과의 문서 집합 또한 너무 커서 사용자의 만족도가 낮다. 고도의 문서 처리 측면에서는 웹 문서들 자체가 갖는 다양한 오류들로 인해 현실적으로 충분히 만족할 만할 우수한 성능의 구문 분석 모듈이 구현되기는 어려우므로 상향식 구문 분석 모듈을 구현하여 완전한 구문 분석 결과를 얻지 못하는 많은 문장에 대해서도 가능한 구 단위 색인을 이용하여 검색 정확률과 재현률이 향상되고 검색 과정의 처리 부하도 줄이는 장점을 얻는다.

Abstract This paper proposes a phrase-based indexing system based on the phrase, the larger syntax unit than a single keyword. Early information retrieval systems with indexing system matching single keyword is simple and popular. But with single keyword matching it is very hard to represent the exact meaning of documents and the set of documents from retrieval is very large, therefore it can't satisfy the user of the information retrieval systems. Web documents include lots of syntactic errors, the natural language parser with high quality cannot be expected in Web. Partial trees, even not a full tree, from fully bottom-up parsing is still useful for extracting phrases, and they are much more discriminative than single keyword for index. It helps the information retrieval system enhance the efficiency and reduce the processing overhead, too.

Key words : information retrieval, indexing, phrase-based, syntax analysis

1. 서 론

정보검색 시스템(information retrieval system)이 성공적으로 사용되기 위해서는 검색의 성능에 초점을 맞추어야 한다. 많은 양의 불필요한 문서보다는 꼭 필요한 문서들이 검색되기를 원하는 사용자들을 위해 문서 검색의 정확도와 검색 속도를 동시에 추구해야 하는 것이다. 많은 초기 정보 검색 시스템이 채택하고 있는 소위 단일 키워드 기반의 정보 검색 시스템은 색인 과정이 단일어의 집합으로 나타내어진다[10]. 단순히 형태소 분석 과정을 통해 명사들을 중심으로 색인하였으며, 이는 명사들이 주로 문장의 주요 의미를 나타낸다고 보기 때문이다. 그런데 이 방법은 단순히 명사의 매칭 관계로 찾고자 하는 문서의 내용을 정확히 표현하는데 한계가 있다. 문서의 양이 엄청나게 많아지는 실정에서 원하는 내용을 정확히 표현하기 어렵고 검색 결과의 문서 집합 또한 너무 커서 사용자의 만족도가 낮다. 정확도가 떨어지는 검색은 검색 결과로 의미 관계가 약한 다량의 결과 문서를 제공하게 되고, 사용자는 제공된 문서들을 다시 읽고 원하는 문서를 선별해야 하기 때문이다.

검색 시스템의 정확도를 향상시키기 위해 단어 이상의 상위 구문 단위인 구(phrase)를 이용하여 색인 시스템을 구축하는 방법을 제안한다. 단일 키워드의 매칭 빈도에만 의존하지 않고 문장을 구성하는 단어간의 관계성을 고려할 수 있게 되어 검색 정확도가 크게 향상될 수 있다.

문서의 문장들로부터 의미 관계성을 갖는 구를 알아내고 색인의 단위로 삼기 위해서는 형태소 분석 이상의 구문 분석 기술을 포함하는 자연언어 처리 기술을 도입해야 한다. 그러나 수많은 웹 문서들은 자체에 다양한 많은 오류를 포함하고 있으므로 문장에 대한 완전한 구

*상명대학교 컴퓨터소프트웨어전공
E-mail : shyoon@smu.ac.kr

문 트리(tree)를 얻을 수 있는 충분히 견고한 구문 분석이 요구되나 현실적으로는 만족할 만한 성능으로 구현되기 어렵다[6, 8]. 구 단위의 색인 방법은 완전한 구문 분석 결과를 얻지 못하는 많은 경우에도 부분 구문 분석의 결과를 이용하여 구(phrase) 단위로 색인할 수 있어서 단일 키워드 색인 방법보다 식별력이 뛰어나 검색 성능이 향상되고 검색 과정의 부하도 크게 줄일 수 있다.

2. 구 단위 색인 기법

검색 정확도를 높이기 위해 문장에서 의미적으로 밀접하게 관련된 구 단위의 단어들을 색인의 단위로 삼는 방법이다. 색인(indexing)이란 문서에 나타난 용어들의 빈도를 조사하고, 검색 모델의 가중치 부여 방법에 따라 용어들에 가중치가 부여되어, 찾기 쉬운 형태로 조직되고 저장되는 과정이다. 색인을 위한 자료구조는 흔히 가변 차수의 B-트리를 이용한 역파일 구조(inverted file)가 많이 사용된다.

보다 단순하게는 단일 단어 이상의 단위로 복합 명사만으로 대상으로 삼아 색인하는 경우도 간혹 있지만, 실질적으로는 자연언어 처리 과정에서 얻어지는 모든 종류의 실질 색인어가 될 수 있다. 본 색인 시스템에서는 다음과 같은 세 가지 종류의 색인 구를 구분한다.

2.1 형태소 분석 과정과 명사 색인

영어권의 자연어 문장들은 주로 공백으로 분리되는 단어들을 중심으로 형태소 분석이 이루어진다. 그러나 한국어는 특성 상 단어의 개념이 명확하지 않고 여러 가지 문법적 기능을표시하는 각종 접사가 결합되어 여러 가지 형태로 나타날 수 있다. 따라서 형태소 분석에서는 매우 많은 중의성(ambiguity)이 나타나서 색인으로서의 제 기능을 하지 못하는 경우가 많다. 다음과 같은 문장의 예에서 많은 형태소 중의적 현상을 볼 수 있다.

"먹이는" 먹이(N)/ 먹(N)+이 / 먹(V)+이+는
"... 겨울새들의 먹이는 작은 벌레들과 ..."
"... 먹이 많이 닳아서 ..."
"... 이런 종류의 사료를 먹인 가축들은 ..."
"... 풀을 먹이는 가축들을 사랑하고"

"종이" 종(N)+이 / 종이(N)
"... 컴퓨터가 종이 시절을 마감하고 ..."
"... 시계의 종이 울리는 ..."

"산만한" 산(N)+만한 / 산만한(AJ)
"... 산만한 구름이 덮쳐서 ..."
"... 매우 산만한 환경에서 ..."

"경기도" 경기도(N)/ 경기(N)+도
"...경기도에 사는 사람들은 ..."
"... 선수들은 경기도 관람하고 ..."

이와 같이 형태소 분석 과정에서 명사만 단일어 키워드 색인으로 추출 가능하지만 형태소적 중의성이 많이 나타나서 색인의 크기가 커지고 정확도가 떨어진다. 따라서 구문 분석의 결과로부터 구문 트리가 생성되는 경우의 형태소적 해석만 그 문서의 색인으로 채택하는 것이 바람직하다. 다음의 예는 구문 분석으로부터 타당한 문장 성분만 분석 결과의 색인으로 남는 예들이다.

"먹이"
"... 풀을 먹이(V)는 가축들은 ..."
"... 겨울새들의 먹이(N)는 작은 벌레들과 ..."
"... 이런 종류의 사료를 먹인(V) 가축들은 ..."
"... 풀을 먹이는(V) 가축들을 사랑하고"

"종이"
"... 컴퓨터가 종이(N) 시절을 마감하고 ..."
"... 시계의 종(N)이 울리는 ..."

"산만한"
"... 매우 산만한(AJ) 환경에서 ..."

복합 명사 색인은 형태소 분석의 결과로부터 색인되거나 구문 단위 인식 과정에서 색인될 수도 있다. 본 색인 시스템은 연속된 명사의 형태로 복합 명사의 처리를 형태소 분석 결과로부터 색인 단위로 추출한다[4, 11]. 복합 명사는 효과적으로 색인의 단위가 될 수 있지만 분리된 형태소들이 구문적 요소로 역할을 가질 수 있기 때문에 구문 분석 과정을 위하여 완전히 결합되지는 않는다.

(예) "컴퓨터 소프트웨어"
"컴퓨터 소프트웨어 개발"
"컴퓨터 소프트웨어 개발 과정"

2.2 명사구 색인

색인 명사구 종류	문장에서의 예
수식어를 갖는 명사구	"...컴퓨터의 뛰어난 성능 ..." (컴퓨터+성능) "... 매우 산만한 환경 ..." (산만하+환경)
접미사로 연결된 명사구	"... 사회적 혼란 ..." (사회+혼란)
병렬적 명사구	"...오류의 검출과 정정 ..." (오류+검출) (오류+정정) (검출+정정)

구문 분석 과정의 결과인 구문 트리로부터 명사구 묶음(chunking)을 실행한다[9]. 복합 명사 외에 다양한 수식을 갖는 앞의 표과 같은 형태의 명사구들이 인식되고, 색인된다.

2.3 {중심-종속} 관계의 구 색인

구문 분석의 결과는 동사나 형용사들을 중심으로 하는 종속어들의 문법적 관계를 인식하여 색인할 수 있게 한다. 중심어-종속어 관계를 보이는 구 단위의 색인 종류는 다음과 같다.

{서술어+주어}
{서술어+목적어}
{서술어+보어}
{피수식어+수식어}

다음의 문장들은 위와 같은 단위의 색인구가 추출되는 문장의 예이다.

"...컴퓨터 과학의 중요한 주제를 논할 것이다..."
{논하/주제}
"... 이 방법으로 알고리즘의 성능을 분석한다 ..."
{분석하/성능}
"... 이 기술은 오류를 검출하고 정정한다 ..."
{오류/검출하} {오류/정정하}
"... 프로그램을 실행하는 과정에서..."
{실행하/프로그램}
"... 잃어버린 정보를 되살리는 ..."
{잃/정보} {되살/정보}
"... 껍질을 까는 기계를 구입하고 ..."
{까/껍질} {까/기계}
"... 서버로 전송되는 정보들은 ..."
{전송되/서버}

앞의 형태소 분석 단계에서 나타나는 많은 수의 형태소적 중의성을 갖는 단어들이 구문 분석 과정에서 구 단위로 분석됨으로써 형태소적 중의성의 상당 부분이 해소될 수 있으며, 결과적으로 색인의 양을 크게 줄일 수 있다. 다음과 같은 문장들에서 그 예를 볼 수 있다.

"풀을 먹이(V)는 가축들은 ..."
"겨울새들의 먹이(N)는 작은 벌레들과 ..."
"사료를 먹인(V) 가축들은 ..."
"컴퓨터가 종이(N) 시절을 마감하고"
"시계의 종이 울리는 ..."

3. 시스템 구성과 자연언어 처리부

많은 웹 문서 처리 과정에서 보듯이 구문적 오류를 많이 포함하는 웹 문서들에 대해서 문법이 전체 문장을 완전히 포용하는 우수한 성능의 자연언어 구문 분석기의 구현은 현실적으로 어렵다. 따라서 구문 분석의 중간 과정에서 생성되는 구 구조들의 결과를 충분히 활용할 수 있는 순수 상향식(bottom-up) 기법을 구문 분석기 구현에 사용한다. 순수 상향식 기법은 문장 전체에 대한 분석을 얻을 수 없는 경우에도 계속 진행하여 나갈 수 있으며 결과적으로 분석이 가능한 모든 부분 분석 결과를 생성해 준다. 인터넷 정보검색에서 처리하는 문서들은 웹 문서이기 때문에 자체적으로도 많은 오류들을 포함하므로 완전 구문 분석이 실패하더라도 성공한 부분까지의 중간 결과를 이용하여 견고하고 유연한 시스템을 구현할 수 있다.

형태소 분석 과정 및 구문 분석 과정을 자연언어 처리부로 포함하는 색인 시스템은 다음과 같은 부분들로 구성된다. 다음의 그림 1에서 각 부분들의 관계를 보여준다.

3.1 전처리부

음운 부호를 인식하여 문장 단위를 분리하고, 각종

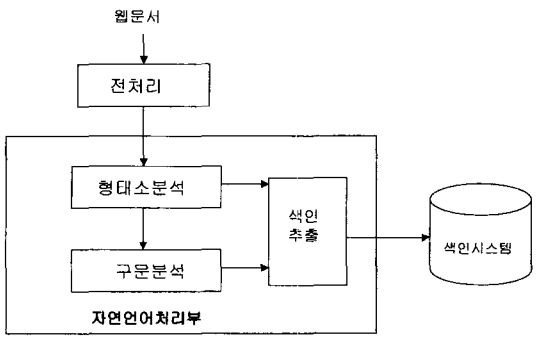

그림 1. 한국어 처리부와 색인시스템

음운 부호와 고빈도의 불용어(stopwords)를 제거하는 과정이다.

3.2 색인어 및 구 추출부

형태소 분석기의 결과로 명사 중심의 색인어 후보가 추출되고, 구문 분석의 결과로 색인구로서 문법적 관계를 갖는 단어들이 추출된다. 완전 상향식 구문 분석부는 완전 분석 성공률이 낮은 웹 문서의 문장들에 대해 중간 분석 결과를 제공함으로써 색인구의 추출이 가능하게 한다.

3.3 색인어/구 저장부

단일 명사 색인은 물론 앞서 제시한 형태의 구 구조 색인들을 색인 역파일로 생성한다. 입력된 사용자 질의문(query)은 마찬가지로 형태소 분석과 구문 분석을 거쳐 추출된 색인어 및 구 색인들을 색인 파일에서 일치시키고 정답 문서를 결과로 얻게 된다.

4. 실험 및 결과

기존에 널리 구현된 명사 중심의 단일어 색인 방법에 비해 구 단의 색인 방법의 장점은 식별력이 뛰어나서 검색 정확도가 향상된다는 것이다. 의미 관계가 깊은 구 단위를 색인으로 채택하여 보다 정확한 의미를 갖는 문서를 검색할 수 있다.

본 논문에서 제안한 구 단위의 색인 방법을 이용한 검색 시스템은 실험 단계에 있으며, 구문분석기의 성능이 가장 중요한 성공 척도가 된다. 현재까지 진행된 실험에 대한 주요 분석 사항들을 기술하면 다음과 같다.

기존의 다른 검색 시스템의 구현과 실험에 사용되기 위해 구축된 바 있는 웹문서 집합의 일부를 사용하여 실험 진행 중이다[5]. 디지털 도서관 환경에서의 검색 기법을 실험한 이 문서 집합에는 내용 상 유사한 주제와 상이한 주제들의 문서들이 고루 포함되어 있다. [2, 5]에서 사용된 문서들을 다시 구 단위 색인 방법으로 색인하여 이미 실험된 바 있는 단일 키워드 색인의 결과와 비교하였다.

구문 분석기의 분석 성공률은 약 61%이지만, 웹 문서 문장 자체가 갖는 구문적 오류에서 비롯되는 실패의 비중이 매우 높다. 완전한 구문 분석의 성공률이 낮아도, 상향식 구문분석 방법을 구현함으로써 중간 과정에서 얻어지는 많은 구 구조들이 색인에 이용되어 구문 분석 단계의 부하를 줄이고 유연한 시스템이 구현될 수 있다.

구 단위 색인을 통한 검색 성공률은 단일 명사만을 색인으로 삼은 검색 시스템에 비해 검색 성능이 향상된다. 색인 시스템의 평가와 검색 시스템의 적합성은 일반적으로 재현률(recall ratio)과 정확률(precision ratio)이라는 척도에 의하여 측정되는데, 그 의미는 다음과 같다.

재현률=검색된 적합 문서의 수/전체 적합 문서의 수
정확률=검색된 적합 문서의 수/검색된 전체 문서의 수

실험 중인 현재의 중간 결과에 의하면 단일 명사 단위의 색인에 비해 검색 결과 문서의 양이 32% 감소하여 검색 정확도가 향상됨을 보여주었다. 단일 키워드의 일치만으로 문서를 검색하는 방법의 경우에 상대적으로 다량의 문서를 검색 결과 문서 집합으로 제공한다는 것은 검색 결과의 문서 중에 검색 요구된 정확한 문서의 범위에 들지 못하는 것이 상당량 포함되어 있으며 사용자가 직접 문서들을 읽어서 선별해야 하는 상황이 된다. 이에 비해 문장을 구조적으로 분석하여 문장의 정확한 의미를 추출하는 구 단위 검색 과정은 사용자가 읽어서 배제해야 하는 상당량의 검색 결과 문서를 선별할 수 있다. 검색 시스템의 성능은 검색 요구에 적합한 문서를 모두 검색하는지 여부와 동시에 부적합한 문서는 검색하지 않는 적합성, 그리고 응답 시간, 경제성에 의하여 평가된다. 적합성에 관한 정의는 명확하게 기술하기 어려우나 검색된 정보 자료와 질의문과의 일치되는 정도를 의미한다[10]. 지금까지 [5]에서 사용된 바 있는 실험 문서들을 본 논문에서 제시한 구 단위 색인 방법으로 색인하여 실험 진행하였으나, 충분한 성능 평가의 결과를 제시하기 위해 웹 문서의 범위를 보다 넓게 확장하여 추가의 실험을 계속하고자 한다.

5. 응용 및 계속 연구의 방향

검색 시스템의 정확도를 향상시키기 위해 단어 이상의 상위 구문 단위인 구(phrase)를 이용하여 색인 시스템을 구축하는 방법을 제안하였다. 단일 키워드의 매칭 빈도에만 의존하지 않고 문장을 구성하는 단어 간의 문법적 관계를 고려함으로써 검색 정확도가 크게 향상되는 방법이다. 문서의 문장들로부터 특정 구조의 의미 관계를 갖는 구를 알아내고 색인의 단위로 삼기 위해서는 형태소 분석 이상의 구문 분석 기술을 도입하였다. 웹 문서들을 충분히 정밀하게 구문 분석하는 구문 분석기는 구현과 처리 과정의 부담이 매우 크다. 따라서 완전한 구문 분석 결과를 얻지 못하는 많은 경우에도 부분 결과를 이용하여 구(phrase) 단위로 색인할 수 있도록 하여 검색 성능이 향상되고 검색 과정의 부하도 크게 줄일 수 있었다.

이미 조사된 바에 의하면 두 사람이 잘 알려진 객체에 대해 같은 키워드를 선택할 확률은 20%가 채 되지

않는다고 한다. 따라서 의미적으로 동일한 질의나 문서의 내용을 처리하는 과정이 검색 성능을 크게 향상시킬 수 있을 것이다. 이는 동일 의미를 갖는 단어들을 처리하는 방법으로부터 문장 구조 이상의 의미 분석 과정의 통해서 처리될 수 있으나 처리 과정의 부담을 고려하여 시스템으로의 확장 구현 여부를 결정해야 할 것으로 보인다[7]. 다음 예들은 구문 분석의 결과에서 얻는 구단위 색인으로 해결하기 어려운 다양한 현상을 보여준다.

- 같은 의미를 가지는 문장이 서로 다른 통사적 형태로 존재할 수 있다.
- 동일한 의미를 갖거나 비슷한 의미를 가지는 문장이 서로 다른 표현 방법으로 나타날 수 있다.
- 완전히 같은 의미는 아닐지라도 의미적으로 매우 가까운 키워드를 포함하는 문서들을 색인하는 기술을 도입할 수 있다. 예를 들어 "암호" "보안" "인증" 등의 단어들은 의미적으로 유사한 내용의 문서들일 가능성이 매우 높다.

자연언어 문장이 본질적으로 갖는 중의성으로 인해 색인의 양이 많아지는 것으로 파악되어, 구문 분석에서 해결되는 이상의 의미 분석이 요구되는 중의성의 해결 방법과 그에 따른 처리 부하에 대해서도 계속 연구할 필요가 있다[3]. 현재까지는 구문적 요소에 기반하여 색인하고 문서의 의미를 추출하였으며 그 결과 검색 결과 정확도가 상당히 향상됨을 보았다. 더 나아가서는 통계적 정보[1, 7]와 의미 분석을 동반하는 연구가 검색 성능을 향상시키는 방법이 될 것으로 보이므로 위에 제시한 바와 같은 여러 경우들을 분석하여 계속되는 연구의 방향으로 삼고자 한다.

참고문헌

[1] 김상범, 이상주, 홍금원, 임해창, "한국어 정보검색에서 위치관계에 기반한 통계적 구 색인", 제14회 한글 및 한국어 정보처리 학술대회 논문집, 76-82, 2002.

[2] 맹성현, 이석훈, 송사광, 박혁로, "정보 검색 시스템 평가를 위한 균형 테스트 컬렉션 구축", Proc. of KOSI, 1998.

[3] 박세영, 강현규 "한글공학 : 정보검색", 한국정보처리학회 특집, 제5권 제5호. 1998.

[4] 원형석 외, "복합명사 분할과 명사구 합성을 이용한 통합 색인 기법", 정보과학회 논문지:소프트웨어 및 응용, 제27권 제1호, 2000.

[5] 윤성희, "디지털 도서관 환경에서의 정보검색을 위한 자연어 문서 및 질의 처리기에 관한 연구", 한국컴퓨터산업교육학회 논문지, 2000. 제2권 제12호, 2001.

[6] 이충희 외, "구 기반 색인 시스템의 구현", 제14회 한글 및 한국어 정보처리 학술대회 논문집, 63-69, 2002.

[7] 장명길 외, "의미기반 정보검색", 정보과학회지 10월호 한글정보처리 특집, 2001.

[8] A. T. Arampatzis, T. Tsoris, C. H. A. Koster and Th. P. van der Weide, "Phrase-based Information Retrieval," Journal of information Processing & Management, vol. 34, Issue 6, 1998.

[9] Jose Perez-Carballo and Tomek Strazalkowski, "Natural Language Information Retrieval : progress report", Information Processing & Management, Vol. 36, Issue 1, 2000.

[10] R. Baeza-Yates, and B. Reberio-Neto, "Modern Information Retrieval", Addison Wesley, 1999.

[11] Won, H., Park, M. and Lee. G. "Integrated indexing method using compound noun segmentation and noun phrase synthesis", Journal of KISS : Software and Applications, vol. 27, no. 1, 2000.