論文
5-9-30

웹 문서 검색을 위한 검색어 추출과 확장에 관한 연구
A Study on Keyword Extraction and Expansion for Web Text Retrieval

윤성희(Sung-Hee Yoon)[1]

요 약

웹 문서 검색 시스템 사용자에게 자연어 질의를 입력하는 방법은 가장 이상적인 인터페이스이다. 본 논문은 자연어 질의를 입력하는 웹 문서 검색 시스템을 위해 자연어 처리 기술에 기반하여 사용자의 입력 질의 문장을 구문 분석한 후 검색어를 추출하고 확장하는 다중검색 기법을 제안한다. 질의문에 대한 형태소 분석 및 구문 분석을 수행하고, 구문 트리를 순회하여 구조적으로 연관된 복합명사를 조합하거나 분할하며, 검색어가 되는 음역어와 축약어들을 확장하여 다중검색함으로써 재현율과 정확도를 향상시킬수 있음을 보였다.

Abstract

Natural language query is the best user interface for the users of web text retrieval systems. This paper proposes a retrieval system with expanded keyword from syntactically-analyzed structures of user's natural language query based on natural language processing technique. Through the steps combining or splitting the compound nouns based on syntactic tree traversal, and expanding the other-formed or shorten-formed keyword into multiple keyword, it shows that precision and correctness of the retrieval system was enhanced.

Key Words: web test retrieval, natural language query, keyword extraction, keyword expansion

논문접수 : 2004. 11. 03.
심사완료 : 2004. 12. 15.

1) 정회원: 상명대학교 컴퓨터소프트웨어공학전공

1. 서 론

웹 정보 검색 시스템은 입력 질의와 문서 내용에 대한 검색어의 형태적 일치를 검사함으로써 적합 문서 여부를 결정하며, 많은 웹 문서들을 주기적으로 방문하고 색인 데이터베이스를 갱신한다. 또한 수많은 상업용 정보 검색 시스템들은 단순 키워드뿐만 아니라 불리언(boolean) 연산식의 형태로 입력할 수 있도록 하지만 실험에 의하면 대부분의 일반 사용자는 불리언 연산식 표현에 익숙하지 않고, 매우 단순하게 적은 수의 검색 키워드 입력을 통해서 검색을 수행하는 경향이 뚜렷하다. 웹 정보 환경은 문서들이 빈번하게 생성되고 소멸되는 매우 유동적인 환경이며, 검색 대상 문서들은 양적으로 방대하고 형식, 주제, 생산자 등에 있어서도 매우 이질적이다. 더욱이 웹 환경 검색 시스템의 사용자들은 다양한 계층 및 연령층이면서도 트랜잭션 로그 분석을 이용한 연구의 결과를 통해 볼 때에는 "웹 검색에 있어서 사용자의 검색 방식의 단순함"이 두드러진 특징으로 나타난다[8].

사용자가 검색 의도를 가장 정확하게, 또 가장 편리하고 자연스럽게 표현할 수 있는 인터페이스(interface)는 자연어 문장을 검색 질의문으로 입력하는 방법이다. 자연어 인터페이스의 검색 시스템은 질의문을 자연어 처리 기술에 따라 처리하는 기능을 필수적으로 수반해야 한다. 구문 분석 과정을 배제한 자연어 질의 검색 시스템은 대부분 형태소 분석, 키워드 추출, 불용어(stopwords) 제거 등의 과정을 갖지만, 통계적 색인과 검색에 비해 보다 정교한 정보 검색을 위해서는 형태소 분석 이상의 구문 분석(syntactic analysis)과 같은 고수준의 자연어 처리 기술의 도입이 필수적이다. 본 논문은 형태소 분석 및 구문 분석하는 자연어 처리 기술 기반의 질의 처리와 검색 방법을 제안한다. 색인 데이터베이스와 질의문에서 색인어 일치, 불일치를 보다 정교하게 처리하기 위해 구문 트리로부터 복합명사의 조합 형태와 분할 형태로 확장하여 다중 검색어로 추출한다. 뿐만 아니라, 축약어들과 음역어의 이형표기들에 대해서도 마찬가지로 검색어를 확장하는 다중검색 방법을 제안한다.

2. 연구 배경

자연어 질의문의 처리는 형태소 분석 및 구문 분석 결과에 기초하여 다양한 문법적, 의미적 관계를 분석하고 유지함으로써 단순하게 형태소 분석으로부터 추출된 검색어들의 집합으로 취급할 때에는 불가능한 관계성을 파악할 수 있다. 검색어의 추출이 검색 엔진의 성능을 결정하는 기본적인 요소임에도 불구하고 정보 검색 연구 분야에서는 색인어와 색인 기법이 검색 결과에 미치는 영향이 검색 모델에 비해 매우 중요함에도 불구하고 상대적으로 중요시되지 않았던 것으로 보인다[6]. 관련 연구 [2]에서는 통계적인 방법과 구문 구조를 이용한 색인의 비교에서 복잡한 자연어 처리를 이용한 방법보다는 방법론이 간단한 통계적인 방법을 선호했지만 통계적 방법과 비교된 구문 구조 분석 방법은 적용 과정이 지나치게 단순하여 적절한 비교가 되지 못한다. 구를 합성하여 색인어로 사용하여 실험 집합에 대해 평균적으로 10% 정도의 정확도를 향상시켰지만 실험 집합에 따라 성능의 차이가 크고, 이용되는 요소들이 대상 도메인에 따라 반복 실험에 의해 구해져야 하는 문제점을 보였다. 관련 연구 [1]은 대규모의 문장을 처리할 수 있는 강건한 자연어 처리기를 사용하기 위해 완전 구문 분석기 대신 명사구 분석기(NP Phrase parser)를 이용하여 명사구를 합성하는 방법으로 정확도와 재현률을 향상시켰다.

기존 검색 시스템 중 일부는 검색 엔진에서 한글 문서 내용을 한글의 교착어적 특성을 반영하여 형태소 분석을 통해 색인한다. 한국어의 문법 형태소 조사/어미/접미사'를 형태소 분석기에 의해 분리한 후에 각 형태소를 색인어로 추출한다[6]. 그런데, 탈락과 불규칙 활용 등 변형된 어간이나 변형된 어미는 원형을 복원하지는 않고, 변형된 어간의 경우에 입력 어

질에서 어간의 길이만큼 어절 단위로 절단하여 색인어로 추출한다. 결과적으로 어미, 의존명사, 조사 등의 형태소들이 색인어나 검색 키워드로 추출되면서, 형태소 분석 후 색인하는 과정이 항상 더 정확한 결과를 얻게 되는 것은 아님을 볼 수 있다. 그림 1에서 기존 웹 문서 검색 시스템에서의 자연어 질의문장의 일부와 검색 결과 문서의 내용을 보이고 있다. 대부분의 시스템들은 키워드 입력을 전제로 하기 때문에 색인어로서 의미가 없는 형태소적, 문법적 요소들을 분리하고 적절한 색인어나 검색어를 추출하는 과정이 뒤따르지 않았기 때문이다.

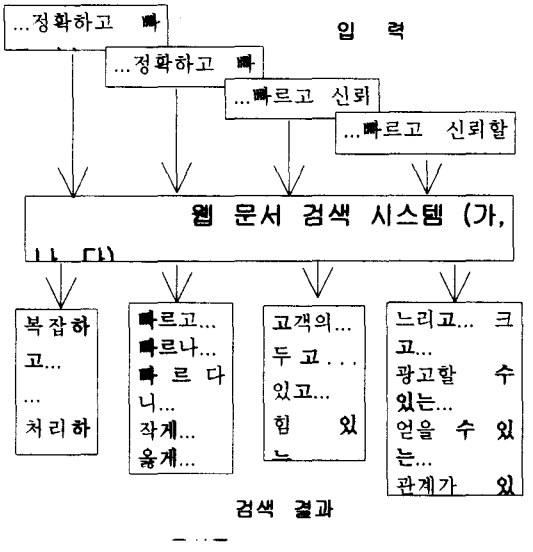


그림 1. 기존 웹 검색 시스템과 자연어 질의 검색 결과 예

3. 웹 문서 검색 시스템의 설계

본 논문에서 제안하는 자연어 질의문에 대한 구문 분석 및 검색어 확장에 의한 다중검색과 문서 색인부의 관계를 그림 2에서 보이고 있다.

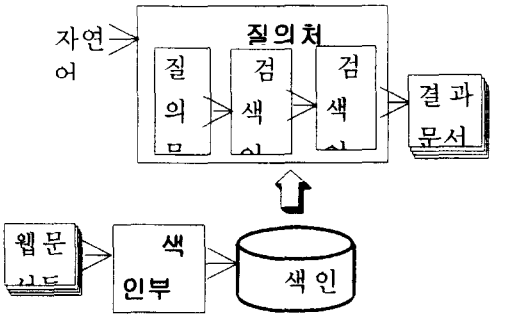


그림 2.웹 문서 검색 시스템의 구조

자연어 질의문의 형태소 분석과 구문 분석의 결과로서 구문 트리(syntactic tree)를 얻고, 트리 순회를 통해 복합명사를 조합하거나 분할하여 검색어를 확장한다. 또한 이형표기 용어들과 축약표기 용어들에 대해서도 질의를 확장하여 다중 검색한다.

3.1 구문 형태소 추출

입력한 자연어 질의문은 형태소 분석, 불용어의 제거, 태깅(tagging) 과정을 우선적으로 거친다. 자연어 문장에서 연속적으로 나타나는 형태소 패턴들 중에는 빈번하게 출현하면서도 구문 분석 과정에서 생성하게 될 문법적 구조가 크게 의미가 없는 관용적 패턴의 성격을 갖는 것들이 있는데, 이들을 구문 형태소열로 정의한다[10]. 구문 분석 과정에 앞서 구문 형태소 열들을 인식하고 단위 구문 요소로 결합하여, 양상 정보를 추출하는 전처리 과정을 도입함으로써 질의문 문법 구조 분석 과정의 효율을 높일 수 있다. 다음의 예들은 구문 형태소로 정의되고 전처리될 수 있는 연속된 형태소열들 중에서 빈도수가 높은 대표적인 예들이다.

구문 형태소열 예

"-애/jca 대하/pvg ㄴ/etm" "-에/jca 대하/pvg 어/ecs" "-을/jco 위한/pvg" "-ㄹ/etm 수/nbn 있/pvg" "-ㄹ/etm 수/nbn

없/paa"

구문 형태소열은 문장에서 내용어의 역할을 수행하지 않고, 기능어의 역할을 하므로 다른 숙어나 관용어에 비해 문장에서 자주 발생하며 다양한 체언이나 용언에 이 구문 형태소열들이 결합할 수 있기 때문에 언어 생성력이 매우 크다. 코퍼스(corpus)에서 일정 이상의 빈도를 가지며, 한 어절 이상에 걸쳐 기능 형태소들을 대상으로 체언이나 용언에 구조적, 의미적 영향을 미치는 상위 빈도 어절로서 그 자체로 하나의 문법 범주가 될 수 있다. 관련 연구 [10]은 관용구 관점에서 빈번하게 출현하는 구문 형태소들의 유형들을 보조 용언을 매개로 한 구문 형태소, 의존 명사를 매개로 한 구문 형태소, 의사 조사, 기타 강한 어휘적 공기 관계를 갖는 형태소 등의 체계로 분류하고, 그 구체적인 예들에 대한 통계적 연구 결과를 제시하고 있다. 구문 분석 과정에서 문법 규칙들의 적용에 앞서 구문 형태소열들을 단일 문법 요소로 결합시키는 전처리 과정은 문장의 구조적 모호성을 크게 감소시킴으로써 전체 검색 시스템의 계산 성능을 향상시킬 수 있다.

3.2 자연어 질의문의 구문 구조 분석

검색부는 형태소 분석과 태깅, 구문 형태소열의 인식 과정에 이어서 구문 분석 과정을 거쳐 단어들의 문법적 관계를 얻기 위해 문장의 문법 구조를 해석하고 그 결과로 구문 트리를 생성한다. 자연어 질의문의 구문 분석 과정은 단순한 형태소 열이 나타낼 수 없는 중요한 구조적, 의미적 관계를 추출할 수 있는 과정으로서, 자연어 처리에서 가장 핵심적이고도 복잡한 처리 과정이다[9,10]. 그림 3은 자연어 질의 입력 예문의 구문 분석에서 얻어지는 구문 트리의 예이다.

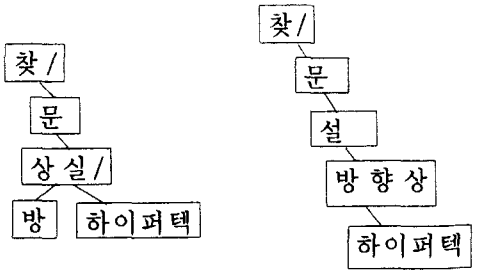


그림 3 자연어 질의문의 구문 구조 분석

자연어 문장을 구문 분석하는 방법은 통계적인 방법, 구문 규칙에 기반한 방법, 학습에 의한 방법 등 여러 가지가 있으나, 본 연구에서는 이미 여러 실험에서 적용된 바 있는 문맥 자유(context free) 형식 구 구조(phrase structure) 규칙을 기반으로 하는 상향식(bottom up) 분석 방법을 채택한다. 상향식 분석은 전체 문장의 구조 해석에 실패하더라도 부분적으로 완성된 중간 구조를 활용할 수 있는 큰 장점이 있기 때문에 자연어 문장의 구문 분석과 같이 완전 분석 성공률이 낮은 경우에도 유용한 레코드를 얻을 수 있다. 일부의 규칙을 이용하여 부분 분석을 우선으로 실행하고 그 결과들로부터 전체 분석을 진행하는 구문 분석 과정이 매우 효율적임을 보인 연구가 있다[9]. 또한 규칙 기반의 구문 분석 방법은 통계적 방법에 비해 문재 분야에 영향을 받지 않고 여러 응용에서 널리 사용될 수 있다.

3.3 복합명사와 검색어 추출

관련 조사 및 연구 [4,8]에 의하면 "한국어의 질의어는 거의 명사 유형이다"라고 한다. 명사가 검색 사용자들의 질의 의도를 나타내는 중심이 된다는 뜻이다. 검색에서 의미의 중심 역할을 하는 명사들은 색인이나 검색에서 복합명사의 다양한 형태로 빈번하게 나타난다. 복합명사의 문법적 규정은 매우 약하고 띄어쓰기 문제와도 긴

밀하게 관련되어 있다. 두 개 이상의 명사를 붙여 쓴 형태이거나 띄어 쓴 두 개의 단일 명사이거나 대부분의 경우 모두 문법적으로 오류가 되지 않는다.

KTSET을 중심으로 문서와 질의문에서 나타나는 복합명사들의 길이에 대한 통계 자료를 참조하면 세 단어까지의 복합명사가 전체의 96.8%를 차지하며, 네 단어 이상의 복합명사는 매우 드물게 나타난다는 실험 결과가 있다. 반대로, 한국어에서 네 음절 이상의 단일어는 극히 드물기 때문에 네 음절 이상의 명사라면 복합어일 확률이 매우 높다고 본다[4].

본 연구는 동적 웹환경의 부하를 고려하여 색인 데이터베이스는 형태소 분석과 태깅 결과로 나타나는 색인어들로 색인하고, 입력 질의문에 나타나는 명사 색인어들에 대한 복합명사를 분할하고 조합한 결과로 색인 데이터베이스를 다중검색하는 방법을 우선 적용하고 실험하였다. 또한 이미 구축되어 사용되고 있는 상용의 검색 시스템에 대한 자연어 메타 검색기로서 자연어 질의문을 입력할 수 있는 사용자 환경을 제공하고, 입력 질의문의 문법적 구조를 분석하여 복합명사를 조합하고 분할한 결과를 검색어를 추출하여 검색 시스템에 제공할 수도 있다.

3.3.1 복합명사 조합과 검색어 확장

구문 트리를 순회하면서 문법 구조적으로 연관된 어휘들을 추출하여 복합명사 형태로 조합하여 색인어로 확장하며, 이들을 검색어로 다중검색한다. 구문 트리에서 두 개 이상 명사가 복합명사로 조합되어 검색어로 확장될 수 있는 구조적인 예들은 다음과 같다.

복합명사 조합의 경우

- 조사 생략된 인접 명사
- 관형격 조사로 연결된 피수식 명사
- 목적격/주격조사로 연결된 서술형 명사
- 피수식 내포문과 명사
- 관형화된 내포문과 명사

복합명사 조합 예

성능을 분석하고 평가하는 / 성능의 분석과 평가	성능분석 / 성능평가
통계 정보 / 통계적 정보 / 통계의 정보	통계정보
실험 결과의 분석 / 실험한 결과의 분석 실험한 결과를 분석 / 실험 결과에 대한 분석	실험결과 분석

형태소 수준의 복합명사 조합 방법이 의미 없는 조합을 많이 생성하는 문제가 있는 반면, 본 연구의 방법과 같이 구문 트리의 구조에 기반한 복합명사 조합은 사용자의 검색 의도를 반영한 조합이 가능하다. 두 단어 이상의 복합명사 확장 방법은 기존의 연구 방법을 따르며, 네 단어 이상의 복합명사는 실제로 사용 빈도가 극히 낮고 오히려 검색 성능을 떨어뜨리는 경향이 있으므로 세 단어까지 조합하는 것이 효율적이다[4]. 다음의 그림 4는 구문 트리로부터 조합되는 복합명사와 띄어쓰기에 따라 조합되고 확장될 수 있는 예이다.

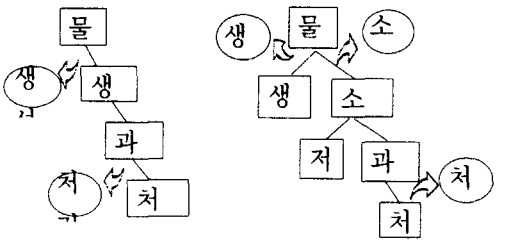


그림 312. 구문 트리와 복합명사 처리

3.3.2 복합명사 분할과 검색어 확장

네 음절 이상의 단일명사는 복합명사인지 여부를 검사하고 단일어 추출과 띄어쓰기의 분할 과정을 거쳐서 입력 질의에 대한 검색어를 확장하고 다중검색을 한다. 앞에서 설명한 바와 같은 이유로 복합명사는 두 개 또는 세 개 단일명사들

의 조합으로 분할한 후 색인 데이터베이스를 다중 검색한다. 이때 분할된 각 단일명사들은 모두 색인 데이터베이스에 독립적으로 등록된 색인어이거나, 접두사 또는 접미사이어야 한다. 분할된 단일명사들은 복합명사에 비해 특정성이 떨어질 수 있지만 실험에서 보이는 바와 같이 검색 결과의 재현률을 높일 수 있다.

복합명사 분할 예

통계정보	통계 정보
실험결과분석	실험 결과 분석 / 실험결과 분석 / 실험 결과 분석
병렬처리시스템	병렬 처리 시스템 / 병렬처리 시스템 / 병렬 처리시스템
음성인식	음성 인식

3.4 음역 및 축약 검색어 확장

음역어에서 빈번하게 나타나는 이형표기의 검색어들을 질의 구문 트리로부터 추출하고 확장하여 이들을 검색어로 다중 검색한다. 빈도수 높은 이형표기 용어 집합이나 축약표기 용어 집합들이 별도의 사전으로 등록되어 처리되면 사용자의 검색 만족도를 높일 수 있다.

이형동의 및 음역어 검색어의 예

불리언 / 부울린 / 불린(n)
데이터베이스 / 데이터베이스
운영체제 / 운영체계
자연어 / 자연언어
모호성 / 중의성 / 애매성
알고리즘 / 알고리듬 / 앨고리듬 / 앨고리즘

검색 성능을 높이기 위해 질의 확장될 수 있는 또 다른 예는 축약표기 용어이다. 축약표기 용어들은 또 다른 종류의 이형표기 용어로 처리할 수 있다. 축약표기 용어와 이에 대응하는 기본 색인어들은 이형표기 용어 사전에 등록되어 질의문에 입력된 검색어들을 양방향으로 확장하고 다중검색 한다.

축약표기 검색어의 예

헌재 / 헌법재판 / 헌법재판소
정통부 / 정보통신부
노조 / 노동조합

기존의 여러 검색 시스템에서 위와 같은 이형표기 용어들이나 축약표기 용어들을 검색어에 포함시켜 검색을 시도하면 검색 결과 문서의 집합은 모두 상이함을 볼 수 있다. 이러한 용어들의 처리는 태깅의 결과에 적용하는 방법도 생각할 수도 있지만, 구문 구조 분석 과정에 기반하는 검색 방법에서는 형태소적 모호성과 구조적 모호성이 크게 해소된 후에 검색어를 확장하는 것이 훨씬 효율적이다. 예를 들어, "불린"은 "불린 연산식", "물에 불린 미역, "바보라고 불린 청년"등의 모호성을 가질 수 있다. 따라서 문법적 구조 분석의 결과를 얻은 후에 이형표기 용어나 축약표기 용어들을 추출하고 다중 검색하는 방법이 타당하다. 다음의 그림 5는 질의문 분석에 기반한 검색어 추출과 확장 과정을 보인다.

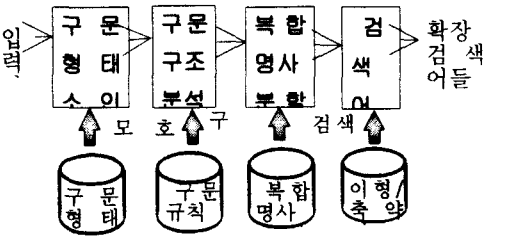


그림 313 질의분석과 검색어 추출 및 확장 과정

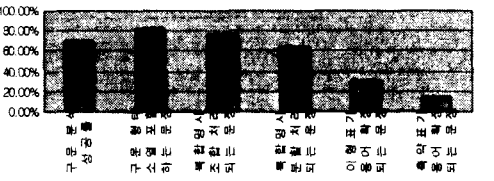
4. 실험 결과 및 평가

정보 검색 시스템의 성능은 일반적으로 재현율(recall)과 정확도(precision)라는 척도로 평가된다. 분할된 단일 명사들은 복합명사에 비해 특정성이 떨어지게 되지만 검색에서 재현율

을 높일 수 있고, 문장에 나타나는 단어들로 명사구를 조합하게 되면 특정성이 큰 색인어를 만들게 되어 정확도를 높일 수 있다. 본 연구의 실험은 여러 실험에서 사용된 바 있는 KTSET 문서[5]들과 개인, 학과, 부서, 학교, 기관 등의 홈페이지들, 학술 내용과 시사 뉴스 등의 내용을 담은 웹 문서들을 대상으로 하였다. 제한 없는 임의의 검색자들로부터 한국어 자연어 질의문들을 수집하여 약 4,900여 질의문을 대상으로 본 연구의 방법을 실험하였다. 입력된 자연어 검색 질의문의 두드러진 특징으로 문장의 길이가 일반 문서의 문장에 비해 비교적 짧고, 복합명사들이 다양한 형태로 빈번하게 출현한다는 점을 볼 수 있었다. 입력된 질의문의 구문 형태소 처리를 거쳐 구문 분석하는 과정은 일반 문서들의 문장을 구문 분석하는 과정에 비해 성공률도 비교적 높아서 68.76%로 나타났으며, 분석 성능도 향상되었다. 일반 웹 문서에는 길이가 매우 긴 문장들이 많고, 문법적으로 오류를 많이 포함하는 특성에 비해 사용자의 질의문은 상대적으로 단순하기 때문인 것으로 분석된다. 각 단계의 검색어 확장 과정을 거치면서 입력 질의문에 대한 검색어의 개수는 검색어 확장을 하지 않았을 때와 비교하여 2.72배로 증가하였다.

총 질의문 수 : 4,894 문장	
구문 분석 성공률	68.76%
구문 형태소열 포함하는 문장	81.44%
복합명사 조합 처리되는 문장	77.31%
복합명사 분할 처리되는 문장	64.28%
이형표기 용어 확장되는 문장	29.52%
축약표기 용어 확장되는 문장	13.54%
평균 검색어 확장 비율	2.72 배

그림 5. 검색어 확장 실험 결과



검색 성능으로 볼 때에는 검색어 확장 과정을 전혀 거치지 않았을 때보다 재현율이 약 11.3%가량의 향상을 보였다. 특히 이형표기 용어들이나 축약표기 용어들에 대한 검색어 확장은 재현율의 향상에 크게 기여하고 있다. 전반적으로는 정확도는 약 4.7% 정도로 다소 높아졌으나, 일부 검색에서는 거리가 먼 명사들의 조합과 의미 없는 분할이 일어나는 경우도 간혹 나타나고 있어서 이에 대한 정밀한 분석이 필요한 것으로 보인다. 다양한 질의를 추가 수집하고, 오류 현상에 대한 유형별 분석과 대응 방법의 연구가 요구된다. 또한 메타검색기로서 적용과 검색 결과 문서의 가중치 계산에 대한 연구가 반영되는 것이 필요하며, 음역 이형태와 축약형 검색어들에 대한 통계적 수집 시스템을 구축할 필요가 있다.

5. 결 론 및 제언

본 논문에서는 웹 문서 검색 시스템의 이상적 인터페이스로서 한국어 자연어 질의문을 입력하고 분석하여 검색어를 추출하고 확장하는 처리 과정을 제안하였다. 형태소 분석과 구문 분석을 비롯한 기본적인 자연어 처리 과정을 포함하며, 명사 검색어뿐만 아니라 문장의 구조가 전달하는 정확한 문법 관계를 얻기 위해 구문 분석 과정도 포함된다. 형태소 분석 및 구문 형태소 추출과 구문 트리를 순회하면서 복합명사를 분할하거나 조합하고 이형태나 축약형 검색어들을 확장하는 과정을 포함한다. 입력 질의문의 구문 분석 결과를 기반으로 검색어들 사이의 계층적 구문 구조가 전달하는 의미 관계를 획득함으로써 정확한 검색어를 추

출하는 것이 목적이다. 각종 문법적 요소들의 분해가 단순히 색인어나 검색어가 되는 것이 아니며, 의미 관계가 정의되는 명사들을 인식하여 복합명사로 조합하거나 분할한다. 또한 이형표기 및 축약표기 검색어들의 확장을 통해 다중 검색함으로써 웹 검색 시스템의 재현율과 정확도를 높일 수 있음을 실험 결과로 보였다.

참 고 문 헌

[1] Gerard Salton, "Automatic text processing," Addison-Wesley publishing company, 1988

[2] Joel L Fagan, "The effective of a non-syntactic approach to automatic phrase indexing for document retrieval," JASIS, 1989.

[3] Lee, G., Park, M., and Won, H., "Using syntactic information in handling natural language queries for extended boolean retrieval model", In proceedings of the 4th international workshop on information retrieval with Asian language(IRAL99), pp.63-70, 1999.

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

[5] KIBS : Korean Information Base System, http://kibs.kaist.ac.kr

[6] 강승식, "한글 문서의 색인어와 색인 기법," 정보과학회지, 제22권 제4호, pp.72-77, 2004.

[7] 강현규, "개념 검색어 확장을 통해 질의 형식화를 도와주는 개념 마법사의 설계 및 구현", 정보처리학회논문지 제9-B권, 제4호, pp. 437-444, 2002. 8.

[8] 박소연, 이준호, "로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구", 정보관리학회지 제19권 제3호, pp.111-122, 2002.

[9] 이공주, 김재훈, "규칙에 기반한 한국어 부분 구문분석기의 구현", 정보처리학회논문지, 제10권 B제4호, pp. 389-396, 2003.

[10] 황이규, 이현영, 이용석, "형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용", 정보과학회논문지 제27권 제7호, 2000.