자연어 질의 분석과 검색어 확장에 기반한 웹 정보 검색*

Web Information Retrieval based on Natural Language Query Analysis and Keyword Expansion

윤 성 희(Sung-Hee Yoon)**
장 혜 진(Hai-Jin Chang)***

초 록

웹 문서 검색을 위해 키워드와 불리언 연산식을 사용하는 것에 비해 자연어 질의 문장을 입력하는 방법은 검색 시스템 사용자에게 훨씬 이상적인 인터페이스이다. 본 논문은 사용자가 입력하는 자연어 질의 문장을 구문 분석하고 그 구문 구조에 기반하여 검색어를 확장하는 다중 검색 기법을 제안한다. 구문 트리를 순회하여 구조적으로 연관된 복합 명사를 조합하거나 분할하는 과정을 거치고, 이형 표기 및 축약 표기 용어들에 대해 확장 다중 검색함으로써 웹 정보 검색 시스템의 재현율과 정확도를 높일 수 있다.

ABSTRACT

For the users of information retrieval systems, natural language query is the more ideal interface, compared with keyword and boolean expressions. This paper proposes a retrieval technique with expanded keyword from syntactically-analyzed structures of natural language query as user input. Through the steps combining or splitting the compound nouns based on syntactic tree traversal of the query, and expanding the other-formed or shorten-formed into multiple keyword, it can enhance the precision and correctness of the retrieval system.

키워드: 웹 정보 검색, 자연어 질의, 구문 분석, 검색어 확장
information retrieval, natural language query, syntactic analysis, keyword expansion

---

* 이 논문은 상명대학교 2004년 교내연구비 지원의 결과임
** 상명대학교 컴퓨터정보통신공학부 부교수(shyoon@smu.ac.kr)
*** 상명대학교 컴퓨터정보통신공학부 부교수(hjchang@smu.ac.kr)
■ 논문접수일자 : 2004년 5월 27일
■ 게재확정일자 : 2004년 6월 14일

1. 서 론

정보의 범람이라고 불릴 만큼 다양하고 방대한 양의 웹 문서들의 집합에서 사용자가 원하는 내용을 담고 있는 문서들을 선별해 주는 것이 정보 검색(information retrieval) 시스템의 역할이다(Baeza-Yates Ricardo and Reberio Neto Berthier 1999). 일반적으로 정보 검색 시스템은 입력 질의와 문서 내용에 대한 색인어의 형태적 일치를 검사함으로써 적합 문서 여부를 결정하며, 많은 웹 문서들을 주기적으로 방문하고 색인 데이터베이스를 갱신한다(강승식 2004). 또한 수많은 상업용 정보 검색 시스템들은 검색 정확도를 높이기 위한 방법으로 사용자의 질의를 단순 검색 키워드뿐만 아니라 불리언 연산식의 형태로 입력할 수 있도록 하고 그 불리언 연산식을 해석한 결과로 색인 데이터베이스를 접근하는 과정을 포함한다. 반면에, 실험에 의하면 대부분의 일반 사용자는 자신의 검색 의도를 불리언 연산식 형태로 표현하는데 익숙하지 않거나, 불편함을 느끼고 있으며, 검색 결과에 만족하지 못하면서도 매우 단순하게 아주 적은 수의 검색 키워드 입력을 통해 검색을 수행하는 경향이 뚜렷하게 나타나고 있다(박소연, 이준호 2002).

웹 정보 환경은 문서들이 빈번하게 생성되고 소멸되는 매우 유동적인 환경이며, 검색 대상 문서들은 양적으로 방대하고 형식, 주제, 생산자 등에 있어서 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 가장 두드러진 특징으로 나타나고 있다(박소연, 이준호 2002). 대부분의 검색 시스템 사용자들은 복잡한 검색식이나 연산자를 사용하지 않고 매우 단순한 검색 키워드를 입력하고, 결과적으로 검색 결과 집합에 대해 충분히 만족하지 못하여 검색 키워드에 대한 사소한 수정과 함께 반복적으로 검색하게 되는 것이다.

사용자가 검색 의도를 가장 정확하게, 또 가장 편리하고 자연스럽게 표현할 수 있는 인터페이스는 자연어 문장을 검색 질의 문장으로 입력하는 방법이다. 이때 형식 질의어와 달리 자연어의 본질적인 특성에 의해 동일한 의도를 갖는 검색 요구에 대해서도 개인의 성향이나 습관에 따라서 다양한 용어나 구조의 자연어 질의 문장으로 입력될 수 있음이 매우 중요한 문제이다. 따라서 자연어 인터페이스를 갖는 검색 시스템은 질의 문장을 자연어 처리 기술에 따라 처리하는 기능을 필수적으로 수반해야 한다.

구문 분석 과정을 배제한 자연어 질의 검색 시스템은 대부분 형태소 분석 키워드 추출, 불용어 제거 등의 과정을 포함하지만, 통계적 색인과 검색에 비해서 보다 정교한 정보 검색을 위해서는 형태소 분석 이상의 구문 분석과 같은 고수준의 자연어 처리 기술의 도입이 필수적이다.

본 논문은 사용자가 입력하는 한국어 자연어 질의 문장을 형태소 분석 및 구문 분석하는 자연어 처리 기술 기반의 질의 처리와 검색 방법을 제안한다. 색인 데이터베이스와 질의 문장에서의 검색어 일치, 불일치를 보다 정교하게 처리하기 위해 복합 명사를 조합 형태와 분할 형태로 확장하여 다중 검색어로 추출한다. 뿐

만 아니라 사용자들이 빈번하게 사용하는 축약 표기 용어들과, 음역어에서 흔하게 나타나는 이형 표기 용어들에 대해서도 마찬가지로 검색어를 확장하는 다중 검색 방법을 제안한다.

2. 관련연구

자연어 질의 문장의 처리는 형태소 분석 및 구문 분석 결과에 기반하여 다양한 문법적, 의미적 관계를 분석하고 유지함으로써 단순하게 형태소 분석으로부터 추출된 검색어들의 집합으로 취급할 때에는 불가능한 관계성을 파악할 수 있다. 동일한 검색 의도를 갖더라도 사용자의 성향과 습관에 따라 다른 구문 구조의 문장으로, 복합 명사들의 상이한 조합으로, 축약 표기 용어 및 이형 표기 용어의 서로 다른 선택으로 각각 다른 형태로 입력되는 자연어 질의 문장의 문제들을 해결하는 기법이 검색 시스템의 성능을 좌우하게 된다.

다음은 검색 시스템 사용자가 자연어로 입력하는 질의 문장의 여러 가지 가능한 형태를 가상의 예로 든 것이다. 여러 문장에서 검색자의 의도는 한 가지이며, 이상적으로 볼 때 검색 결과 집합은 동일해야 할 것이다. 하지만 기존 검색 시스템에 아래와 같은 다양한 형식의 질의 문장들을 입력하면 결과 문서 집합이 상당한 차이를 보임을 알 수 있다.

자연어 질의 예문

"하이퍼 텍스트의 방향 상실 문제에 대한 자료들을 찾아 주시오."

"하이퍼텍스트와 방향상실에 관해 설명한 문서를 찾아라."

"하이퍼텍스트에서 방향상실에 대해 소개한 문서들을 찾아라."

"하이퍼 텍스트에서의 방향 상실에 관한 문서를 찾아라."

"하이퍼 텍스트의 방향성 상실에 대한 문서를 찾아라."

"하이퍼 텍스트와 방향성 상실에 관해 다룬 문서를 찾아라."

"하이퍼 텍스트에서의 방향상실 문제에 대해 소개한 문서들을 찾아라."

한편 검색어의 추출이 검색 엔진의 성능을 결정하는 기본적인 요소임에도 불구하고 정보 검색 연구 분야에서는 정보 자료의 색인 기법보다는 좀 더 정교한 검색 모델에 의해 성능 향상을 추구하는 방향으로 연구가 진행되어 왔다(강승식 2004). 즉, 색인어와 색인 기법이 검색 결과에 미치는 영향이 검색 모델에 비해 매우 중요함에도 불구하고 상대적으로 중요시되지 않았던 것으로 보인다. 이는 정보 검색 학문 분야가 영어권 문서들을 중심으로 연구되어 왔고, 한국어, 중국어, 일본어 등의 언어들과 구조가 크게 달라 색인 기법에 의해 성능을 향상시킬 수 있는 여지가 적기 때문일 것이다.

Salton의 연구(Salton Gerard 1988)에서는 통계적인 방법과 구문 구조를 이용한 색인의 비교에서 복잡한 자연어 처리를 이용한 방법보다는 방법론이 간단한 통계적인 방법을 선호했다. 하지만 통계적 방법과 비교된 구문 구조 분석 방법은 적용 과정이 지나치게 단순하여 적절한 비교가 되지 못하였다. 또 다른 연구는 구를 색인어로 합성해 내기 위해 구문 규

칙 대신 단일어들의 출현 빈도를 기반으로, 단어의 출현 위치, 단일어 간의 거리, 수식어로의 출현 빈도 등 여러 가지 요소들을 사용하였다(Joel L Fagan 1987). 구를 합성하여 색인어로 사용하여 실험 집합에 대해 평균적으로 10% 정도의 정확도를 향상시켰다. 하지만 실험 집합에 따라 성능의 차이가 크고, 이용되는 요소들이 대상 도메인에 따라 반복 실험에 의해 구해져야 하는 문제점을 보였다. Fragan (1989)는 앞의 통계적 방법 기반의 단점을 보완하기 위해 완전 구문 분석기를 이용하여 구문 구조를 생성하기도 하였으나, 너무 적은 수의 구를 생성하여 검색 성능에 거의 영향을 주지 못했다. 최근의 연구에서는 대규모의 문장을 처리할 수 있는 강건한 자연어 처리기를 사용하기 위해 완전 구문 분석기 대신 명사구 분석기(NP Phrase parser)를 이용하여 명사구를 합성하였다. 250MB의 대규모 문서 집합을 실험하여 단일어만을 색인했을 때 보다 합성한 명사구를 색인어에 추가했을 때 정확도에서 최고 18%, 재현율 13%의 성능 향상을 얻었다고 한다(Zhai Chengxiang 1997).

구글 검색 엔진에서 한글 문서의 색인은 한글의 교착어적 특성을 반영하여 형태소 분석을 통해 색인한다(<http://www.google.co.kr>). 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 복원하지는 않고, 변형된 어간의 경우에 입력 어절에서 어간의 길이만큼 어절 단위로 절단하여 색인어로 추출한다. 결과적으로 의존명사, 어미, 조사 등의 형태소들이 색인어 및 검색 키워드로 추출되기도 한다. 어미 '고'나 '게', 또 의존명사 '수' 등에서 그 예를 볼 수 있다.

아래의 <표 1>은 현재 널리 사용되고 있는 세 가지의 검색 시스템 A, B, C(<http://www.google.co.kr>, <http://www.naver.com>, <http://kr.yahoo.com>) 에서 자연어 형식의 검색 구(phrase)를 입력한 검색 결과의 일부 예이다. 검색 문서나 질의어의 형태소들을 각

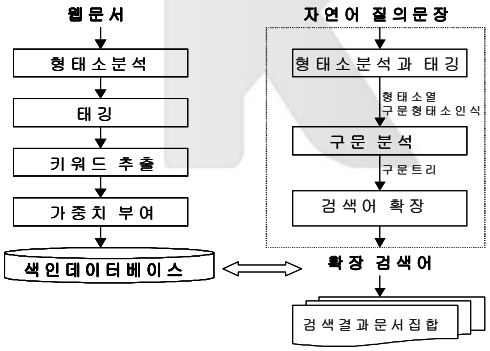
〈표 1〉 입력구와 검색 결과의 부분 예

입력 예	검색 문서 내용 예	검색시스템
정확 **하고** 빠른 데이터의 전송	복잡 **하고**, 처리 **하고**	A
정확하고 **빠르게** 데이터를 전송	**빠르**나, **빠르**고, **빠르**다니 쉽**게**, 작**게**, 옳**게**, 어떻 **게**	A
빠르 **고** 신뢰성 **있는**	**고**객의, 두 **고**, 있**고** 힘 **있는**, 가치 **있는**	A
빠르 **고** 신뢰할 **수** 는	느리 **고**, 크 **고** 제공할 **수** 있으며, 광고할 **수 있는** 실효성 **있는**,	A
	맡길 **수 있는**, 작성할 **수 있는**, 얻을 **수 있는**, 관계가 **있는**	B
	사용자 **수**준의, 만들 **수** 있다, 설득력 **있는**, 즐길 **수 있는**, **있는** 대로 밟아	C

각 색인하는 과정이 항상 더 정확한 결과를 얻게 되는 것은 아님을 볼 수 있다. 또한, 검색 시스템에 따라서는 형태소들을 전혀 분리하지 않기 때문이다. 또는 색인부나 검색부에서 문장이나 구의 형태소들을 분리하는 과정을 거친 후 색인어로서 의미가 없는 형태소적, 문법적 요소들을 구 단위나 문장으로부터 분리하고 적절한 색인어나 검색어를 추출하는 과정이 뒤따르지 않았기 때문이다.

3. 자연어 질의 문장의 분석

다음의 <그림 1>은 본 논문에서 제안하는 자연어 질의 문장에 대한 구문 분석 및 검색어 확장에 의한 다중 검색과 문서 색인부의 관계를 보여준다.



〈그림 1〉 자연어 검색부와 색인부

자연어 질의 문장의 형태소 분석과 구문 분석의 결과로서 구문 트리를 얻고, 트리 순회를 통해 복합 명사를 조합하거나 분할하여 검색어를 확장한다. 또한 음역어 등에서 흔히 나타나는 이형 표기 용어들과 축약 표기 용어들에 대해서도 질의 검색어를 확장하여 다중 검색한다.

3.1 구문 형태소열의 인식

<그림 1>에서 보는 바와 같이 사용자가 입력한 자연어 질의 문장은 형태소 분석, 불용어의 제거, 태깅 과정을 우선적으로 거친다. 자연어 문장에서 연속적으로 나타나는 형태소 패턴들 중에는 빈번하게 출현하면서도 구문 분석 과정에서 생성하게 될 문법적 구조에 크게 의미가 없는 관용적 패턴의 성격을 갖는 것들이 있는데, 이들을 구문 형태소열로 정의하기로 한다.

구문 구조 규칙들을 반복적으로 적용하는 구문 분석 과정 앞서서 구문 형태소 열들을 인식하고 단위 구문 요소로 결합하여, 양상 정보를 추출하는 전처리 과정을 도입함으로써 질의 문장 문법 구조 분석 과정의 계산 효율을 높일 수 있다.

다음 예들은 구문 형태소로 정의되고 전처리될 수 있는 연속된 형태소열들 중에서 빈도수가 높은 대표적인 예들이다(황이규, 이현영, 이용석 2000, <http://kibs.kaist.ac.kr>).

구문 형태소열 예

"-에/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 수행하므로

다른 숙어나 관용어에 비해 문장에서 자주 발생하며, 다양한 체언이나 용언에 이 구문 형태소열들이 결합할 수 있기 때문에 언어 생성력이 매우 크다(황이규, 이현영, 이용석 2000). 코퍼스로부터 일정 회수 이상 빈도를 가지며, 한 어절 이상에 걸쳐 기능 형태소들을 대상으로 체언이나 용언에 구조적, 의미적 영향을 미치는 상위 빈도 어절들이다. 또한 여러 기능 형태소들이 결합하여 하나의 구문, 의미적 단위를 형성하는 형태소열로서, 그 자체로 하나의 문법적 범주가 될 수 있다.

황이규, 이현영, 이용석(2000)은 관용구 관점에서 매우 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.

자연어 문장의 구문 분석 과정은 본질적인 형태소적, 구문 구조적, 의미적 모호성으로 인해 계산 과정이 아주 복잡하고, 생성되는 레코드의 양이 엄청나게 많을 수 있다. 구문 분석 과정에서 문법 규칙들의 직접적인 적용에 앞서 이와 같은 성격의 구문 형태소열들을 미리 인식하고 단일 문법 요소로 결합시키는 전처리 과정은 구조적 모호성을 크게 감소시킴으로서 질의 문장의 구문 분석 과정을 통해 전체 검색 시스템의 계산 성능을 향상시킬 수 있다. 특히 본 논문에서 제안하는 규칙 기반의 구문 분석 과정은 구문 형태소열들에 대해 문법 구조, 즉 트리를 생성하기 위해 많은 중간 구조들을 만들게 되기 때문에 문법적 의미가 거의 없는 단순 구조들을 미리 결합하는 전처리가 매우 효율적인 방법이다.

3. 2 자연어 질의 문장의 구문 분석

형태소 분석과 태깅, 구문 형태소열의 인식 과정을 마친 질의 문장은 구문 분석 과정을 거쳐서 단어들의 문법적 관계를 얻기 위해 문장의 문법 구조를 해석하고 그 결과로 구문 트리를 생성한다. 자연어 질의 문장의 구문 분석 과정은 단순한 형태소 열이 나타낼 수 없는 중요한 구조적, 의미적 관계를 추출할 수 있는 과정으로서, 자연어 처리에서 가장 핵심적이고도 복잡한 처리 과정이다(이공주, 김재훈 2003, 황이규, 이현영, 이용석 2000). 다음의 <그림 2>는 자연어 질의 입력 예문의 구문 분석에서 얻어지는 구문 트리의 예들이다.

자연어 문장을 구문 분석하는 방법은 통계적인 방법, 구문 규칙에 기반한 방법, 학습에 의한 방법 등 여러 가지가 있으나, 본 연구에서는 이미 여러 실험에서 구현되고 적용된 바 있는 문맥 자유(context free) 형식 구 구조(phrase structure) 규칙을 기반으로 하는 상향식(bottom up) 분석 방법을 채택하였다. 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(이공주, 김재훈 2003). 또한 규칙 기반의 구문 분석 방법은 통계적 방법에 비해 문제 분

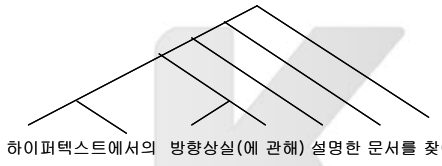
야에 영향을 받지 않고 여러 응용에서 널리 사용될 수 있다.

```
물질/을
    +---생성/되/는
            +---과정/에서
                    +---처리/하/는


물질/을
    +---소멸/되/는
    +---생성/되/는
            +---저절로
            +---과정/에서
                    +---처리


찾/아/라/
    +---문서/를
            +---상실/(에 대한)
                    +---방향
                    +---하이퍼텍스트/에서/의

찾/아/라/
    +---문서/를
            +---설명/하/ㄴ
                    +---방향상실/(에 관해)
                            +---하이퍼텍스트/에서/의
```



〈그림 2〉 자연어 질의와 구문 트리

3 .3 복합 명사 처리와 검색어 확장

한국어 자연어 질의 문장 사용 행태에 대한 조사 및 연구에 의하면 “한국어의 질의어는 거의 명사 유형이다”라고 한다(강현규 2002, Lee G., Park, M., and Won, H. 1999) 명사가 검색 사용자들의 질의 의도를 나타내는 중심이 된다는 뜻이다.

질의 문장에 의한 검색에서 의미의 중심 역할을 하는 명사들은 색인이나 검색에서 복합 명사의 다양한 형태로 빈번하게 나타난다. 복합 명사의 문법적 규정은 매우 약하고 띄어쓰기 문제와도 긴밀하게 관련되어 있다. 두 개 이상의 명사를 붙여 쓴 형태이거나, 띄어 쓴 두 개의 단일 명사이거나 대부분의 경우 모두 문법적으로 오류가 되지 않는다.

KTSET 을 중심으로 다수 문서와 질의 문장에서 나타나는 복합명사들의 길이에 대한 통계 자료를 참조하면 길이가 3까지인 복합 명사가 전체의 96.8%를 차지하며, 길이가 4 이상인 복합 명사 매우 드물게 나타난다는 실험 결과가 있다. 반대로, 한국어에서 4음절 이상의 단일어는 극히 드물기 때문에 4음절 이상의 명사라면 복합어일 확률이 매우 높다고 볼 수 있다(Won, H., Park, M, and Lee, G. 2000).

다량의 문서에 대한 색인 데이터베이스 구축을 위해 문서에 나타나는 색인어들에 대해 복합 명사 분할 및 조합 과정을 적용하는 것은 주기적으로 갱신되어야 하는 동적인 웹 환경의 색인 시스템에 대한 부하가 매우 크다. 따라서 본 연구는 <그림 1> 에서와 같이 색인 데이터베이스는 형태소 분석과 태깅 결과로 나타나는 색인어들로 색인하고, 입력 질의 문장에 나타나는 명사 검색어들에 대한 복합 명사를 분할하고 조합한 결과로 색인 데이터베이스를 다중 검색하는 방법을 우선 적용하고 실험하였다. 검색어의 추출은 품사열 패턴 매칭의 한계를 극복하기 위해 구문 분석 결과인 구문 트리에 의존한다. 또한 이미 구축되어 사용되고 있는 상용의 검색 시스템에 대한 자연어 메타 검색기로서의 역할을 하여 자연어 질의 문장을 입력할 수 있는 사용자 환경을 제공하고, 입력 질의 문장의 문법적 구조를 분석하여 복합 명사를 조합하고 분할한 결과를 검색어를 추출하여 검색 시스템에 제공할 수도 있다.

3. 3. 1 복합 명사 조합에 의한 다중 검색

구문 분석 결과인 구문 트리에서 다양한 구조로 나타나는 어휘들은 같은 검색 의도를 말하는 검색어들일 수 있다. 구문 트리를 순회하면서 문법 구조적으로 연관된 어휘들을 추출하여 복합 명사 형태로 조합하여 색인어로 확장하며, 이들을 검색어로 다중 검색한다. 구문 트리에서 2개 이상의 명사가 복합 명사로 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	실험 결과 분석
· 성능을 분석하고 효율을 검증하는 · 성능을 분석한 후 효율을 검증하는	성능분석 효율검증
· 처리하는 과정에서 생성되는 물질	처리과정 생성물질
· 결과를 유형별로 분류하여	결과분류

형태소 수준의 복합 명사 조합 방법이 의미 없는 조합을 많이 생성하는 문제가 있는 반면, 본 연구의 방법과 같이 구문 트리의 구조에 기반한 복합 명사 조합은 사용자의 검색 의도를 반영한 조합이 가능하다. 2 단어 이상의 복합 명사 확장 방법은 기존의 연구 방법을 따르며, 길이 4 이상의 복합 명사는 실제로 사용 빈도가 극히 낮고 오히려 검색 성능을 떨어뜨리는 경향이 있으므로 길이 3까지 조합하는 것이 효율적이다(Won, H., Park, M, and Lee, G. 2000).

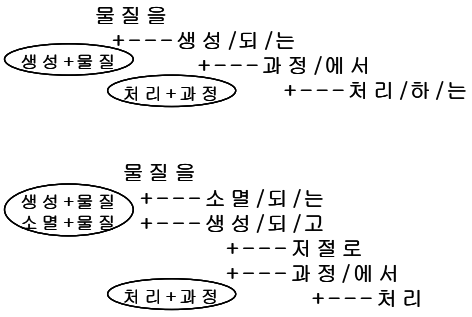
3. 3. 2 복합 명사 분할에 의한 다중 검색

4 음절 이상의 단일 명사는 복합 명사인지 여부를 검사하고 단일어 추출과 띄어쓰기의 분할 과정을 거쳐서 입력 질의에 대한 검색어를 확장하고 다중 검색을 한다. 앞에서 설명한 바와 같은 이유로 복합 명사는 두 개 또는 세 개 단일 명사들의 조합으로 분할한 다음 색인 데이터베이스를 다중 검색한다. 이때 분할된 각 단일 명사들은 모두 색인 데이터베이스에 독립적으로 등록된 색인어이거나, 접두사 또는 접미사이어야 한다. 분할된 단일 명사들은 복합 명사에 비해 특정성이 떨어질 수 있지만 실험에서 보이는 바와 같이 검색 결과의 재현율을 상당히 높일 수 있다.

복합명사 분할 예

통계정보	통계 정보
실험결과분석	실험 결과 분석 실험결과 분석 실험 결과분석
병렬처리시스템	병렬 처리 시스템 병렬처리 시스템 병렬 처리시스템
음성인식	음성 인식
효율검증	효율 검증
결과분류	결과 분류

다음의 <그림 3>은 구문 트리로부터 조합되는 복합 명사의 예와 띄어쓰기에 따라 확장될 수 있는 검색어들의 예로 분할된 형태들의 예를 보여준다. 이와 같이 조합되고 분할된 복합 명사의 여러 형태들은 모두 검색어로 추출되고 확장되어 색인 데이터베이스에 대한 다중 검색을 시도한다.



〈그림 3〉 구문 분석과 복합 명사 조합

복합 명사의 분할과 검색은 문서 색인 과정에서 이루어질 수도 있다. 하지만 구문 트리로부터 조합된 복합명사의 경우 붙여 쓴 형태와 띄어 쓴 형태의 검색어 및 개별 검색어들이 추출 가능하며, 각 경우 검색 정답 문서의 우선순위를 계산하는데 다른 가중치로 적용될 수 있다. <그림 3>의 예에서 조합된 복합 명사 '처리과정'뿐만 아니라 단일어 '처리' 및 '과정'으로도 검색될 수 있다. 검색 문서에 대한 정밀한 우선순위 계산에 대한 연구는 본 논문의 주제와는 별도의 과제로 삼고자 한다.

3. 4 이형 표기 및 축약 표기 용어 확장에 의한 다중 검색

음역어 등의 예에서 흔히 볼 수 있는 이형 표기의 검색어들을 질의 구문 트리로부터 추출하고 확장하여 이들을 검색어로 다중 검색한다. 이와 같은 이형 표기 색인어들은 미리 수집되어 별도의 사전에 등록되어야 한다. 다음의 예는 각종 문서와 질의 문장에서 빈번하게 혼용되고 있는 이형 표기 용어들의 몇 가지 예이다.

이형표기 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 한 가지의 예는 축약 표기 용어이다. 축약 표기 용어들은 또 다른 종류의 이형 표기 용어로 처리할 수 있다. 축약 표기 용어와 이에 대응하는 기본 색인어들은 이형 표기 용어 사전에 등록되어 질의 문장에 입력된 검색어들을 양방향으로(축약 표기 형태 및 원래 형태) 확장하고, 검색 시스템은 이들을 다중 검색 한다.

축약표기 검색어의 예

헌재 / 헌법재판 / 헌법재판소

정통부 / 정보통신부

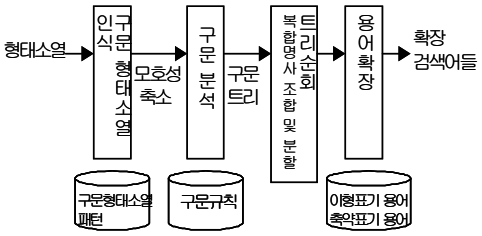
노조 / 노동조합

야후, 구글, 네이버 등의 기존의 여러 검색 시스템에서 위 예의 이형 표기 용어들이나 축

약 표기 용어들을 검색어에 포함시켜 검색을 시도하면 검색 결과 문서의 집합은 모두 상이함을 볼 수 있다(<http://www.google.co.kr>, <http://www.naver.com>, <http://kr.yahoo.com>). 예를 들어 이형 표기어인 "알고리즘"과 "알고리듬"을 각각 검색 키워드로 입력하였을 때, 검색 결과의 정답 문서들의 집합은 완전히 다른 문서 집합들이다. 단지 표기 형태만 다를 뿐 완전히 동일한 용어임이 전혀 반영되지 못한 결과이다. 이들 이형 표기 용어들의 집합이나 축약 표기 용어 집합들이 별도의 사전으로 등록되어 처리되면 사용자의 검색 만족도를 높일 수 있다.

이형 표기 용어들과 축약 표기 용어들은 모두 같은 방법으로 검색어 확장된다. 이와 같은 용어들을 여러 단계를 거치는 자연어 질의 문장 처리 과정의 앞 단계, 즉 태깅의 결과에 적용하는 방법도 생각할 수도 있지만, 구문 구조 분석 과정에 기반하는 본 연구의 질의 분석과 검색어 확장 방법에서는 형태소적 모호성과 구조적 모호성이 상당 부분 해소된 뒤 단계에서 검색어를 확장하는 것이 훨씬 효과적이다. 예를 들어, "불린"은 "불린 연산식(boolean expression)", "물에 불린 미역", "바보라고 불린 청년" 등의 모호성을 가질 수 있다. 따라서 문법적 구조 분석의 결과를 얻은 후에 이형 표기 용어나 축약 표기 용어들을 추출하고 다중 검색하는 방법이 타당하다.

다음의 <그림 4>는 질의 문장 분석에 기반한 검색어 추출과 확장 과정들을 보여 준다. 그림에서 보이는 각 검색어 확장 과정들은 질의 문장의 트리 구조에서 중복하여 적용될 수 있다.



〈그림 4〉질의 문장 분석과 검색어 확장

4. 실험 및 평가

정보 검색 시스템의 성능은 흔히 재현율(recall)과 정확도(precision)라는 척도로 평가된다. 재현율이란 검색에서 관련 문서를 추출해 낸 비율이다. 따라서 같은 내용을 담고 있는 문서들에 대해 그에 관련된 질의어로 모든 문서를 검색할 수 있는지를 말하는 척도이며, 정확도란 검색에서 추출한 문서 중 관련된 문서의 비율이다. 비슷하지만 조금씩 다른 내용을 담고 있는 문서들을 같은 질의어로 검색을 해서 추출하였을 때, 관련 문서가 어떤 비율을 차지하는지를 말한다(Baeza-Yates Ricardo, and Reberio-Neto Berthier 1999).

분할된 단일 명사들은 복합 명사에 비해 특정성이 떨어지게 되지만 검색에서 재현율을 높일 수 있고, 문장 상에 나타나는 단어들로 명사구를 조합하게 되면 특정성이 큰 색인어를 만들게 되어 정확도를 높일 수 있다. 따라서 복합 명사의 분할 및 조합 과정과 기타 검색어 확장 방법이 재현율과 정확도 측면에서 정보 검색 시스템의 성능 향상시킬 수 있다.

본 연구의 방법은 여러 실험에서 사용된 바 있는 KTSET 문서들(약 4,400여 문장을 포

함)과 개인, 학과, 부서, 학교, 기관 등의 홈페이지들, 학술 내용과 시사 뉴스 등의 내용을 담은 웹 문서들을 대상으로 검색자에게 임의의 질의 문장을 입력하도록 하였다. 제한 없는 임의의 검색자들이 원하는 내용의 문서들을 검색하기 위해 입력하는 한국어 자연어 질의 문장들을 수집하여 약 4,900여 질의 문장을 대상으로 본 연구의 방법을 실험하였다.

구현된 형태소 분석기와 규칙 기반의 상향식 구문 분석기의 계산 결과로 구문 트리를 생성하였으며, 중간에 구문 형태소열의 인식 과정을 둠으로써 구문 분석의 부하를 크게 줄일 수 있었다. 실험에서 수집된 자연어 검색 질의 문장의 두드러진 특징으로 문장의 길이가 일반 문서의 문장에 비해 비교적 짧고, 복합 명사들이 다양한 형태로 빈번하게 출현한다는 점을 볼 수 있었다. 입력된 질의 문장의 구문 형태소 처리를 거쳐 구문 분석하는 과정은 일반 문서들에 포함된 문장을 구문 분석 하는 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 질의 문장은 상대적으로 단순하기 때문인 것으로 분석된다. 하지만 여전히 사용자의 자연어 질의 문장은 문법적 오류를 많이 담고 있음을 볼 수 있었다. 각 단계의 검색어 확장 과정을 거치면서 입력 질의 문장에 대한 검색어의 개수는 검색어 확장을 하지 않았을 때와 비교하여 2.72배로 증가하였다. 질의 문장의 구문 분석 과정과 검색어 확장에 의한 다중 검색 과정이 처리 속도 면에서 다소 성능의 저하를 가져오지만, 재현율과 정확도의 향상을 본 연구에서 제안하는 검색 기법의 성능 평가의 우선 척도로 삼고자 한다. 이와 함께 실용적으로 성능이 우수한 구문 분석기의 개발을 본 연구의 중요한 과제로 삼고 있다. 앞에서 기술한 바와 같이 본 연구에서는 구문 형태소열의 인식 과정에 의해 처리 부하가 크게 감소될 수 있었으며, 상향식 구문 분석 방법으로 접근하여 성공하지 못한 입력 문장의 경우에 대해서도 부분 분석의 중간 결과를 이용함으로써 전체 시스템의 유연성을 확보하였다.

검색 성능의 비교를 위해 질의문장 구문 분석을 거치지 않으며, 또한 복합명사 분할이나 조합, 이형 표기어, 축약 표기어 등에 대한 검색어 확장 과정을 전혀 거치지 않고, 형태소 분석의 결과로부터 명사 키워드들을 검색어로 간주하여 관련 문서를 검색하는 과정과 본 논문에서 제시한 방법대로 구문 분석 과정을 거치고 검색어 확장 과정을 통한 검색 결과에 대한 재현율과 정확도를 비교하는 방법을 취하였다.

검색 성능으로 볼 때에는 검색어 확장 과정을 전혀 거치지 않았을 때보다 재현율이 약 11.3%가량의 향상을 보였다. 특히 이형 표기 용어들이나 축약표기 용어들에 대한 검색어 확장은 재현율의 향상에 크게 기여하고 있다. 전반적으로는 정확도는 약 4.7% 정도로 다소 높아졌으나, 일부 검색 시도에서는 거리가 먼 명사들의 조합과 의미 없는 분할이 일어나는 경우도 나타나고 있어서 이와 같은 경우에 대한 분석이 필요한 것으로 보인다.

이형 표기 용어들이나 축약 표기 용어들에 대한 대응 사전이 시소러스를 통해 구축되어야 하는 것이 최종적인 목표이지만 현재의 실험에서는 대상 문서들과 수집된 질의 문장들에 출

현하는 용어들을 대상으로 수동의 실험 사전을 이용하였다. 이러한 용어들은 검색 정확도나 재현율을 높이기 위해 동의어 사전을 구축하는 것과 같은 개념에서 접근될 수 있다.

〈표 2〉 질의 문장 분석과 검색어 확장처리 비율

총 질의 문장 수 : 4,894 문장	
구문 분석 성공률	68.76%
구문 형태소열 포함하는 문장	81.44%
복합 명사 조합 처리되는 문장	77.31%
복합 명사 분할 처리되는 문장	64.28%
이형 표기 용어 확장되는 문장	29.52%
축약 표기 용어 확장되는 문장	13.54%
평균 검색어 확장 비율	2.72 배

검색 대상 문서들의 문장들에 대한 색인 데이터베이스 구축 과정에서도 구문 분석과 복합 명사 조합이나 분할, 이형 표기 및 축약 표기 용어 확장 등 색인어를 추출하기 위한 동일 과정을 거치는 것이 검색어 일치 측면에서 이상적이다. 하지만 웹 문서의 특성 상 문법적 오류를 갖는 문장들이 많이 포함되고, 수시로 생성되고 소멸되며, 내용이 변경되는 동적인 웹 환경은 색인데이터베이스 구축 시스템에서 구문 분석의 성능의 저하로 인해 검색어 확장 과정과 결과에서 상당한 오류를 발생시킬 수 있으므로 타당성에 대한 연구가 필요하다고 본다. 본 실험에서 사용한 검색 대상 문서의 규모를 확대하여 보다 다양한 웹 문서를 검색하는 사용자들의 자연어 질의를 추가 수집하고, 단독 검색기로서, 또 기존 검색 시스템의 메타 검색기로서 검색어 확장의 성능 실험을 계속하고 있다. 또한 오류 현상에 대한 유형별 분석과 대응 방법도 모색하고 있다.

검색 시스템 사용자의 행태로 볼 때, 사용자들은 대부분 검색 결과 중 상위 20개가량의 문서만을 참조하는 것으로 분석된다(이공주, 김재훈 2003). 따라서 검색 목표가 되는 문서들이 실제로 상위 랭킹 문서가 되도록 가중치 계산에 대한 연구가 반영되는 것이 필요하며, 이형 표기 용어와 축약 표기 용어들에 대한 사전은 실험적 규모 이상의 통계적 수집 시스템을 구축할 필요가 있다.

5. 결 론

본 논문에서는 정보 검색 시스템의 이상적 인터페이스로서 한국어 자연어 질의 문장을 입력하고 분석하여 검색어를 추출하고 확장하는 처리 과정을 제안하였다. 형태소 분석과 구문 분석을 비롯한 기본적인 자연어 처리 과정을 포함하며, 단순히 명사 검색어뿐만 아니라 문장의 구조가 전달하는 정확한 문법 관계를 얻기 위해 구문 분석 과정도 포함된다. 형태소 분석, 구문 형태소 추출, 구문 분석 결과로부터 검색어 추출과 확장 과정이 주요 처리 과정이다. 구문 트리를 순회하면서 복합 명사를 분할, 또는 조합하고, 이형 표기 용어 및 축약 표기 용어에 대한 검색어 확장 과정을 포함한다.

입력 질의 문장의 구문 분석 과정과 그 결과를 기반으로 검색어들 사이의 계층적 구문 구조가 전달하는 의미 관계를 획득함으로써 정확한 검색어를 추출하는 것이 목적이다. 각종 문법적 요소들의 분해가 단순히 색인어나 검색어가 되는 것이 아니며, 의미 관계가 있는 명사들을 인식하여 복합 명사로 조합하여 검색하

고, 분할 형태나 이형 표기 및 축약 표기 검색어들의 확장을 통해 다중 검색함으로써 검색 시스템의 재현율과 정확도를 높일 수 있음을 실험을 통해 보여주었으며, 계속되는 연구의 방향을 제시하였다.

참 고 문 헌

강승식. 2004. 한글 문서의 색인어와 색인 기법. 『정보과학회지』, 22(4): 72-77.

강현규. 2002. 개념 검색어 확장을 통해 질의 형식화를 도와주는 개념 마법사의 설계 및 구현. 『정보처리학회논문지』, 9-B(4): 437-444.

박미화, 원형석, 이근배. 1999. 구문분석에 기반한 한글 자연어 질의로부터의 불리언 질의 생성. 『정보과학회논문지』, 26(10): 1219-1229.

박세영, 강현규. 1998. 한글공학:정보검색. 『한국정보처리학회』, 특집 5(5).

박소연, 이준호. 2002. 로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구. 『정보관리학회지』, 19(3): 111-122.

이공주, 김재훈. 2003. 규칙에 기반한 한국어 부분 구문분석기의 구현. 『정보처리학회논문지』, 10-B(4): 389-396.

장명길, 김현지, 장문수 외. 2001. 의미기반 정보검색. 『정보과학회지』 한글정보처리 특집, 19(10): 7-18.

황이규, 이현영, 이용석. 2000. 형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용. 『정보과학회논문지』, 27(7).

Arampatzis, A.T., Tsoris, T., Koster. C.H.A. and Th.P. van der Weide, 1998. "Phrase-based Information Retrieval." *Journal of information Processing & Management*, 34(6).

Baeza-Yates Ricardo. and Reberio-Neto Berthier. 1999. "Modern Information Retrieval," *Addison Wesley*.

Fagan Joel L. 1987. "Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods," *Ph.D. thesis, Cornell University.*

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

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

Perez-Carballo Jose and Strazalkowski Tomek. 2000. "Natural Language

Information Retrieval : progress report." *Information Processing & Management.* 36(1).

Salton Gerard. 1988. "Automatic text processing." *Addison-Wesley publishing company.*

Won, H., Park, M. and Lee. G. 2000. "Integrated indexing method using compound noun segmentation and noun phrase synthesis," *Journal of KISS: Software and Applications,* 27(1).

Zhai Chengxiang. 1997. "Fast statistical parsing of noun phrases for document indexing." *Fifth conference on applied natural language processing*: 312-319, KIBS : Korean Information Base System. <http://kibs.kaist.ac.kr>

<http://www.google.co.kr>

<http://www.naver.com>

<http://kr.yahoo.com >