정보 검색 시스템의 성능 향상을 위한 구문 분석과 검색어 확장

윤 성 희
상명대학교 컴퓨터소프트웨어공학전공
shyoon@smu.ac.kr

Syntactic Analysis and Keyword Expansion for Performance Enhancement of Information Retrieval System

Sung-Hee Yoon
Dept. of Computer Software Engineering, Sangmyung University

요 약

자연어 질의 문장을 입력하는 방법은 정보 검색 시스템 사용자에게 아주 이상적인 인터페이스이다. 검색을 위해 색인어를 입력하거나 불리언 질의식을 사용하는 것에 비해 훨씬 친밀하지만, 동일한 의도의 검색 요구에 대해서도 개인의 성향에 따라서 다양한 형태나 구조의 자연어 질의문장으로 입력될 수 있는 본질적인 특성이 있다. 본 논문은 자연어 질의문장을 입력으로 하는 검색 시스템을 위해 사용자의 입력 질의 문장을 분석하고 검색어를 확장하는 다중 검색 기법을 제안한다. 질의 문장에 대한 형태소 분석 및 구문 분석을 수행하고, 구문 트리를 순회하여 구조적으로 연관된 복합명사를 조합하거나 분할하고, 이형 표기 용어와 축약 표기 용어들을 확장하여 다중 검색함으로써 재현율과 정확도를 높일 수 있다.

1. 서론

일반적으로 정보 검색 시스템은 입력 질의(query)와 문서 내용에 대한 색인어(index)의 형태적 일치를 검사함으로써 검색 결과 문서 여부를 결정하며, 많은 웹 문서들은 주기적으로 방문되어 색인데이터베이스를 갱신한다. 실험에 의하면 대부분의 일반 사용자는 원하는 정보를 불리언 연산식 형태로 표현하는데 불편함을 느끼고 있으며, 복잡한 검색식이나 연산자를 사용하지 않고 적은 수의 검색어로 구성된 단순한 질의를 통해 정보 검색하는 경향이 뚜렷하여 검색 키워드에 대한 사소한 수정과 함께 반복적으로 검색한다[7].

사용자가 검색 의도를 가장 정확하게, 또 가장 편리하고 자연스럽게 표현할 수 있는 인터페이스는 자연어 문장을 검색 질의로 입력하는 것이다. 이때 형식 질의어와 달리 자연어의 본질적인 특성으로서 같은 의도의 검색 요구에 대해서도 개인의 성향이나 습관에 따라서 다양한 형태나 구조의 자연어 질의 문장으로 입력될 수 있음이 매우 중요한 문제이다. 따라서 자연어 검색 시스템은 자연어 질의 문장을 자연어 처리 기술에 따라 처리하는 기능을 필수적으로 수반해야 한다.

본 논문은 사용자가 입력하는 한국어 자연어 질의 문장을 형태소 분석 및 구문 분석하는 자연어 처리 기술 기반의 질의 처리와 검색 시스템을 제안한다. 색인 데이터베이스와 질의 문장에서의 색인어 일치, 불일치를 보다 정교하게 처리하기 위해 복합명사의 조합 형태와 분할 형태로 확장하여 다중 검색어로 추출한다. 또한 음역어에서 흔히 나타나는 이형 표기 검색어들과 축약 표기 검색어들을 확장하는 다중 검색 방법을 제안한다.

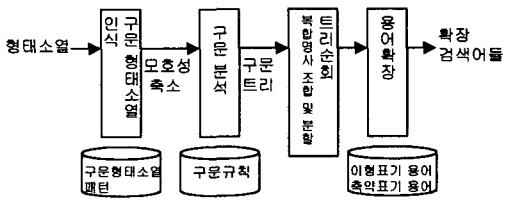
2. 자연어 질의와 구문 분석

다음의 표 1은 현재 널리 사용되고 있는 검색기 야후, 구글, 네이버 등에 다음과 같은 자연어 질의를 입력했을 때의 유의할 만한 검색 결과를 보여준다. 일부 검색기는 형태소들을 전혀 분리하지 않으며, 검색 시스템에 따라서는 색인어나 검색어를 구할 때 형태소 분석을 하기도 하는데, 항상 더 정확한 결과를 얻게 되는 것은 아님을 볼 수 있다. 색인부나 검색부에서 문장이나 구의 형태소들을 분리하는 과정을 거친 후 색인어로서 의미가 없는 형태소적, 문법적 요소들을 구(phrase)나 문장으로부터 분리하고 적절한 색인어나 검색어를 추출하는 과정이 뒤따르지 않았기 때문이다.

입력 예	검색 문서 내용 예
정확<u>하고</u> 빠른 데이터의 전송	복잡<u>하고</u>, 처리<u>하고</u>
정확하고 <u>빠르게</u> 데이터를 전송	<u>빠르</u>나, <u>빠르</u>고, <u>빠르</u>다니 <u>쉽게</u>, 작<u>게</u>, 옳<u>게</u>, 어떻<u>게</u>
빠르<u>고</u> 신뢰성 <u>있는</u>	<u>고</u>객의, 두<u>고</u>, 있<u>고</u> 힘 <u>있는</u>, 가치 <u>있는</u>
빠르<u>고</u> 신뢰할 수 <u>는</u>	느리<u>고</u>, 크<u>고</u> 제공할 수 있으며, 실효성 <u>있는</u>,
	맡길 수 <u>있는</u>, 작성할 수 <u>있는</u>, 얻을 수 <u>있는</u>, 관계가 <u>있는</u>
	사용자 수준의, 만들 수 있다, 설득력 <u>있는</u>, <u>있는</u> 대로 밝아

[표 1] 기존 검색시스템에서의 입력과 결과 예

본 연구에서 제안하는 자연어 질의 검색은 형태소 분석 및 구문 분석 등 자연어 처리 기술을 기반으로 하며, 다음 그림 1에서 보이는 바와 같은 검색어 확장 과정을 포함한다.



[그림 1] 질의 문장 분석과 검색어 확장

2.1. 구문 형태소열 인식

사용자로부터 입력된 자연어 질의 문장은 형태소 분석, 불용어의 제거, 태깅 과정을 거친다. 자연어 문장에서 빈번하게 출현하고, 그 구문 구조가 의미가 없는 관용적으로 연속된 형태소 열 패턴은 구문 구조 규칙을 적용하기 전에 구문 형태소로 추출하여 전처리한 다음 일반 구문 규칙을 적용한다. 구문 형태소 열은 문장에서 내용어의 역할을 수행하지 않고 기능어의 역할을 수행하며,, 다양한 체언이나 용언에 이 구문형태소가 결합할 수 있기 때문에 언어생성력이 매우 크다[10].

관용구 관점에서 매우 빈번하게 출현하는 구문 형태소들의 유형과 그 구체적인 예들에 대한 통계적 연구 결과가 있으며, **"-에/jca 대하/pvg ㄴ/etm","-을/jco 위한/pvg", "-ㄹ/etm 수/nbn 있/pvg", "-ㄹ/etm 수/nbn 없/paa"**등은 구문 형태소 열로 추출될 수 있는 연속된 최소 형태소들의 대표적인 예이다[10].

구문 형태소열을 인식하는 전처리 과정은 구문 분석 과정이 갖는 처리 부하를 크게 감소시킴으로서 전체 검색시스템의 성능을 향상시킬 수 있다.

2.2. 규칙 기반의 구문 분석

형태소 분석과 태깅을 마친 질의 문장은 구문 분석 과정을 거쳐서 단어들의 관계성을 얻기 위해 문장의 문법 구조를 해석하게 된다. 자연어 질의 문장의 구문 분석 과정은 단순한 형태소 열이 나타낼 수 없는 중요한 의미적 관계를 추출할 수 있다[2].

자연어 문장을 구문 분석하는 방법은 여러 가지가 있으나, 본 연구에서는 구 구조 규칙을 기반으로 하는 상향식 분석 방법을 변형하여 적용한다. 상향식 분석은 전체 문장의 구조 해석에 실패하더라도 부분적으로 완성된 중간 구조를 다음 과정에 활용할 수 있기 때문에 완전 분석 성공률이 낮은 자연어 문장의 분석에 매우 적합하다.

2.3. 구문 트리 순회와 복합 명사

한국어 질의 문장에 의한 검색에서 중심 역할을 하는 명사 색인어들은 복합 명사의 형태로 빈번하게 사용된다. 복합 명사의 문법적 규정은 매우 약하고 띄어쓰기 문제와도 긴밀하게 관련되어 있다. 두 개 이상의 명사를 붙여 쓴 형태이거나, 띄어 쓴 두 개의 단일명사이거나 대부분의 경우 모두 문법적으로 오류가 되지 않는다[4].

문서와 질의 문장에서 나타나는 복합 명사들의 길이에 대한 통계자료를 참조하면 길이가 3까지인 복합명사가 전체의 96.8%를 차지하며, 길이가 4 이상인 복합명사 매우 드물게 나타난다. 반대로, 한국어에서 4음절 이상의 단일어는 극히 드물기 때문에 4음절 이상의 명사라면 복합어일 확률이 매우 높다고

볼 수 있다[4,6].

2.3.1. 복합명사 조합에 의한 다중 검색

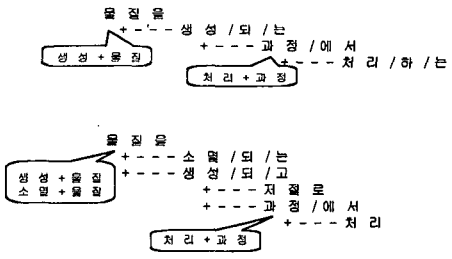
구문 분석 결과인 구문 트리에서 다양한 구조로 나타나는 어휘들은 같은 검색 의도를 말하는 색인어들일 수 있다. 구문 트리를 순회하면서 문법 구조적으로 연관된 어휘들을 추출하여 복합 명사 형태로 조합하여 색인어로 확장하고, 이들을 색인어로 다중 색인한다. 조사가 생략된 인접 명사, 관형격 조사로 연결된 피수식 명사, 피수식 내포문과 명사 등 복합 명사로 조합되는 문법적 구조를 분류하였다.

형태소 수준의 복합 명사 조합 방법이 의미 없는 조합을 많이 생성하는 것에 비해서 구문 트리의 구조에 기반한 복합 명사 조합은 사용자의 검색 의도를 반영한 조합이 가능하다. 2단어 이상의 복합 명사 확장 방법은 기존의 연구 방법을 따르며, 길이 4 이상의 복합 명사는 실제로 사용빈도가 극히 낮고 오히려 검색 성능을 떨어뜨리는 경향이 있으므로 길이 3까지 조합한다[2,4].

2.3.2. 복합명사 분할에 의한 다중 검색

4음절 이상의 단일 명사는 복합 명사는 복합어임을 검사하고 분할 과정을 거쳐서 입력 질의를 확장하고, 이들을 검색어로 다중 검색을 한다. 복합어는 두 개 또는 세 개 단어의 조합으로 분할한 다음 색인데이터베이스를 다중 검색한다.

이때 분할된 각 단일 명사들은 모두 색인 데이터베이스에 독립적으로 등록된 색인어이거나, 접두사 또는 접미사이어야 한다. 분할된 단일 명사들은 복합명사에 비해 특정성이 떨어지게 되지만 검색에서 재현율을 높일 수 있다.



[그림 2] 구문트리와 복합명사 조합 및 분할

2.4. 이형 표기 및 축약 표기 검색어의 확장

음역어 등의 예에서 흔히 볼 수 있는 이형 표기 용어들은 검색어로서 확장되고, 다중 검색한다. 이와 같은 이형태들은 미리 수집되어 이형태 사전으로 등록되어야 한다. 흔히 **'알고리즘/알고리듬/앨고리든/앨고리즘','데이터/데이타','운영체제/운영체계'**등의 용어가 문서나 질의 문장에서 이형 표기로 빈번하게 출현한다.

검색을 위해 질의 확장될 수 있는 또 다른 한 가지의 예는 축약 표기 용어들이며, 다른 종류의 이형 표기 용어로 취급할 수 있다. 축약 표기 용어와 이에 대응하는 기본 색인어들은 사전에 등록되어 질의 문장에 입력된 색인어를 양방향으로 확장하고, 검색 시스템은 이들을 다중 검색한다. **'정통부(정보통신부)','노조(노동조합)'**등이 축약 표기 검색어의 예가 된다. 야후, 구글, 네이버 등의 기존의 여러 검색시스템에서 예의 이형 표기 용어들을 질의에 포함시켜 검색을 시도하면 결과 문서의 집합이 크게 다름을 볼 수 있다.

3. 실 험

분할된 단일 명사들은 복합명사에 비해 특정성이 떨어지게 되지만 검색에서 재현율을 높일 수 있고, 문장에 나타나는 단어들로 명사구를 조합하게 되면 특정성이 큰 색인어를 만드는 것이 되어 정확도를 높일 수 있다. 따라서 복합명사의 분할 및 조합이 정보 검색 시스템의 성능 향상에 도움이 된다.

본 연구의 방법은 여러 실험에서 사용된 바 있는 4,400여 문장을 포함하는 KTSET 문서들과 교내의 개인, 학과, 부서 등의 홈페이지 등의 내용를 대상으로 검색자에게 임의의 질의 문장을 입력하도록 하였다. 제한없는 임의의 검색자들에게 이들 문서들을 검색하기 위한 한국어 자연어 질의 문장들을 수집하여 약 4,900여 질의 문장을 대상으로 본 연구의 방법을 시도하였다.

입력된 질의 문장을 구문 형태소 처리를 거쳐 구문 분석하는 과정은 일반 문서들에 포함된 문장을 구문 분석 하는 과정에 비해 성공률도 높고, 분석 성능도 높았다[10]. 검색 성능으로 볼 때, 검색 재현율은 11.3%가량의 향상을 보였다. 특히 이형 표기나 축약 표기 용어의 등록은 재현율 향상에 크게 기여하고 있다. 전반적으로는 정확도는 약 4.7% 정도로 다소 높아졌으나, 일부 검색 시도에서는 거리가 먼 명사들의 조합과 의미없는 분할이 일어나는 경우도 나타나고 있어서 이에 대한 분석이 필요하다. 반면에 질의 문장에 대한 검색어의 개수는 검색어 확장을 하지 않는 경우에 비해 약 2.7배 증가하였다.

검색 대상 문서들의 문장들에 대한 색인 데이터베이스 구축 과정에서도 구문 분석과 복합 명사 조합이나 분할 등의 동일한 과정을 거치는 것이 검색어 일치 측면에서 이상적이지만, 입력되는 질의 문장에 비해 웹 문서의 특성 상 문법적 오류를 갖는 문장들이 많이 포함되어 구문분석의 성능을 저하시키고 분할과 조합 등 과정에 상당한 오류를 발생시킬 수 있으므로 추가적인 실험이 필요하다. 본 실험에서 검색 대상 문서의 분량이 실험 코퍼스로서는 적은 양이기 때문에 웹 문서를 검색하는 사용자들의 자연어 질의를 추가 수집하고, 메타검색기의 역할로서 검색 키워드를 추출하는 방법으로 실험을 계속하고 있다.

4. 결론

본 논문에서는 정보 검색 시스템의 인터페이스로서 한국어 자연어 질의 문장을 입력하고 검색하기 위한 처리 과정을 제안하였다. 자연어 질의 문장에 대한 기본적인 자연어 처리 과정을 포함하며, 명사 형태의 색인어 뿐만 아니라 문장의 구조가 전달하는 정확한 문법 관계를 얻기 위해 구문 분석 과정도 포함된다. 형태소 분석, 구문 형태소 추출, 구문 분석 결과로부터 검색 키워드 추출 과정이 주요 처리 과정이다. 입력 질의 문장을 구문 분석 과정을 통해 키워드들 사이의 의미 관계를 획득할 수 있고, 이를 반영한 복합명사의 분할과 조합, 이형태 질의 확장을 통해 다중 검색함으로서 검색 결과의 재현율과 정확도를 높일 수 있음을 보여준다.

검색 시스템 사용자의 행태로 볼 때, 사용자들은 대부분 검색 결과 중 상위 20개 가량의 문서만을 참조하는 것으로 분석된다[7]. 따라서 검색 목표가 되는 문서들이 실제로 상위 랭킹 문서가 되도록 가중치 계산에 대한 정밀한 연구가 병행되는 것이 필요하며, 이형 표기의 색인어나 축약 표기 용어들에 대한 사전은 실험적 규모 이상의 통계적 수집에 의한 구축이 필요하다.

참고문헌

[1] Chengxiang Zhai, "Fast statistical parsing of noun phrases for document indexing," Fifth conference on applied natural language processing, pp.312-319, 1997

[2] 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.

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

[4] 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.

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

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

[7] 박소연, 이준호, "로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구", 정보관리학회지 제19권 제3호, 2002.

[8] 이공주, 김재훈, "규칙에 기반한 한국어 부분 구문분석기의 구현", 정보처리학회논문지, 제10권 B제4호, 2003.

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

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