의미 기반의 질의 분석 및 확장

Question Analysis and Expansion based on Semantics

신승은*, 박희근**, 서영훈**
(주)코난테크놀로지*, 충북대학교 전기전자컴퓨터공학부**

Seung-Eun Shin(seshin@nlp.chungbuk.ac.kr)*, Hee-Guen Park(pinetree@nlp.chungbuk.ac.kr)**,
Young-Hoon Seo(yhseo@chungbuk.ac.kr)**

요약

본 논문에서는 효율적인 정보검색을 위한 의미 기반의 질의 분석 및 확장을 제안한다. 기존의 정보검색 시스템들은 사용자 질의로 자연언어 질의를 허용하고 있지만 단순히 명사 단어의 색인어를 사용자 질의로부터 추출하여 정보검색에 활용하기 때문에 사용자의 질의 의도를 반영한 정보검색을 하지 못한다. 이러한 문제점을 해결하기 위해서 의미 기반 질의 분석 및 확장은 사용자의 질의를 의미적으로 분석하여, 질의 유형을 결정하고 의미 자질들을 추출한다. 추출된 의미 자질들과 정답을 표현하기 위해 사용되는 구문 구조를 이용하여 사용자 질의를 확장한다. 또한 확장된 질의를 이용하여 정답을 포함하는 관련문서들을 정보검색 결과의 상위에 랭크시킬 수 있는 방법을 제시한다. 비교적 짧지만 사용자의 질의 의도를 충분히 표현하고 있는 자연언어 질의에 대한 의미 기반의 질의 분석 및 확장을 통해 정보검색의 정확률을 향상시킬 수 있음을 보였다.

■ 중심어 : | 질의 분석 | 질의 확장 | 의미 | 정보검색 |

Abstract

This paper describes a question analysis and expansion based on semantics for en efficient information retrieval. Results of all information retrieval systems include many non-relevant documents because the index cannot naturally reflect the contents of documents and because queries used in information retrieval systems cannot represent enough information in user's question. To solve this problem, we analyze user's question semantically, determine the answer type, and extract semantic features. And then we expand user's question using them and syntactic structures which are used to represent the answer. Our similarity is to rank documents which include expanded queries in high position. Especially, we found that an efficient document retrieval is possible by a question analysis and expansion based on semantics on natural language questions which are comparatively short but fully expressing the information demand of users.

■ keyword : | Question Analysis | Query Expansion | Semantics | Information Retrieval |

I. 서론

인터넷과 더불어 정보의 양이 급증하고 정보에 대한 요구 형태가 다양해지면서 정보검색에 관한 관심은 폭발적으로 증가하고 있다. 정보검색 기술은 90년대 후반부터 인터넷의 발전과 더불어 상업적 응용이 확대되면

* 본 논문은 2006년도 충북대학교 학술연구 지원사업의 연구비지원에 의하여 연구되었습니다.
(This work was supported by the research grant of the Chungbuk National University in 2006)

접수번호 : #070329-001
접수일자 : 2007년 03월 29일
심사완료일 : 2007년 05월 15일
교신저자 : 서영훈, e-mail : yhseo@chungbuk.ac.kr

서 급속히 발전하고 있다. 현재까지 정보검색의 발전은 단어에 대한 통계적인 모델을 기반으로 이루어졌다. 텍스트, 문장, 문장 구성 성분이 가지고 있는 의미가 명시적으로 분석, 표현되어 사용되기 보다는 통계적인 방법으로 정보검색에서 비슷한 효과를 얻으려고 노력하고 있다. 그러나 효율적인 정보검색을 위해 자연언어처리 기술이 보다 적극적으로 적용되어 양질의 색인어를 추출하는 방법에만 그치지 않고, 텍스트 및 문장의 구성 요소의 의미를 분석하여 보다 정교한 검색이 수행될 수 있도록 하여야 한다.

최근에는 웹 문서의 양이 급격히 증가하면서 대용량 문서 색인 기술과 함께 수만에서 수십만의 검색 결과 중에서 사용자가 원하는 의도에 맞는 정보를 정확하게 찾아주는 효과적인 검색 기술이 요구되고 있다. 특히 웹과 같은 영역에서의 정보검색은 다양한 분야의 정보들이 서로 연결되어 있는 상황에서 빠르고 정확하게 찾아주는 점에 초점을 맞추어 기술 개발이 집중적으로 이루어지고 있다.

그러나 모든 정보검색 시스템들은 검색 결과에 원하지 않는 문서가 포함되어 있다. 이러한 문제점의 근본 원인 중 한 가지는 시스템이 검색에 사용하는 질의가 사용자의 정보 요구를 제대로 표현하지 못하고 있기 때문이다[1]. 이것은 단순히 명사 단어의 색인어를 사용자 질의로부터 추출하여 정보검색에 활용하기 때문이다. 따라서 질의 의도를 충분히 표현하고 있는 사용자 질의라도 검색 결과에 원하지 않는 문서가 포함된다. 이러한 문제를 해결하기 위해서는 사용자 질의로부터 질의 의도를 정확하게 파악하고 단어의 사용 의미를 분석하는 것이 필요하다. 즉 사용자 질의의 구성 성분이 가지고 있는 의미가 명시적으로 분석되어 정보검색에 활용되어야 한다.

II. 관련 연구

정보검색에서 질의 분석은 사용자의 질의를 분석하여 검색에 사용할 질의를 구성하는 과정이다. 대부분의 정보검색 시스템들은 사용자의 질의로부터 색인어들을 추출하고, 추출한 색인어들에 대해 질의 확장을 수행하여 시스템 질의를 구성한다. 그러나 통계적인 정보검색 모델에서는 질의에서 색인어의 사용 의미를 고려하지 않기 때문에 사용자의 질의 의도를 파악하지 못하고 정확한 정보검색을 어렵게 한다. 이것은 부정확한 정보검색의 근본적인 원인이 된다[2-4].

자연언어처리 기술을 이용하여 단어의 의미를 정보검색에 활용하고자 하는 연구들이 있었다. 기존의 의미 기반 정보검색에서는 기본적으로 자연언어 텍스트로부터 적절히 색인어를 추출하여 이를 통계적인 검색 모델에 반영하고 있다. 이를 위해서는 텍스트 문서의 분석에 자연언어처리 기술을 적용하여 의미있는 색인어를 추출하는 것이 무엇보다도 중요하다[5].

90년대 중반에 와서 TREC-5부터 NLP(Natural Language Processing) SIT(Special Interest Track)가 만들어지면서 본격적으로 자연언어처리 기술을 이용하여 정보검색의 효과 향상을 위한 다양한 연구들이 진행되었다. CLARIT[6]나 IRENA[7] 시스템에서는 문서 텍스트의 명사구 구문분석을 통한 구 기반 색인(phrase-based indexing)이 시도되었으며, TSAs (Tree Structure Analytics)라는 구문구조 분석에 의한 색인 방법에서는 질의와 텍스트 문서의 구문분석 결과를 트리 구문구조로 매칭하여 검색을 수행하였다[8]. 전반적으로 이 시기에는 구절 색인어(phrasal term)와 고유명사(proper name) 색인어 추출에 자연언어처리 기술을 적용하는 다양한 색인 방법들에 관한 연구가 본격적으로 이루어졌다. 90년대 후반에 들어와서는 FERRET 등과 같이 프레임, 스크립트 등의 지식 표현을 이용한 개념 색인(conceptual indexing)에 관한 연구도 실험적으로 시도되었다[9]. 이러한 연구는 단어의 의미를 직접 색인에 반영하는 방법이었다.

이렇듯 지금까지의 자연언어처리 기술을 이용한 색인 방법에 관한 연구는 근본적으로 문서의 의미적 내용을 명사 구 수준의 색인단위로 표현하는 한계를 가지고 있었다. 기존의 의미 기반 정보검색은 의미있는 색인어를 추출하기 위해 단어의 의미를 분석하여 색인에 반영하고자 하는 연구들이었다.

질의 확장(query expansion)은 사용자가 제시한 질의

에 이와 관련된 단어들을 추가해서 문서를 검색함으로써 보다 연관된 문서들을 검색하고자 하는 것이다. 시소러스를 이용해서 질의를 확장하거나 코퍼스에 나타나는 단어들의 형태에 따라 단어들의 상호관계를 분석해서 질의를 확장하는 방법 등이 있다. 자동적인 질의 확장 방법은 재현율은 높일 수 있으나 높은 순위의 문서들에서 정확률은 일반적으로 낮아지기 때문에 실용적이지 않다[10]. 이에 대한 보완기법으로 사용자의 적합성 피드백을 이용하여 새로운 질의어를 형성하는 방법은 검색된 문서를 기반으로 해서 사용자가 직접 관련 문서와 비관련 문서를 판단해야 하고 사용자의 판단의 질에 매우 종속적이다. 최근에는 사용자가 제시한 질의에 대한 검색된 문서들을 분석해서 질의를 자동으로 확장하는 방법 등이 연구되고 있다[11][12].

질의에 대한 어휘 대역어(lexical paraphrase)를 생성하는 연구[13]는 워드넷(WordNet)과 품사 정보(part-of- speech information)를 이용하여 질의에 있는 내용을 나타내는 명사와 동사들(content words)에 대한 동의어/유의어를 질의의 어휘 대역어로써 생성하는 연구로, 질의의 어휘 대역어를 순위화하기 위해 말뭉치로부터 얻은 통계 정보를 사용한다. 그러나 이러한 방법은 통계 정보를 이용하여 어휘 중의성을 해결할 수 있지만, 질의의 의미를 고려하지 않고 단순히 질의에 대한 어휘 대역어만을 생성한다. 따라서 술어가 생략된 축약된 형태의 질의의 경우에 동의어/유의어를 이용한 질의 확장 방법과 유사하다.

기존의 질의 확장은 사용자 질의보다 확장된 질의에 낮은 가중치를 부여하기 때문에 확장된 질의에 의해 검색된 결과들을 정보검색 결과의 상위에서 찾아보기 어렵다. 이것은 실제로 질의 확장이 수만에서 수십만의 검색 결과를 제공하는 정보검색에 별 영향을 주지 못함을 의미한다. 따라서 효율적인 정보검색을 위해 사용자 질의의 색인어들을 직접 포함하지 않더라도 관련도가 높은 문서들을 상위로 랭크시킬 수 있는 방법이 필요하다.

III. 의미 기반의 질의 분석 및 확장

기존의 정보검색에서의 질의 분석은 사용자의 질의로부터 색인어를 추출하여 확장을 하거나 정답 유형을 결정하는 역할만을 했으며, 색인에 의미를 반영하더라도 단어 자체의 의미만을 고려하였다. 그러나 의미 기반 질의 분석은 질의로부터 정답의 세부 유형 결정과 단어가 어떤 의미로 사용되었는지 그 사용 의도를 나타내는 의미 자질 추출 과정을 수행하고, 세부 정답 유형과 의미 자질을 이용한 질의 확장을 통해 정보검색 시스템의 상위 문서에서의 정확률 향상을 목적으로 한다.

정답 유형이 '저자'인 질의들을 살펴보자.
(저자 1) 동의보감을 저술한 사람은 누구인가?
(저자 2) 햄릿의 저자는?

정답 유형이 '저자'인 위의 자연언어 질의를 살펴보면 '저자'에 대한 질의에 공통적으로 사용되는 의미 자질들이 있음을 알 수 있다. '저자'에 대한 질의에 사용된 공통적인 의미 자질들은 저서명(동의보감, 햄릿), 저자를 표현하기 위한 명사(저자), 저자를 표현하기 위한 용언(저술하), 인물을 표현하기 위한 명사(사람), 인물의문사(누구) 등이며, 질의는 이러한 공통된 의미 자질들을 이용하여 구성되고 있다.

통계적인 정보검색 모델의 경우, (저자 1)에서 색인어(동의보감, 저술, 사람, 누구)를 추출하여 문서와의 유사도를 계산함으로써 관련문서를 검색한다. 그러므로 "허준은 동의보감을 편찬하였다"라는 문장을 포함하는 문서보다 단순히 색인어를 많이 포함하는 문서를 유사도가 높은 관련문서로 선택하는 문제점을 가지고 있다. 심지어 (저자 2)의 경우 '저자'란 단어가 많이 포함되어 있는 인터넷 서점이 상위에 랭크되기도 한다.

정보검색 시스템들은 유사도를 계산할 때 확장한 질의보다 사용자 질의의 색인어에 높은 가중치를 부여한다. 따라서 (저자 1)에 대한 질의 확장을 통해 '저술'을 '편찬'으로 확장하더라도 위 문제를 근본적으로 해결할 수는 없다. 이를 해결하기 위해서는 단순히 색인어의 통계 정보를 이용하는 것이 아니라 (저자 1)이 '저서명

(동의보감)'의 '저자'를 묻고 있다는 것을 분석하고, '동의보감'의 '저자'를 나타내는 문서를 상위 관련문서로 제공해야 한다. 따라서 자연언어 질의에 대한 정답 유형이 '인물-저자'이며, '동의보감'이 '저서명'이라는 것을 의미적으로 분석하고 정보검색에 활용해야 한다.

정보검색의 성능 향상을 위해서 위의 문제점들을 해결할 수 있는 방안으로 질의의 의도를 파악하고 단어의 의미를 분석하는 것이 필요하다. 따라서 의미 기반 질의 분석은 (저자 1)에 대해 다음과 같이 질의에서 의미 자질들을 추출하고, 이를 활용한 질의 확장을 수행하여 정보검색의 정확률을 향상시키고자 한다.

<의미 기반 질의 분석 및 확장 결과>
(저자 1) 동의보감을 저술한 사람은 누구인가?
- 정답 유형 : 인물
- 세부 정답 유형 : 저자
- 저서명 : 동의보감
- 질의 확장 : 동의보감을 (저술하|편찬하|쓰|.....)
  동의보감의 (저자|작가|글쓴이|.....)

(저자 1)에 대해 위와 같이 세부 정답 유형과 질의에 내포된 의미 자질들을 추출하고, 의미 자질들을 이용하여 질의 확장을 수행하여 정보검색에 활용한다면 "허준은 동의보감을 편찬하였다"라는 문장을 포함하는 문서를 유사도가 높은 관련 문서로 추출할 수 있으며, 상위 문서에서의 정확률을 향상시킬 수 있다.

1. 의미 기반의 질의 분석

위의 <의미 기반 질의 분석 및 확장 결과>와 같이 사용자 질의로부터 세부 정답 유형을 결정하고, 의미 자질을 추출하기 위해 세부 정답 유형별 의미 자질들을 정의하고, 의미 자질 추출 규칙을 구성하였다.

의미 자질은 비슷한 유형의 질의에 공통적으로 나타나는 개념으로, 사용자의 질의 의도를 표현하기 위해 해당 질의에 나타나는 개념들이다. 의미 자질은 단어 자체의 사전적 의미가 아니라 단어의 의미를 포괄하는 문장에서의 단어의 역할을 나타내는 개념이라 할 수 있다.

TREC-8/TREC-9의 질의 말뭉치와 웹에서 643개의 인물 관련 질의들로부터 질의 유형과 의미 자질들을 정의하고, 동의어/유의어 사전을 이용하여 질의 유형에 따라 명사와 용언의 의미 자질 사전을 구축하였다. 세부 정답 유형은 전체 24개의 유형(author, family, prizewinner, politician, developer, inventor, scholar, entertainer, player 등)으로 분류되었으며, 의미 자질은 전체 125개가 정의되었고, 의미 자질 사전은 1761개의 명사와 278개의 용언들로 구성되었다.

의미 자질 추출 규칙은 사용자 질의의 의미와 구문구조를 고려하기 위해 세부 정답 유형에 따라 의미 자질과 문법형태소로 구성되었다. [그림 1]은 의미 자질 추출 규칙의 BNF(Backus- Naur Form) 표현이고, 그림 아래는 '저자'에 대한 의미 자질 추출 규칙의 예를 보여주고 있다. 의미 자질 추출 규칙은 643개의 인물 관련 자연언어 질의 말뭉치를 기반으로 작성되었으며, 세부 정답 유형에 따라 의미 자질들과 문법형태소로 구성된 어절 정보들로 구성된다.

```
〈Extraction Rule〉 ::= 〈Word List〉
〈Word List〉 ::= 〈Word Information〉 |
                 〈Word List〉〈Word Information〉
〈Word Information〉 ::= '(' 〈Semantic Feature〉 ')' |
        '(' 〈Semantic Feature〉〈Grammatical Morpheme〉 ')'
〈Semantic Feature〉 ::= '@Who' | '@Person' |
                 '@Author_N' | '@Author_V' | '#Book' |·····
〈Grammatical Morpheme〉 ::= 'jc' | 'jx' | 'jm' | 'etm' | 'co' |
                 'ef' | 'cj' | 'co+etm'
```

그림 1. 의미 자질 추출 규칙의 BNF 표현

'저자'에 대한 의미 자질 추출 규칙의 예

.....

4. (#Book jc) (@Author_V etm)
   (@Person|@Author_N jx?) (@Who)?
5. (#Book jm) (@Author_N jx?) (@Who)?

.....

규칙의 구성에서 문장에서의 위치가 자유로운 의미 자질들은 제외하였다. 이러한 의미 자질들은 사전이나 규칙을 통해 먼저 추출되고, 남은 질의에 대해 규칙을

이용한 의미 자질 추출을 수행한다. 이것은 자유로운 어순을 갖는 한국어의 특성을 반영하고 규칙의 수를 줄이기 위함이다. 어절 정보는 실질형태소의 의미 자질로만 구성되는 것과 실질형태소의 의미 자질과 문법형태소로 구성되는 것이 있으며, 의미 자질 추출 규칙은 어절 정보들의 리스트로 구성된다.

다음은 의미 자질 추출 규칙을 이용한 의미 자질 추출 결과이다. 사용자 질의로부터 세부 정답 유형을 결정하고, 의미 자질들을 추출할 수 있다. 또한, 구문구조가 다른 사용자 질의를 똑같이 분석하여 의미 자질들을 추출할 수 있다.

<의미 자질 추출 결과>

- 질의 1 : 동의보감을 저술한 사람은 누구인가?
- 적용규칙 : 4. (#Book jc) (@Author_V etm) (@Person|@Author_N jx?) (@Who)?
- 세부 정답 유형 : 저자
- 의미 자질 : 저서명(동의보감)

- 질의 2 : 햄릿의 저자는?
- 적용규칙 : 5. (#Book jm) (@Author_N jx?) (@Who)?
- 세부 정답 유형 : 저자
- 의미 자질 : 저서명(햄릿)

인물 관련 질의들을 세부 정답 유형에 따라 전체 24개의 세부 정답 유형으로 분류하였다. 24개의 세부 정답 유형 중 '기타' 유형을 제외한 23개의 세부 정답 유형에 대한 의미 자질 추출 규칙을 구축하였고, 규칙을 적용하여 의미 자질을 추출한다. 따라서 규칙이 정의되지 않은 세부 정답 유형에 대한 질의들과 규칙이 적용되지 않은 질의들에 대한 의미 자질 추출 방법이 필요하다. 따라서 의미 자질 추출 규칙에 의해 분석되지 않는 질의들에 대해서는 자연언어 질의의 구문구조 특성을 이용하여 의미 자질을 추출한다.

인물 관련 질의들을 용언을 포함하는 질의와 용언을 포함하지 않는 질의로 구분하였다. 용언을 포함하는 질의는 이벤트 관련 용언(Event_V), 인물명사(Person) 혹은 특성명사(Property), 인물 의문사(Who)로 구성되며, 용언을 포함하지 않는 질의는 특성명사와 특성명사를 한정하는 명사구로 구성된다. 따라서 구문구조를 이용한 의미 자질 추출은 이벤트용언, 특성명사, 인물명사, 관련 명사구를 구문구조에 따라 추출한다. 질의에서 명사구를 정확하게 분석하는 것 자체가 어려운 연구이나 질의의 어절 수가 적은 자연언어 질의에 대해서는 단순 명사나 명사 나열 등 간단한 명사구 형태를 적용하여 분석하는 것이 가능하다.

용언을 포함하는 질의와 용언을 포함하지 않는 질의의 구문구조 특성에 따른 어절 정보 리스트는 다음과 같다.

<용언을 포함하는 질의 어절 정보 리스트>

1. (NP jc) (#Event_V etm) (@Person|@Property_N) (@Who)?
2. (NP jx) (@Who) (#Event_V ef)
3. (@Who) (NP jc) (#Event_V ef)

<용언을 포함하지 않는 질의 어절 정보 리스트>

1. (NP jm) (@Property_N) (@Who)?
2. (NP jc) (@Property_N) (@Who)?
3. (@Property_N) (@Who)?

'#Event_V'는 사용자 질의로부터 추출할 용언을 의미하고 '@' 속성의 의미 자질은 의미 자질 사전을 이용하여 의미 자질을 부여한다.

의미 자질 추출 규칙이 적용되지 않는 질의들에 대해서는 위와 같이 질의 어절 정보 리스트에 따라 의미 자질 추출을 수행한다. 사용자 질의의 세부 질의 범주를 결정할 수 없고, '%' 속성을 갖는 의미 자질과 특성명사 이외의 의미 자질들을 추출할 수는 없다. 그러나 동의어/유의어 사전을 이용하여 이벤트용언과 특성명사에 대한 질의 확장을 수행하여 의미 자질 추출 규칙을 적용한 질의들과 유사하게 정보검색에서 성능 향상을 보일 수 있다.

다음은 질의 어절 정보 리스트를 이용한 의미 자질 추출 결과이다.

<의미 자질 추출 결과>

- 질의 3 : 남극에 도착한 최초의 사람은?
- 적용 어절 정보 리스트 :
  (NP jc) (#Event_V etm)
  (@Person|@Property_N) (@Who)?
- 정답 유형 : 인물
- 의미 자질 : 단서부사(최초), 이벤트용언(도착하), 명사구(남극)

2. 의미 기반의 질의 확장

일반적인 정보검색 시스템은 확장된 질의에 대해 사용자 질의보다 낮은 가중치를 부여한다. 이것은 질의에서 색인어의 사용 의미를 분석하지 않고 단지 사용자 질의에 대한 색인어가 확장된 질의보다 질의 의도를 잘 표현하고 있다고 판단하기 때문이다. 따라서 수십에서 수십만 이상의 정보검색 결과 중 상위에 랭크된 문서들은 대부분 사용자 질의로부터 추출된 색인어들을 통계적으로 많이 포함하는 문서들이다. 이러한 문제점을 해결하고 관련된 문서들을 검색하고자 하는 질의 확장의 목적을 위해서는 사용자 질의의 질의 의도를 정확하게 파악하고 의미를 분석하여 질의 확장에 활용하고 생성된 질의에 높은 가중치를 부여함으로써 관련 문서를 상위에 랭크시켜야 한다.

상위 문서에서의 정확률을 높이기 위해 본 논문에서는 의미 자질을 이용하여 질의 확장을 수행한다. 질의 확장은 사용자 질의로부터 추출된 의미 자질을 이용하여 정보검색의 성능 향상을 위해 질의를 생성하는 과정으로, 질의 확장 규칙과 의미 자질 사전, 사용자 질의로부터 추출된 의미 자질들을 이용하여 질의를 생성한다. 생성된 질의를 통해 문서를 검색하고 검색된 문서를 검색 결과의 상위에 랭크시킴으로써 상위 문서에서의 정확률을 높이고, 사용자의 검색 만족도를 높일 수 있다.

질의 확장 규칙은 인물 관련 자연언어 질의 말뭉치들의 정답문서를 수집하고, 각각의 문서에서 정답을 표현하기 위해 사용되는 의미 자질들과 구문구조를 이용하여 구축되었으며, 의미 자질과 문법형태소로 표현된다. 의미 자질 추출 결과, 의미 자질 사전, 동의어/유의어 사전을 이용하여 질의 확장 규칙의 의미 자질 부분을 생성하고 문법형태소와 결합하여 질의 확장을 수행한다. 다음은 '저자'에 대한 질의 확장 규칙의 예이다.

<'저자'에 대한 질의 확장 규칙의 예>

1. ($Book oj) (&Author_V)
2. ($Book jm) (&Author_N)
3. ($Book_Info jc) (&About) (!Genre oj)
   (&Author_V)
4. ($Book co+etm) (!Genre oj) (&Author_V)
5. ($Book co+etm) (!Genre jm) (&Author_N)

질의 1의 의미 자질 추출 결과와 '저자'에 대한 질의 확장 규칙을 이용하여 질의 확장을 수행한 결과는 다음과 같다.

- 질의 확장 규칙 1 : ($Book oj) (&Author_V)
  - $Book : 동의보감
  - &Author_V : 저술하, 편찬하, 쓰, 짓, ...
  - 질의 확장 결과
    - 동의보감을 (저술하|편찬하|쓰|짓|.....)
- 질의 확장 규칙 2 : ($Book jm) (&Author_N)
  - $Book : 동의보감
  - &Author_N : 저자, 작가, 지은이, ...
  - 질의 확장 결과
    - 동의보감의 (저자|작가|지은이|.....)

의미 자질 추출 규칙이 적용되지 않고 질의 어절 정보 리스트를 이용하여 의미 자질을 추출한 질의에 대해서는 구문구조 특성을 이용하여 질의 확장 규칙을 작성하고, 의미 자질의 동의어/유의어를 이용하여 생성한다. 다음은 구문구조 특성을 이용한 질의 확장 규칙이다.

<용언을 포함하는 질의의 질의 확장 규칙>

1. (NP jc) (#Event_V etm)
   (@Person|@Property_N) (@Who)?
   ⇒ (NP jc) (Event_V)
2. (NP jx) (@Who) (#Event_V ef)

⇒ (NP oj) (Event_V)

3. (@Who) <NP jc> (#Event_V ef)
   ⇒ (NP jc) (Event_V)

<용언을 포함하지 않는 질의의 질의 확장 규칙>

1. (NP jm) (@Property_N) (@Who)?
   ⇒ (NP jm) (Property_N)
2. (NP jc) (@Property_N) (@Who)?
   ⇒ (NP jc) (Property_N)
3. (@Property_N) (Who)?
   ⇒ (Property_N)

자연언어 질의의 구문구조 특성에 따른 질의 확장 규칙에서 사용된 격조사(jc)는 질의 생성과정에서 사용자 질의에서 쓰인 격조사를 그대로 사용하며, 'NP', 'Event_V', 'Property_N'는 동의어/유사어 사전을 이용하여 생성한다.

질의 3의 의미 자질 추출 결과와 구문구조 특성에 따른 질의 확장 규칙을 이용하여 질의 확장을 수행한 결과는 다음과 같다.

<질의 3 질의 확장 결과>

- 질의 확장 규칙 : (NP jc) (Event_V)
  - NP : 남극, 에스 극, 남쪽 끝, ..., 남극점
  - Event_V : 도착하, 닿, 다다르, ..., 이르
  - 질의 확장 결과
    - (남극|에스 극|남쪽 끝|.....|남극점)에 (도착하|닿|다다르|.....|이르)

구문구조 특성에 따른 질의 확장은 질의 확장을 위해 의미 자질 사전을 이용하는 것이 아니라 동의어/유의어 사전을 이용한다. 이벤트용언, 특성명사, 명사구 등의 의미를 분석하지 않기 때문에 의미 자질을 이용한 다양한 형태의 질의 확장이 어렵다. 그러나 의미 자질 추출 규칙이 작성되지 않은 질의들에 자연언어 질의의 구문구조 특성을 이용한 질의 확장 방법으로 유연성과 효율성을 동시에 제공할 수 있다.

3. 질의와 문서의 유사도

의미 기반 질의 확장 결과를 정보검색에 사용할 때, 일반적인 색인어와 달리 높은 가중치를 부여해야 한다. 이것은 자연언어 질의로부터 각각의 의미 자질들을 추출하고, 의미 자질들과 정답을 표현하기 위해 사용되는 구문 구조에 의해 확장된 질의이기 때문이다. 질의 1에 대해서 일반적인 색인어인 '동의보감', '저술', '사람'이 많이 출현하는 문서보다 한번만 출현하더라도 '동의보감을 편찬하다'를 포함하는 문서가 관련문서일 가능성이 높다. 또한, 자연언어 질의의 구문 구조 특성을 이용하여 생성한 질의에도 일반 색인어보다 높은 가중치를 부여해야 한다.

사용자 질의의 의미 자질을 이용하여 생성한 질의는 일반적인 질의와 문서간의 유사도로 계산하지 않고, 질의를 포함하는 문서를 상위 관련 문서로 랭크시키기 위해 식 (1)과 같이 유사도를 계산한다. 식 (1)은 벡터 모델에서의 코사인 유사도와 결합된 형태이다. 생성된 질의를 포함하는 문서를 그렇지 않은 문서보다 상위에 랭크시키는 유사도 계산 방법으로써, d_i는 문서, q는 질의를 나타내며, 각 첨자는 질의의 종류를 나타낸다. q_u는 사용자 질의를 나타내고, $\vec{d_i}$는 문서에 대한 벡터, $\vec{q_{se}}$는 의미 기반 질의 벡터이며, $\vec{q_e}$ 일반적인 질의 확장에 의한 벡터이다. 'else'에 해당하는 유사도 계산 부분을 수정하여 다른 모델에서의 유사도 계산과 결합할 수도 있다.

실험에서 의미 기반 질의의 가중치는 1로 설정하고, 식(1)의 유사도 계산방식에 따라 질의와 문서간의 유사도를 계산하였다.

$$if\ \vec{d_i}\cdot\vec{q_{se}} \neq 0\ then\ \sim(d_i,q_u)=\vec{d_i}\cdot\vec{q_{se}} \qquad (1)$$

$$else\ \sim(d_i,q_u)=\frac{\vec{d_i}\cdot\vec{q_e}}{|\vec{d_i}||\vec{q_e}|}$$

IV. 실험 및 평가

의미 기반의 질의 분석 및 확장을 통한 정보검색의

성능 향상을 평가하기 위해 상용 정보검색 시스템인 Google과 Yahoo의 상위 N-문서에서 정확률을 비교하는 실험을 수행하였다. Google과 Yahoo를 통해 실험 질의 말뭉치로 추출된 100개의 인물 관련 자연언어 질의에 대한 검색을 수행하고, 각각의 질의에 대해 상위 30개의 웹 문서를 추출하여 실험 문서 집합을 구축하였다. 100개의 질의에 대해 상위 30개의 웹 문서를 검색한 결과, Google에서 1,896개의 웹 문서와 Yahoo에서 2,281개의 웹 문서를 수집하였다. 실험 문서 집합 구축 과정에서 링크 오류나 웹 문서 오류와 같은 비정상적인 웹 문서는 제외하였으며, 하나의 질의에 대한 모든 검색 결과가 30개 미만인 경우도 있었다. 검색 정확률 실험에서 객관적인 관련 문서 판단을 위해 질의에 대한 정답을 포함하는 문서만을 관련 문서로 판단하였다.

상위 N-문서에서의 정확률 비교 실험 결과는 [표 1]과 같다. Google과 Yahoo의 정보검색 결과와 각각에 제안한 방법을 적용하여 정보검색 결과를 재순위화하여 평가하였다.

표 1. 상위 N-문서에서의 정확률 비교 실험 결과

N	상위 N-문서에서의 정확률			
	Google	Google + 제안한 방법	Yahoo	Yahoo + 제안한 방법
3	0.584	0.803(+0.219)	0.580	0.804(+0.224)
5	0.585	0.770(+0.185)	0.557	0.743(+0.186)
10	0.548	0.646(+0.098)	0.511	0.604(+0.093)
15	0.523	0.591(+0.068)	0.478	0.541(+0.063)
20	0.489	0.541(+0.052)	0.458	0.506(+0.048)

[표 1]은 제안한 방법이 정보검색의 성능 향상에 효율적이라는 것을 보인다. 본 논문에서 제안한 방법을 Google과 Yahoo의 정보검색 결과에 적용했을 때 평균 +0.2215 (N=3), +0.1850 (N=5), +0.0955 (N=10)의 높은 정확률 향상을 보였다. 이것은 Google과 Yahoo의 상위 30개의 웹 문서를 이용하였을 때의 정확률 향상이며, 정보검색 결과 전체에 적용한다면 보다 높은 정확률 향상을 기대할 수 있을 것이다.

웹의 특성을 반영하여 Page Rank라는 랭킹 시스템을 도입하여 사용자 입장에서 높은 검색 효과를 얻는 서비스를 제공하고 있는 Google과 대표적인 주제별 검색엔진으로 데이터베이스의 크기는 작으나 양질의 검색결과를 제공하는 Yahoo보다 상위 문서에서 높은 정확률을 보였다. 이는 사용자 질의 의도를 파악하고, 문장에서 단어가 어떤 의미로 사용되었는가를 나타내는 의미 자질을 추출하여 정보검색에 활용함으로써 관련 문서를 상위에 직접적으로 랭크시켜 상위 문서에서의 정확률 향상을 목적으로 하기 때문이다.

V. 결론 및 향후 연구

정보검색에서 질의 의도를 충분히 반영하지 못하는 문제점을 해결하기 위해 본 논문에서는 의미 기반의 질의 분석 및 확장을 제안하였다. 정보검색의 성능 향상을 위해 의미 기반 질의 분석 및 확장 결과를 이용하여 관련문서를 정보검색 결과의 상위에 바로 랭크시킬 수 있는 문서 순위화 방법을 제시하였다.

웹에서 추출한 100개의 자연언어 질의에 대한 실험을 통해 제안한 방법을 평가하였다. Google과 Yahoo의 정보검색 결과와 상위 문서에 대한 정확률 비교 실험에서는 평균 +0.2215 (N=3), +0.1850 (N=5), +0.0955 (N=10)의 높은 정확률 향상을 보였다. 이 실험에서는 사용자 질의의도를 파악하고, 문장에서 단어가 어떤 의미로 사용되었는가를 나타내는 의미 자질을 추출하여 정보검색에 활용함으로써 관련 문서를 상위에 직접적으로 랭크시켜 상위 문서에서의 정확률을 향상시킬 수 있음을 보였다. 특히, 상위 문서에서의 정확률 향상은 정보검색 시스템의 성능 향상과 사용자의 검색 만족도도 크게 높일 수 있음을 의미한다.

본 논문에서 제안한 의미 기반 질의 분석 및 확장은 정답유형이 인물인 질의를 대상으로 실험하였기 때문에 다른 정답유형에 대한 실험이 필요하며, 상위 문서에서의 정확률뿐만 아니라 의미 자질의 다양한 활용 방안을 고안하여 정보검색의 성능 향상을 위해 적용해야 할 것이다.

참 고 문 헌

[1] 맹성현, "정보검색 기술의 현황과 발전방향", 정보과학회지, 제22권, 제4호, pp.6-14, 2004(4).

[2] G. Salton, E. Fox, and H. Wu, "Extended boolean information retrieval," Communication of the ACM, Vol.26, No.11, pp.1022-1036, 1983.

[3] G. Salton, *Automatic Text Processing*, Addison-Wesley, 1989.

[4] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," Journal of the ACM, pp.216-244, 1960.

[5] 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정, "의미 기반 정보검색", 정보과학회지, 제19권, 제10호, pp.7-18, 2001.

[6] C. Zhai, "Fast Statistical Parsing of Noun Phrases for Document Indexing," In Proceedings of the Fifth Conference of Applied Natural Language Processing, 1997.

[7] A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and T. P. van der Weide, "Phrase-based Information Retrieval," Journal of Information Processing & Management, Vol.34, Issue 6, pp.693-707, 1998(11).

[8] A. F. Smeaton, R. Odonnel, and F. Kelledy, "Indexing Structures Derived from Syntax in TREC-3: System Description," The Third Text Retrieval Conference (TREC-3), NIST Special Publication, pp.55-67, 1994.

[9] B. V. Dobrow, N. V. Loukachevitch, and T. N. Yudina, "Conceptual Indexing Using thematic Representation of Texts," TREC-6, 1997.

[10] L. Fitzpatrick and M. Dent, "Automatic Feedback Using Past Queries: Social Searching?," In Proc. 20'th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.306-313, 1997.

[11] J. J. Rocchio, "Relevance feedback in information retrieval," In The SMART Retrieval System-Experiments in Automatic Document Processing, Prentice Hall, pp.313-323, 1971.

[12] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," In Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp.4-11, 1996.

[13] I. Zukerman and B. Raskutti, "Lexical Query Paraphrasing for Document Retrieval," The 17th International Conference on Computational Linguistics, COLING 2002.

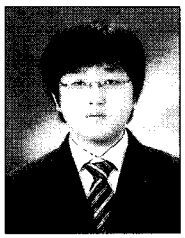
저 자 소 개

신 승 은(Seung-Eun Shin) 정회원

- 1999년 : 충북대학교 컴퓨터공학과(공학사)
- 2001년 : 충북대학교 컴퓨터공학과(공학석사)
- 2006년 : 충북대학교 컴퓨터공학과(공학박사)
- 2006년 ~ 2007년 4월 : 충북대학교 BK21 충북정보기술사업단 PostDoc
- 2007년 4월 ~ 현재 : (주)코난테크놀로지 검색4팀 선임연구원

<관심분야> : 정보검색, 자연언어처리

박 희 근(Hee-Guen Park) 준회원



- 2006년 : 충북대학교 컴퓨터공학과(공학사)
- 2006년 ~ 현재 : 충북대학교 컴퓨터공학과 석사과정

<관심분야> : 자연언어처리, 형태소분석, 구문분석

서 영 훈(Young-Hoon Seo) 종신회원

- 1983년 : 서울대학교 컴퓨터공학과(공학사)
- 1985년 : 서울대학교 컴퓨터공학과(공학석사)
- 1991년 : 서울대학교 컴퓨터공학과(공학박사)
- 1994년 ~ 1995년 : 미국 Carnegie-Mellon 대학 기계번역센터 객원교수
- 1988년 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 교수

<관심분야> : 정보검색, 자연언어처리, 기계번역