복합명사확장을 이용한 KRISTAL2000 DBMS 검색 성능 향상

서정현[0] 김광영 최성필
한국정보과학회 사무국
{kykim[0], jerryseo, spchoi}@kisti.re.kr

Using the Extension of Korean Compound Noun the improvement of KRISTAL2000 DBMS Retrieval System

Kwang-Young Kim[0] Jerry - Seo Cho Sung Pil - Choi
Group for Intelligent Information System, Korea Institute of Science and Technology Information

요 약

복합명사는 한국어에서 가장 빈번하게 나타나는 색인어의 한 형태로서, 영어권 중심의 정보검색모델로는 다루기가 어려운 언어 현상의 하나이다. 복합명사는 2개 이상의 단일어들의 조합으로 이루어져 있고, 그 형태 또한 여러 가지로 나타나기 때문에 색인과 검색의 큰 문제로 여겨져 왔다. 특히 한국어에서는 복합명사 분석이 어렵고 복잡하다. 그러므로 본 논문에서는 복합명사 질의어 대해서 질의어를 확장 또는 최적 방법을 이용하여 KRISTAL2000 DBMS의 성능을 향상 연구에 중점을 두었다.

1. 서 론

복합명사는 한국어에서 가장 빈번하게 나타나는 색인어의 한 형태로서, 영어권 중심의 정보검색모델로는 다루기가 어려운 언어 현상의 하나이다. 복합명사는 2개 이상의 단일어들의 조합으로 이루어져 있고, 그 형태 또한 여러 가지로 나타나기 때문에 색인과 검색의 큰 문제로 여겨져 왔다.[1] 특히 본 논문에서 제시한 사용자 질의어 확장은 일반 명사일 때 보다 복합 명사일 때 그 성능이 더 뛰어났다.

본 논문에서는 사용자 질의어 확장 방법를 제시하고, HANTEC Version2.0를 이용하여 사용자 질의어 확장 한 것과 확장을 하지 않은 결과를 실험하였다.

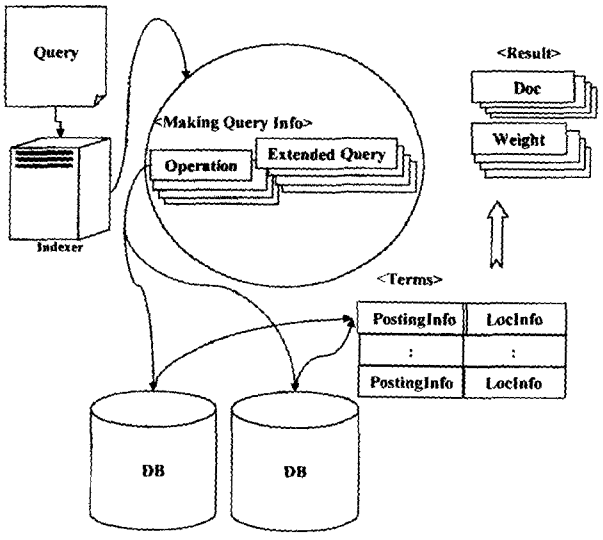
2. 사용자 질의어 확장

2.1. 시스템 구성도

사용자가 입력한 어절에 대해서 질의어 확장을 위한 시스템은 [그림1]과 같다.

사용자 질의어가 들어오면 질의어에 대해서 질의 정보에 대한 Query를 생성하기 위해서 Indexer로 질의어를 넘겨준다. Indexer서버에서 색인 Type에 따른 색인어를 생성하게 되며 색인 Type에 따라 복합명사를 확장할 것인지 판단하여 형태소(morphology) 분석 정보를 이용하여 처리를 하게 된다. 색인 처리가 완료되면 질의 처리기에서 Query를 생성하게 된다.

검색 시스템은 질의 처리기에서 제공해주는 Query Info를 이용하여 DatsBase를 검색하여 사용자가 지정한 모델에 맞게 검색를 처리하여 그 검색된 결과를 사용자에게 제공을 하게 된다.



[그림 1] 시스템 흐름도

2.2 확장 질의어 방법

사용자 질의어 확장은 색인 시스템과 밀접한 관계를 지게 된다. 본 논문의 색인기 시스템은 형태소 분석기로 하는 고성능 색인 시스템이다.

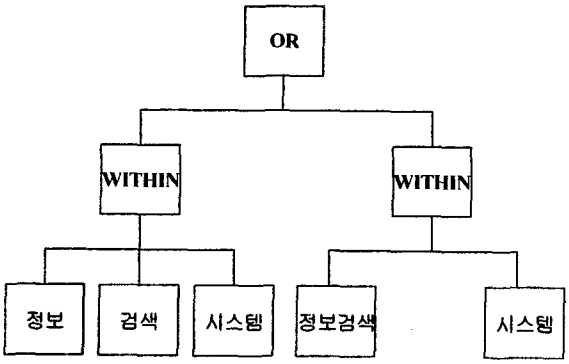
사용자 질의어가 "정보검색 시스템"를 입력할 때 각

확장은 [표 1]과 같다.

[표1] Extended Query

Query: 정보검색 시스템 Jeongbogeomsaek Siseutem Meaning: information_retrieval system
Extended Query : ((정보/W1검색/W1시스템) | (정보검색 /W1 시스템)) Meaning: ((information /W1 retrieval) | (information_retrieval /W1 system))

확장된 질의어는 검색 시스템에서
그림2와 같이 확장된 질의어 노드를 생성하게 된다.



[그림 2] Extended Query Tree

일반적으로 한국어 복합명사에 대해서 띄어쓰기만 고려하면 질의어 확장은 다음과 표2와 같다.

[표2] Normal Expansion

Query: 정보검색 시스템 Jeongbogeomsaek Siseutem Meaning: Information_retrieval system
Normal Query : (정보검색 | 시스템) Meaning: (information_retrieval or system)

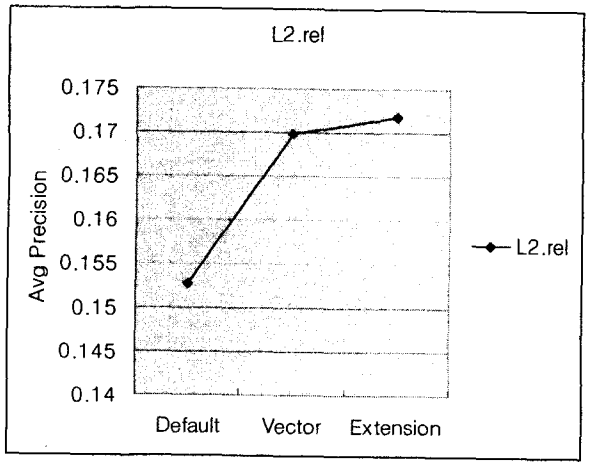
일반적인 한국어 복합명사에 대한 질의어 확장 방법으로도 상위 1000 까지precision를 높일 수 있으나 본 논문에서 제안된 질의어 확장 방법으로 검색을 했을 때가 더 좋은 Precision과 검색 성능을 보여 주었다.

3. 실험

본 논문에서는HANTEC Version 2.0를 이용하여 실험을 하였다. 검색 모델은Vector Model을 이용하였고 질의어는 quer Section를 이용하여50개의 질의어를 이용하여 검색을 하고 검색 대상은 TITLE과 Content Section를 검색하였다.
실험 결과는 [그림 3]과 같다.

[표 3] 실험 Avg Precision

	L2.rel	비고
Default	0.1527	
Vector	0.1699	
Extension	0.1717	



[그림 3] 실험 결과

Default: 질의어에 포함된 복합명사를 하나의 단일명사로 본 것.
예) "정보검색" -> "정보검색"
Vectort: 질의어에 포함된 복합명사를 하나의 단일명사로 본 것.
예) "정보검색" -> "정보 검색 정보검색"
Expand: 질의어에 포함된 복합명사를 분리하여 확장한 것.
예) "정보검색" -> "정보/W검색 정보검색"

실험결과 본 논문에서 제시한 질의어 확장 방법으로 검색을 했을 때 상위 검색 문서가 Precision(0.1717)이 높게 평가가 되었다.

4. 결론

한국어에서는 복합 명사 분석과 색인은 형태가 시스템에 따라 여러 가지로 나타나기 때문에 색인과 검색의 큰 문제로 여겨져 왔다. 본 논문에서는 제시한 질의어 확장 방법을 이용하여 한국어에서 가장 빈번하게 나타나는 복합명사에 대한 정확한 질의어 확장 방법을 이용하여 검색 처리방법을 제시하였다. 그 결과 상위 문서가 Relevant한 문서가 많이 나타나는 결과를 볼 수가 있었다. 또한 인접한 단어를 검색해 줌으로서 사용자가 신뢰하는 검색 결과를 가져왔다.

향후 과제로는 본 논문에서 제시한 질의어 확장에서 예) "한국과학기술정보 연구원"이라는 질의어 대해서 확장 처리를 했을 때 [표 4]과 같이 확장이 된다.
이와 같이 질의어 대해서 너무 많은 확장을 할 때는 검색에 속도에 많은 영향을 미치게 된다. 향후 과제로는 보도 최적화된 질의어 확장에 대한 연구가 필요하다.

[표 4] 한국과학기술정보연구원

Query: 한국과학기술정보연구원 Hangukgwahakgisuljeongboyeonguwon Meaning: Korea Institute of Science and Technology Information
Extended Query : (한국과학 /W1 기술정보 /W1 연구원) (한국과학 /W1 기술 /W1 정보 /W1 연구원) (한국 /W1 과학기술 /W1 정보 /W1 연구원) (한국 /W1 과학 /W1 기술 /W1 정보 /W1 연구원) (한국 /W1 과학 /W1 기술정보 /W1 연구원) (한국과학기술정보연구원) Meaning: : (Korea_Science /W1 Technology_Information /W1 Institute) (Science /W1 Technology /W1 Information /W1 Institute) (Korea /W1 Science_Technology /W1 Information /W1 Institute) (Korea /W1 Science /W1 Technology /W1 Information /W1 Institute) (Korea /W1 Science /W1 Technology_Information /W1 Institute) (Korea_Science_ Technology_Information_ Institute)

참고문헌

[1] Park, Y. C., Choi, K. S., "*A Korean Compound Noun Retrieval Model Using Statistical NounPattern Categorization*," ICCPOL-97, pp. 361-365, 1997.

[2] Cho, M. J., Yun, B.H.,Rim, H.C.,"A Korean Document Retrieval Model considering Compound Nouns and Derived Nouns," Proc. Of the 22nd KISS conference, pp. 499-502, 1997

[3] Park, Y. C.,Choi, K. S., " A Korean Compound Noun Retrieval Model Using Statistical Noun-Pattern Categorization,"

[4] 박대원, "용언 색인을 적용한 한국어 정보 검색시스템의 검색효율 향상" 석사학위논문 2000년 8월