통계적 명사패턴 분류를 이용한 복합명사 검색 모델

박영찬, 최기선
한국과학기술원 전산학과

A Compound Term Retrieval Model Using Statistical Noun-Pattern Categorization

Young C. Park and Key-Sun Choi
Dept. of Computer Science
Koread Advanced Institute of Science and Technology

요약

복합명사는 한국어에서 가장 빈번하게 나타나는 색인어의 한 형태로서, 영어권 중심의 정보검색 모델로는 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. 본 논문에서는 복합명사의 어휘적 정보를 단위명사들의 통계적 행태(statistical behavior)에 기반 하여 자동 획득하고, 이러한 어휘적 정보를 검색에 적용하는 모델을 제시하고자 한다. 본 방법은 색인시의 복합명사 인식의 어려움과 검색시의 형태의 다양성을 극복하는 모델로서 한국어를 포함한 동양권의 언어적 특징을 고려한 모델이다.

1. 서론

정보검색 시스템이란 사용자의 질의를 입력으로 받아 원하는 정보에 해당하는 문헌을 검색하여 주는 시스템을 의미한다. 일반적으로 문헌은 색인과정을 거쳐 색인어들로 축약되어 저장되며, 사용자 질의와의 비교를 통하여 적합도를 계산한다. 이러한 적합도를 바탕으로 검색 시스템은 축적된 문헌을 걸러서 사용자에게 제시한다 [18, 20]. 사용자 질의와 문헌간의 적합도는 사용자 질의에서 사용된 어휘와 문헌을 표현하는 어휘들간의 비교에 근거한다. 따라서 문헌을 나타내는 어휘와 사용자 질의에 사용되는 어휘는 일정한 형식상의 공통점을 갖고 있어야 한다. 이러한 어휘 형식은 크게 통제어휘(controlled vocabulary) 방식과 비통제어휘(uncontrolled vocabulary 또는 free term) 방식의 두 가지가 있다.

통제어휘 방식이란 문헌을 나타내는 어휘와 사용자 질의에 사용하는 어휘를 미리 구축된 어휘로만 사용하도록 강제적인 어휘 제한을 두는 방법이다 [5, 20]. 이 방법은 상당한 양의 어휘를 보유하고 있어야 하며, 신조어 발생 시에는 새로운 어휘를 계속적으로 추가해 주어야 검색시스템과 실제 문헌간의 어휘차이를 극복할 수 있다. 또한 사용자는 시스템에 내장된 어휘로만 질의를 해

야 하므로 사용자의 생각을 충분히 질의로 표현할 수 없는 단점도 있다. 이와 반대로 어휘 통제를 하지 않는 비통제어휘 방식이 있다. 이 방법은 사용자 질의 어휘나 색인어의 어휘를 제한하지 않는다. 따라서 대규모 어휘를 유지할 필요가 없고, 사용자의 질의 구성도 자유로운 장점을 갖는다. 그러나 이 방법 또한 단점을 갖고 있다. 즉 비통제어휘 방법에서는 같은 내용을 갖는 서로 다른 표현이 존재하게 되어 색인과 검색시의 형태상의 불일치가 자주 발생한다 [16].

한국어, 일본어 등의 동양권 언어에서는 기존 어휘들의 결합에 의하여 새로운 단어를 생성하며, 이러한 단어는 결합한 두 단어간의 조합이기는 하나 전혀 다른 의미를 갖는 경우가 많다 [1, 3]. 복합명사는 이와 같이 2개 이상의 단일어로 이루어진 단어이다. 단일어란 단일 의미를 가진 가장 작은 단위의 단어이다. 복합명사는 한국어의 경우 띄어쓰기의 자유로움으로 인하여 2개 이상의 어절로 분리가 되기도 하고, 구(phrase)의 성격을 가지므로 한 문장에 걸쳐서 나타나기도 한다. 이러한 복합명사는 색인어와 사용자 질의간의 형식상의 단위를 다르게 하므로 검색시에 문헌을 제대로 검색할 수 없게 만드는 요인으로 작용해 왔다 [1, 2, 3, 16, 11].

위에 살펴본 바와 같이, 복합명사 문제는 통제어휘와 비통제어휘 접근방법으로 나누어 생각할 수 있다 [1, 3, 16]. 통제어휘 접근방법은 앞에서 언급한 바와 유사하게 색인어로서 가능한 모든 어휘를 등록하여 색인어 사전을 구축하고, 구축된 사전에 기반 하여 복합명사를 색인시에 인식하고, 검색에 사용하는 방법이다 [6, 7, 17]. 그러나 이 방법은 단순한 규칙만으로는 모든 복합명사의 인식이 어렵고, 색인어가 과생성되는 경우도 생긴다. 비통제어휘 접근방법은 모든 복합명사를 단일어로 분리하여, 단일어만으로 문헌을 표현하고 질의를 구성하는 방법이다. 이러한 접근방법은 복합명사를 색인과정에서 따로 인식할 필요가 없고, 어휘의 형태도 단일어로만 구성되어 기존 영어권의 모델을 그대로 수용할 수 있는 장점이 있지만, 복합명사가 가지는 어휘적 특성을 무시하므로, 관련 없는 문헌을 검색하게 되어 시스템의 정확률이 저하되는 단점을 갖는다 [16].

본 논문에서는 복합명사를 구성하는 단일어들의 통계적 행태 분석을 통하여 복합명사 개개의 어휘적 특성을 자동 획득하며, 이러한 어휘적 특성을 이용하여 검색을 수행하는 모델을 제시하고자 한다.

2장에서는 복합명사로 인해 생기는 정보 검색에서의 문제를 알아보고, 이러한 문제를 해결하기 위한 기존의 연구들을 살펴본다. 이어 3장에서 본 논문에서 제안하는 복합명사의 어휘적 특성의 자동 획득 방법과 이를 이용한 검색 모델에 대하여 설명한다. 4장에서는 기존의 방법들과 본 논문에서 제시하는 방법과의 비교 실험을 통하여 모델의 유효성을 입증한다. 이어서 결론을 맺는다.

2. 기존 연구

한국어, 일본어를 포함한 동양권 언어에 있어서 단어 형태의 변화는 복합명사에서 가장 많이 나타난다 [1, 3, 16, 17]. 복합명사란 앞에서도 언급한 듯이 최소 의미를 가진 단일어가 결합하여 새로운 하나의 의미를 생성하는 단어이다. 복합명사는 단일어의 자유로운 합성에 의해 여러 가지 형태로 문헌에 나타나기 때문에 이를 해결하려는 연구가 활발히 진행되어 왔다 [1, 3, 7, 8, 10 ,11, 12, 13, 16, 17, 21]. 본 장에서는 복합명사로 인해 생기는 정보검색에서의 문제를 알아보고, 이 문제에 대한 기존의 해결 방법에 대해 알아본다.

2.1 복합명사로 인한 검색시의 문제점

복합명사 문제는 동양권의 언어에서 그 현상이 두드러지는데 이는 한자에 기반한 단어의 사용으

로 인해 단어의 합성이 자유롭기 때문이다. 영어권의 경우 단일어로부터 복합명사가 합성되더라도 공백으로 모든 단일어가 분리되어 사용되기 때문에 하나의 구로 여겨지게 된다. 따라서 색인과 검색시에 생기는 형태의 불일치가 발생하지 않는다. 동양권의 경우 복합명사 자체가 하나의 단어이면서 구의 구조를 내포하므로 형태의 불일치 문제가 발생한다. 이러한 형태의 불일치 외에도 복합명사는 색인시에는 인식의 문제를 야기시키며, 검색시에는 가중치 부여 등의 문제를 발생시킨다. 다음은 "정보검색"이라는 하나의 복합명사가 문헌 내에서 쓰이는 다양한 표현을 예로 든 것이다.

a. 정보검색
b. 정보의 검색
c. 정보 검색
d. 정보를 효율적으로 검색하는
e. 정보를 검색하는
f. 정보검색 시스템
g. 문헌 정보 검색 시스템
h. 정보를 색인하고 검색하는
i. 정보를 색인 한다. 이를 검색하는 것은...

위의 예에서도 보듯이 각각의 표현들은 모두 "정보검색"이라는 하나의 개념을 지칭한다. "정보검색"이라는 단어는 그 어휘적 특성으로 인해 단일어간의 자유로운 분리가 가능하다. 이러한 다양한 형태를 하나의 복합명사인 "정보검색"이라고 인식하기엔 거의 불가능하다. 또한 모두 인식하였다 하더라도 f, g, i의 예에서 보듯이 다른 복합명사의 일부로 사용되므로, 가중치의 부여 등이 어렵다.

2.2 복합명사를 다루는 기존 연구방법

복합명사는 그것을 이루는 단일어들간의 어휘적 관련성을 검색에 사용함으로써 그 특성을 정보검색에 사용할 수 있다. 하나의 개념은 단일어의 나열이기보다는 구의 구조를 이루는 경우가 많다. 영어권 언어의 경우 각각의 구는 단일어로 확연히 나눌 수 있다. 그러나 동양권 언어의 경우 하나의 구는 하나의 어휘, 즉 복합명사를 이루게 된다. 예를 들어 "인공지능"이라는 하나의 개념은 한국어에서는 하나의 단어로 표현이 되나, 영어는 "Artificial Intelligence"라는 두 개의 단일어로 이루어진다 [9, 4, 14, 15]. 이러한 현상은 색인과정에 있어서 색인어를 인식하고자 할 때 단일어로의 분할을 필요하다. 즉 같은 문제에 대한 해결을 위하여 동양권에서는 색인과정에서부터 색인어의 인식이라는 문제가 필요하다 [16]. 복합명사의 검색에 있어서 영어권의 문제 해결 방법과 동양어권의 문제 해결 방법의 차이는 색인어 형태라고 할 수 있다.

복합명사를 위한 검색 모델은 영어권의 경우 앞서 언급한 바와 같이 인식의 과정을 단순화하여 단일어로만 색인과 검색을 수행하는 모델이 가장 먼저 제시되었다 [18, 19]. 이 모델은 영어가 가지는 특성인 단일어로의 구성을 사용하여, 색인의 단위와 검색의 단위를 단일어로만 한정하는 경우이다. 이러한 경우 불리언 연산자(boolean operator)를 사용하여 복합적인 의미를 표현하여 사용하기도 한다. 이러한 모델은 복합명사의 어휘적 특성을 무시하고 일률적으로 단일어를 사용함으로써 복합적인 의미를 나타내는 기술력(descriptive power)이 저하되어 검색시 정확률(precision)을 저하시키는 요인이 되기도 한다 [16]. 이러한 방법은 복합명사가 색인어의 대부분을 차지하는 동양어권에 직접 사용되기에는 어려움이 있으므로, 사전에 기반하여 단어를 색인하고 검색하는 방법이 가장 먼저 제시되었다 [1]. 즉 사전에는 모든 개념을 표현하는 단어들이 있다고

가정하고 이러한 단어들만을 색인과 검색에 사용하는 방법이다. 그러나 이러한 방법은 사용자로 하여금 사용 가능한 단어를 제한시키며, 문헌 내에서 나타난 단어를 사전에 나타난 단어로 인식하는 색인과정이 복잡하게 되는 단점이 있다.

Ogawa [16]는 이러한 색인의 과정에서 생기는 인식의 문제를 규칙을 통하여 해결하고자 하였다. 즉 각각의 단일어를 이용하여 복합명사를 인식한 후 검색에서 이를 사용하는 방법이다. 먼저 형태소 해석을 거쳐 하나의 긴 문자열(String)에서 단일어를 분리한다. 이렇게 분리된 단일어를 가지고 색인어로 사용 가능한 후보어를 생성하고, 규칙에 기반하여 최종 색인어들을 골라내는 방법을 사용하였다. 이 방법은 단일어의 어휘정보만을 사용하므로 복합명사 개개의 특성을 충분히 고려하지 못하는 단점을 갖는다. 또한, 일률적인 규칙의 적용은 필요 없는 단어를 생성하는 역효과를 낳게 된다.

Croft [6]은 복합명사를 검색하는데 있어서 복합명사를 구조화된 단일어의 집합으로 나타내어 사용하였다. 이러한 단일어들의 구조화를 추론망(inference network)을 사용하여 표현하였다. 이 연구에서는 복합명사를 4개의 구조화된 모델을 사용하여 구조화하였다. 각각의 모델은 다음과 같다.

① 복합명사와 단일어와의 관계를 무시한 모델
② 복합명사를 단일어의 조합으로 나타낸 모델
③ 복합명사를 단일어들간의 관계로 나타낸 모델
④ 복합명사는 단일어들의 구 구조로 나타낸 모델

그러나 위 모델들은 복합명사 개개의 어휘적 현상을 고찰한 것이 아니라, 모든 단어들이 위 모델중 한가지라고 가정하였다. 따라서 복합명사의 개별적인 어휘적 특성이 고려되지 못한 단점을 가지게 되어 기존의 연구모델보다 현저한 성능의 향상을 보이지는 못하였다.

Fujii [8]은 Croft [6]의 연구를 일본어에 적용하여 동양권 언어의 모델에 복합명사의 사용이 검색의 효율을 가져옴을 실증하였다. 즉 동양권의 언어에서는 복합명사를 단일어로 자르는 과정이 필요하게 되므로, 이러한 과정에서 기존의 모델보다 더 좋은 검색 성능을 보였다. 그러나 위의 모델 또한 개개의 복합명사에 따른 어휘적 특성을 고려하지 못하는 단점은 그대로 남아있다.

위의 연구들은 복합명사 문제를 검색에서 해결하고자 한 연구들로, 이외에 복합명사를 인식하는데 주안점을 둔 연구로는 Su *et al.* [21]의 연구가 있다. 이 연구는 단일어간의 상호정보(mutual information)를 이용하여 코퍼스(corpus)로부터 복합명사를 자동 인식하는 방법이다. 기존에 제시되어 온 규칙기반의 복합명사 인식이 아닌 통계에 기반한 방법으로, 복합명사 개개의 어휘적 특성을 코퍼스로 부터 자동 획득할 수 있음을 제시하였다. 그러나 이러한 방법으로 복합명사를 인식하더라도 복합명사가 구성되어 색인에 사용되면 검색에서는 이러한 복합명사를 이루는 단일어들의 관계를 잃어버리므로 검색 시스템에서의 적용은 기존의 사전 기반 검색 모델과 같은 단점을 그대로 갖는다.

3. 통계적 복합명사 검색 모델

본 장에서는 정보검색에서의 복합명사 문제해결을 위해 복합명사를 단일어의 형태적 조합과 단일어간의 어휘적 조합관계로 표현하는 모델을 제시한다. 이 모델은 복합명사의 어휘적 특성을 코퍼스로부터 획득하고, 이를 검색에서 사용한다. 또한, 이 모델은 색인과 검색과정에서 어휘 형태를 일정하게 하기 위해 색인과정에서는 단일어와 그들 간의 위치정보만을 추출한다. 검색에서는 단

일어만을 이용하여 검색하되, 복합명사의 어휘적 특성을 같이 고려하여 단일어 사용으로 인한 복합명사의 어휘정보 손실을 최소화한다.

3.1 복합명사 어휘 지식 획득

복합명사는 단일어로 표현할 수 있다. 그러나 이러한 표현은 단일어들의 단순한 나열만이 아닌 복합명사 개개에 대한 어휘 지식도 같이 표현되어야 원래의 뜻을 유지할 수 있다. 본 연구에서는 복합명사의 인식과 검색의 문제를 해결하기 위해 먼저 복합명사의 유형을 분류하고, 이러한 복합명사의 유형을 문헌으로부터 자동 인식하는 방법을 제시한다. 먼저 복합명사의 유형을 복합명사의 어휘적 특성에 따라 다음과 같이 3가지로 나눈다.

- Type 1 : 항상 일정한 형태로 쓰이는 복합명사.
  복합명사가 하나의 완전한 형태로 굳어져, 하나의 단일어화한 경우이다. 고유명사 등이 이에 속한다. 이러한 단어는 단일어로 분리하여 색인, 검색하면 원래 의도와는 다른 의미가 된다. 예를 들어 '멀티미디어'라는 단어는 단어의 특성상 '멀티'와 '미디어'라는 두 개의 단일어로 구성되어 있다. 그러나 이는 외형상의 분할일 뿐, 두 단어를 따로 분리하면 원래의 의미와는 다른 문헌이 검색되어, 정확률을 저하시킨다.

- Type 2 : 단일 단어로 나누어도 전혀 뜻의 변화가 없는 복합명사.
  이러한 단어는 기술용어와 같은 전문용어 등에 많이 나타나며 단일어화 하지 않은 복합명사로서 자유로운 조합과 분할이 가능하다. '컴퓨터바이러스'라는 단어의 경우 '컴퓨터'와 '바이러스'라는 2개의 단일어로 구성된다. 이 단어는 단일어들간의 독립성이 강해 문헌 내에서 서로 분리되어 나타난 경우라도 이는 하나의 복합명사 의미를 갖는다. 예를 들어 "컴퓨터의 사용에서 조심할 것은 시스템의 안정성이다. 이러한 안정성에 있어 바이러스의 감염은 치명적이다."라는 문장은 '컴퓨터바이러스'라는 복합명사를 내재하고 있다. 따라서 이러한 복합명사를 검색하기 위해서는 단일어들을 모두 분리하여 색인, 검색해야 한다.

- Type 3 : Type 1과 Type 2 이외의 복합명사.
  복합명사가 뚜렷한 분리나 결합의 형태는 없지만 어느 정도의 연관성이 부여된 경우이다. 예를 들어 '이동통신'이라는 단어는 서로 완전히 분리를 하면 원래의 뜻이 사라지게 된다. 그러나 한 문장정도의 일정한 거리 내에서 쓰여진 경우는 의미적 연관성을 갖는 경우이다.

위의 3가지의 단어 유형은 정보검색 과정에서 각각 다른 접근방법을 사용해야 한다. Type 1의 복합명사는 색인과정에서 단일어로 분리해서는 안된다. 이와는 달리 Type 2의 경우는 서로 분리해서 검색하여야 함을 알 수 있다.

이와 같이, 복합명사를 이루는 단일어가 서로 떨어져 있더라도 이를 하나로 고려해야 한다는 것이다. Type 3의 경우는 일정한 경계 내에서 단일어가 같이 발생하면 이를 하나의 복합명사로 인식하여야 하나, 일정한 경계를 벗어나는 경우 이를 하나의 복합명사로 고려해서는 안된다. 즉 복합명사를 위의 유형으로 분류하고, 이를 검색에 사용하여 복합명사 마다 다른 검색방법을 사용함으로써 복합명사에 대한 문제를 해결할 수 있다.

복합명사 유형에 대한 자동 결정은 복합명사를 단일어로 분리한 후 이를 Type 1이라고 먼저 가설을 내리고 이에 대한 타당한 정도를 문헌집합을 관찰하여 얻어낸 후 최종적인 유형 결정을 내린다. 먼저 문헌을 단일어를 기반으로 하여 검색한다. 이러한 검색 결과는 문헌을 유형별로 따로 검색함을 의미한다. 먼저 복합명사 α가 $\beta_1+\beta_2+...+\beta_n$ 형태로 합성된 경우, 단일어 $\beta_1\beta_2...\beta_n$ 을 이용하여 다음과 같은 3개의 문헌집합으로 나누어 검색한다. 다음의 집합 A_α, B_α, C_α를 복합명사 α에 대한 **검색집합**이라고 정의한다.

검색집합 A_α ={ x| x는 $\beta_1, \beta_2, ... \beta_n$ 모두 인접하여 나타난 문헌}
검색집합 B_α ={ x| x는 $\beta_1, \beta_2, ... \beta_n$이 모두 인접하지 않지만 한 문장 내에서는 같이 나타난 문헌}
검색집합 C_α ={ x| x는 $\beta_1, \beta_2, ... \beta_n$이 한 문장 내에서 나타나지 않지만 한 문헌 내에서 같이 나타난 문헌}

위의 검색집합은 서로 포함관계가 없다. 위의 검색집합 A_α, B_α, C_α에서 각각의 문헌집합은 단일어를 통하여 검색된다. 만약 단어 α가 복합명사이고 이를 이루는 단일단어들이 서로 분리될 수 없는 Type 1의 어휘라면 이러한 경우 검색집합 A_α에는 복합명사 그 자체의 의미를 가진 문헌들이 있게 된다. 검색집합 B_α, C_α에는 복합명사를 이루는 단일단어가 서로 떨어져서 존재하므로 단일어들은 같으나 서로 다른 의미의 문헌들이 여기에 있다. 이러한 경우 검색집합 A_α와 B_α, A_α와 C_α사이에는 집합간의 유사도가 상당히 낮게 나타난다. 예를 들어 '멀티미디어'는 '멀티'와 '미디어'라는 두 단일어로 이루어진 복합명사이다. 그러나 이러한 단일어를 서로 분리할 경우 원래의 의미와는 전혀 다른 의미를 갖는다. 따라서 검색집합 $A_{멀티미디어}$에는 멀티미디어에 관련된 문헌들이 속한다. 그러나 검색집합 $B_{멀티미디어}$, $C_{멀티미디어}$에는 '멀티미디어'와는 전혀 상관없는 문헌들이 위치한다. 따라서 한 복합명사의 검색집합들 간의 유사도 계산을 한다면 단어의 유형을 분류할 수가 있다.

검색집합 간의 유사도는 한 검색집합의 대표값과 또 다른 검색집합의 대표값을 비교함으로써 얻는다. 이러한 검색집합의 대표값은 중심점(centroid) 개념으로 결정할 수 있다. 중심점이란 문헌집합 내의 문헌을 개개 단어를 축으로 하는 n-차원 벡터 공간에 나타내고 문헌들이 표현된 벡터의 평균값, 즉 공간상의 구심점을 말한다. 이러한 중심점은 벡터로 표현한다. 검색집합 A_α의 centroid $\hat{C}(A\alpha)$를 구하는 식은 다음과 같다.

$$\hat{C}(A_\alpha)=\frac{\sum_{a\in A_\alpha}\vec{a}}{|A_\alpha|},$$

$|A_\alpha|$ 는 검색집합 A_α 의 집합크기
$\vec{a}$ 는 문서 a에 대한벡터표현

이러한 centroid간의 비교는 두 벡터간의 유사도를 계산하는 방법으로 많이 사용되어 온 코사인 내적 (Cosine Inner Product) 혹은 다이스 계수 (Dice Coefficient) 등을 사용한다 [18].

그림 1은 3개의 복합명사 "멀티미디어", "컴퓨터바이러스", "이동통신"에 대하여 실제로 검색집합을 구하여, 복합명사와 검색집합의 centroid 사이의 위치 상관 관계를 보이고 있다. 그림에서 검색집합 A는 복합명사를 이루는 단일어가 붙어 나타나는 문헌들의 집합이고, 검색집합 B는 문헌 내에 떨어져서 나오는 경우, 검색집합 C는 한 문장 내에서 같이 나타나는 문헌들의 집합을 표시한다. 그림 1에서 Type 1 복합명사라고 생각되는 "멀티미디어" 단어의 경우 (그림 1의 (1)) 복합명사를 이루는 단일어를 분리하여 검색한 경우와 결합하여 검색한 경우의 중심점들 사이에 상당한

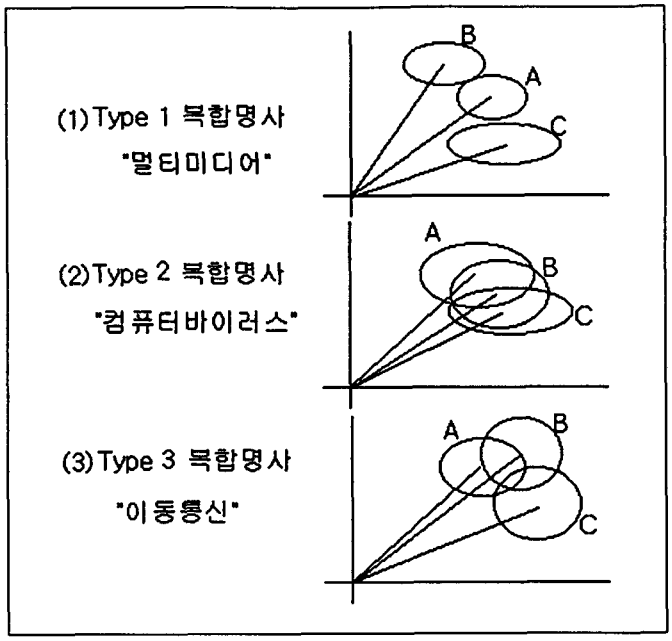

그림 1. 복합명사 Type별 검색집합 centroid 관계

거리가 있음을 알 수 있다. "컴퓨터바이러스"의 경우 (그림 1의 (2)) 복합명사를 이루는 단일어를 분리, 결합해도 같은 뜻을 지니므로 이 경우 서로 분리하여 검색해도 됨을 의미한다. "이동통신"의 경우는 위의 (1)(2) 경우와는 달리 한 문장 내에 발생한 경우는 붙여 쓴 경우의 의미와 유사하고, 한 문장이상 떨어져 나타나면 원래의 복합명사의 의미와 멀어짐을 알 수 있다. 이러한 (1),(2),(3)의 경우와 같이 복합명사를 검색하기 위해서는 단순한 단일어로의 무조건적인 결합 또는 분리를 하는 일률적인 방법이 아닌 복합명사 개개의 특성을 고려해야 함을 알 수 있다.

3.2 검색 모델

검색은 색인어들과 색인어들간의 위치정보, 복합명사의 어휘 구성 정보를 이용하여 다음과 같은 과정을 통해 이루어 진다.

① 복합명사 질의어를 단일어 질의어로 분리
② 검색집합 A, B, C 생성
③ 검색집합들간의 유사도 비교를 통한 어휘지식획득
④ 검색집합과 복합명사 어휘지식을 이용한 복합명사 어휘 가중치 생성
⑤ 확장 불리언 모델을 이용한 문헌 순위화

복합 명사의 어휘 가중치 식은 다음과 같다.

$$W(term|D) = G_{tf} * G_{idf}$$

$$G_{tf} = tf_A + tf_B * typesim_{(A,B)} + tf_B * typesim_{(A,C)}$$

$$G_{idf} = \frac{1}{df_A + df_B * typesim_{(A,B)} + df_C * typesim_{(A,C)}}$$

$typesim(A, B)$: 검색집합 A, B centroid간의 벡터 유사도
$typesim(A, C)$: 검색집합 A, C centroid간의 벡터 유사도

tf_A, tf_B, tf_C : 각 검색집합별 두 단어의 단어공기빈도(word cooccurrence frequency)
df_A, df_B, df_C : 각 검색집합별 문헌내 빈도(document frequency)

위 식은 한 단어가 어떠한 유형의 복합명사인지를 결정하여 검색의 방향을 설정함을 보인다. 질의어의 복합명사가 Type 1, 즉 서로 분리해서는 안되는 복합명사의 경우 검색집합 A만을 문헌 가중치에 포함하며, 서로 완전 분리한 Type 2 복합명사라면 검색 집합 A, B, C를 모두 고려하며, 그 중간인 경우를 0~1의 값으로 곱하여 Type 3 복합명사를 검색 한다.

이러한 방법은 다양한 복합명사의 어휘구성 지식을 실제로 검색에 사용할 수 있다. 또한 신조어의 경우도 문헌분석을 통해 유형결정을 하므로, 신조어를 위한 별도의 처리 모듈이 필요하지 않다.

4. 실험

본 논문에서는 1995년에 개발된 한국어 정보 검색 실험용 데이터 모음 (KTSET2.0) [2]을 사용하여, 이전에 제안된 방법들과 비교실험을 행함으로써 제안하는 모델을 유효성을 입증하고자 한다.

본 실험에서 사용된 KTSET2.0의 특성은 표 1과 같다.

주제	전산, 정보학
문서 구성	논문, 신문기사, 잡지기사
문서 수	4,414
질의 수	50

표 1. KTSET 2.0의 특성

비교 실험을 행한 모델은 다음과 같다

- 복합명사를 고려하지 않은 모델(conjunctive)
  모든 복합명사를 단일어로 분리하여 AND로 표현하는 방법이다. 예를 들어 "인공지능"이라는 복합명사는 '인공 AND 지능' 으로 바꾸어 색인과 검색을 수행한다.

- 사전에 기반한 모델(max exact matching)
  사전의 복합명사 등재의 여부에 따라 가장 길게 일치되는 명사들을 색인과 검색의 어휘로 사용하며, 만약 등재되지 않은 어휘라면 단일어로 잘라서 색인과 검색의 단위로 수행한다.

- 제안하는 모델

단일어와 위치정보로 문헌을 색인하고, 복합명사마다 각각의 어휘정보를 구축하여 검색에 적용한 모델

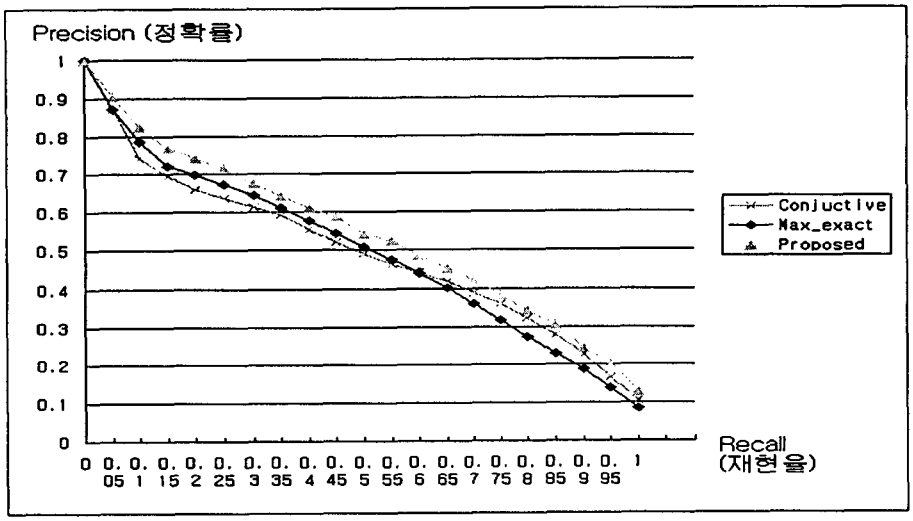

그림 2. 제안하는 모델과 기존 모델과의 검색 결과 (recall-precision)

검색 결과에 대한 평가는 검색 시스템 평가에서 가장 많이 사용되는, 테스트 질의를 통하여 검색된 문헌과 실험용 데이터 모음에 지정된 관련문헌을 정확률(precision)과 재현율(recall)을 비교함으로써 행하였다. 그림 2는 세개의 모델에 대한 검색 결과를 나타낸다.

위의 결과에서 복합명사 정보를 고려하지 않는 모델 (conjunctive)의 경우, 재현율 0.5이상에서 정확도가 떨어지는 현상을 보이고 있다. 그러나 전체적인 재현율은 사전을 사용한 경우보다 높음을 알 수. 반면에 사전을 사용한 모델의 경우 재현율은 낮으나 정확률은 높음을 알 수 있다. 제안하는 모델의 경우 복합명사의 각각의 어휘를 모두 구분하여 독립적으로 지식을 사용하므로, 재현율의 증가와 정확률의 증가를 얻을 수 있다. 따라서 복합명사를 많이 포함하는 한국어의 검색에 적합한 검색모델임을 알 수 있다. 표2는 각각 모델의 실험 결과에 대한 전체 정확률과 재현율 0.5 이상, 재현율 0.5 이하 구간에서의 정확률을 나타낸다.

실험 모델	전체 정확률 평균	재현율 0.5 이상에서의 정확률 평균	재현율 0.5 이하에서의 정확률 평균
conjunctive	0.502	0.670	0.317
max exact	0.503	0.693	0.293
제안하는 모델	0.542	0.723	0.341
제안하는 모델의 성능 향상 (%) (conjunctive와 max exact 중 더 좋은 것과 비교)	7.75%	4.3%	7.57%

표 2. 재현율 구간에서의 각 모델의 정확률 평균

5. 결론

본 연구에서는 한국어 정보처리에서의 가장 중요한 문제 중의 하나인 복합명사를 검색에 사용하는 모델을 개발하였다. 기존의 영어권을 중심으로 한 연구에서는 복합명사라기보다는 구구조(phrase structure)를 검색 단위로서 다루었으나, 실제의 응용에서는 그다지 성능의 향상을 보이지 못하였다. 그러나 동양어권의 검색의 경우 복합명사라는 것은 일반적인 명사로 여겨질 만큼 독립된 어휘로 사용된다. 이러한 복합명사를 다루는데 있어, 색인과 검색에서 이러한 문제를 유연하게 처리하는 모델을 개발하였다.

본 모델은 기존의 연구 방법들보다 정확률 평균 7.75%의 향상을 보인다. 또한 기존의 모델에서 재현율이 높은 모델(conjunctive)과 정확률이 높은 모델(max exact)들 보다 모두 높은 정확률을 보인다. 따라서 기존 두 모델의 장점만을 모두 갖춘 모델이라 할 수 있다.

이 모델은 복합명사 각각의 어휘지식을 자동으로 획득하므로 정확률을 높이기 위해 새로운 어휘를 계속적으로 추가 등재해야 하는 비용을 최소화할 수 있다. 따라서 신조어 등이 나타난다 하더라도 검색 시스템을 수정할 필요가 없다. 또한 재현율을 높이기 위해 정확률을 감소시키지 않는 장점도 갖는다.

참고문헌

[1] 최기선, 한글 문서를 위한 자동 색인어 검출 시스템 개발, 보고서, 한국데이타통신, 1991

[2] 최기선, 지능형 정보 검색 환경, 보고서, 한국통신, 1995

[3] 한성현, 구문해석을 이용한 색인어 자동 추출 시스템의 설계 및 구현, 한국과학기술원 석사학위 논문, 1992

[4] Martin Dillon and Ann S. Gray, FASIT: A Fully Automatic Syntactically based Indexing System, JASIS, 34, 99-108, 1983

[5] W. B. Frakes, Information Retrieval - Data Structures and Algorithms, edited by B. Yates, Prentice-Hall, NJ, 1992

[6] W. Bruce Croft, H. R. Turtle, D. Lewis, The Use of Phrases and Structured Queries in Information Retrieval System, ACM SIGIR-91, pp32-45, 1991

[7] J. Fagan, Experiments in Automatic Phrase Indexing for Document Retrieval, A Comparison of Syntactic and Non-Syntactic Methods, Ph.D. dissertation, Cornell University, 1987

[8] Hideo Fujii and W. Bruce Croft, A Comparison of Indexing Techniques for Japanese Text Retrieval, ACM SIGIR-93, pp237-246, 1993

[9] L. Gay, W. B Croft, Interpreting Nomial Compounds for Information Retrieval, *Information Processing and Management*, 26:10, pp21-38, 1990

[10] Young S. Han, Key-Sun Choi, Syntactic Analysis based Automatic Indexing for Korean Texts, TKE 91, Shanghai, China, 1991

[11] A. Hatakeyama, H. Kawaguchi, H Fujigawa, Processing of synonym and Variant for Free Term Search, IPSJ 39th Zenkoku Taikai, p.1077, 1989

[12] K. Sparc Jones, J. Tait, Automatic Search Term Variant Generation, *Journal of Documentation*, 40:50-66,1984

[13] T. Kamio, Automated indexing for maitaining of a News Paper Article Database. *Journal of Information Processing and Management*, Vol 32, No. 4, pp283-293, 1989

[14] R. Krovetz, W. B. Croft, Lexical Ambiguity and Information Retrieval, *ACM Transactions on Information Systems*, 1991

[15] D. Lewis, W. B. Croft, Term Clustering of Syntactic Phrases, *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, 385-404, 1990

[16] Y. Ogawa, A. Bessho and M. Hirose, Simple String as Compound Keywords: An Indexing and Ranking Method for Japanese Texts, *In Prof. of ACM-SIGIR*, Pittsburg, PA, USA, pp. 227-236, 1993

[17] Hyukro Park, Gene-Sung Chung, Key-Sun Choi, Syntactic Information Based Automatc Indexing for Korean Text, NLPRS 91 , Singapore, 1991

[18] Gerard Salton, Automatic Text Processing, Addison-Wesley Publishing Company, 1988

[19] erard Salton and Maria Smith, On the Application of Syntactic Methodologies in Automatic Text Analysis, ACM, 1989, pp 137-150

[20] Gerard Salton, M. McGill, *Introduction to Modern Information Retrieval,* McGraw-Hill, New york, 1983

[21] K. Su, J. Chang, A Corpus-ased Approach to Automatic Compound Extraction, In *Proc. of ACL*, 242-247, 1994