한국어 정보검색에서 위치관계에 기반한 통계적 구 색인

홍금원ᵘ 김상범 이상주 임해창
고 려 대 학 교 컴 퓨 터 학 과
{gwhong, sbkim, zoo, rim}@nlp.korea.ac.kr

Statistical Phrase Indexing Based on Positional Relation for Korean Information Retrieval

Gum-Won Hong^{U} Sang-Bum Kim Sang-Zoo Lee Hae-Chang Rim
Dept. of Computer Science, Korea University

요 약

최근 웹 문서의 규모가 커짐에 따라 높은 정확도를 필요로하는 정보검색시스템이 요구되고 있다. 구 색인은 정확도를 향상시킬 수 있는 방법으로 전통적으로 많이 사용되어 왔으며, 정보검색에서 사용하는 구는 크게 통계적인 구와 구문적인 구로 나눌 수 있다. 한국에서는 주로 복합명사를 처리하거나, 구문적인 구를 이용한 방법들만이 사용되어 왔고, 통계적인 구를 이용한 검색은 연구되지 않았다. 질의에 존재하는 구의 위치관계와 문서에 존재하는 구의 위치관계가 서로 동일하다면 그 문서는 그 질의와 더욱 유사할 것이라 판단하고, 본 논문에서는 통계적인 구에서 구 구성요소간의 위치관계를 고려한 정보검색 시스템을 제안한다. 명사구 이외의 유용한 구를 생성하기 위하여 내용어를 색인했으며 색인어간의 거리와 순서를 고려하여 가중치를 부여하였다. 명사구와 내용어에 기반한 구를 사용한 각각의 실험에서 거리에 따른 가중치를 부여하는 방법이 거리를 무시한 방법에 비해서 효과적이었고, 구 구성요소간의 위치관계를 고려하는 것이 성능향상의 주요한 요인임을 알 수 있었다. 또한 명사위주의 질의에서는 내용어보다는 명사만을 색인하는 것이 효과적임을 알 수 있었다.

1. 서론

정보검색 시스템의 목표는 사용자가 입력한 질의와 가장 유사한 문서를 빠르게 찾는 것이다. 즉, 정보검색 시스템의 성능은 검색속도와 검색정확도로 구분할 수 있다. 최근 들어 인터넷의 확산과 컴퓨터 시스템의 급속한 발전과 더불어 대용량의 웹 문서를 수 초 내로 쉽게 찾을 수 있지만, 사용자가 진정으로 원하는 문서를 얻기 위해서는 검색 결과로 제공된 문서를 또다시 읽어야하는 번거로움이 있었다. 즉, 많은 양의 불필요한 문서보다는 하나의 꼭 필요한 문서가 더 낫다는 관점에서 보면, 검색 시스템의 정확도는 속도보다도 더 중요하게 다루어질 필요가 있다.

정보검색 시스템의 정확도를 향상시키기 위한 최근의 방법들 중의 하나로 구 색인이 있다. 과거 단일어 색인의 관점에서 보면 질의와 문서를 키워드들을 담고 있는 '주머니'로 표현을 하였다. 그리고 각각의 주머니가 서로 공통되는 키워드를 담고 있다면, 질의와 문서는 어느 정도 유사하다고 보고 있다. 하지만 이러한 단순한 표현방법은 키워드간의 매칭빈도에 의존하고 있기 때문에 문장을 구성하는 단어간의 관계성을 고려하지 않아서 검색의 정확성에 문제점을 지니고 있다. 구는 단어들이 인접해서 어떠한 관계성을 가지고 있어서, 단일어가 표현하는 의미보다 좀더 구체적인 의미를 전달할 수 있다. 예를 들어, '컴퓨터 과학'이나, '컴퓨터 프로그램'은 '컴퓨터'라는 단일어 보다 더 구체적인 의미를 지니고 있다. 이러한 구를 문서간의 내용을 구별해 주는 구별자로서 사용하고자하는 의도는, 단일어만을 사용할 때 발생되는 모호성을 해결함으로써 좀더 정확한 검색성능을 얻고자 함이다.

만약 질의 내 구를 구성하는 요소와, 문서 내 구를 구성하는 요소간의 거리와 순서를 비교하여 고려한다면 좀더 의미 있는 구를 식별해 낼 수 있을 것이고, 이를 이용해서 검색 성능을 향상시킬 수 있을 것이다. 예를 들어 다음 세가지 구는 '정보'와 '검색'사이의 위치관계가 서로 다르다.

(1) '**정보**를 **검색**하는 시스템'
(2) '**정보**를 빠르고 정확하게 **검색**하는'
(3) '**검색**된 **정보**를 사용자에게 제공하는'

(1),(2),(3)에는 각각 '정보'와 '검색'으로 이루어진 통계적인 구가 존재하지만, 색인어 사이의 거리와 순서가 서로 다르다. 그리고 '정보검색'이란 질의로 검색을 할 경우 거리와 순서를 고려한다면 (2),(3)보다 (1)에 더 큰 가중치를 부여할 수 있다.

한국어 정보검색에서 구를 사용했던 기존의 연구들은 주로 복합명사의 처리와 관련된 시도들이었다. 한국어의 특성상 복합명사는 띄어쓰기가 자유롭기 때문에 복

합명사의 분해나 다른 다양한 형태의 명사구들을 정규화하는 연구들이 대부분이었다.

한국어 문서에서 문장의 주된 내용을 명사가 나타내고 있고 복합명사를 처리하는 경우 실제로 성능이 향상되는 예가 많았다[8][11]. 하지만 명사만을 색인해서는 검색이 힘든 경우도 있다. 예를 들어 "토끼는 무엇을 먹고사는가?"라는 질의에 다음 두 개의 문장, (문서 1)"토끼는 당근을 먹고산다.", (문서 2)"토끼야, 토끼야, 산 속에 토끼야!"가 있을 때 색인어를 어떻게 추출하느냐에 따라 검색 순위가 바뀔 수 있다.

[표 1] 색인어 추출방법의 예

	명사추출	실질형태소 추출
문서 1	토끼, 당근	토끼, 당근, 먹, 살
문서 2	토끼, 산	토끼, 산, 속

만일 명사만을 색인 했을 경우에는 다음과 같은 경우가 발생할 수 있다. (문서 2)에 '토끼'라는 명사가 가장 많이 출현했기 때문에 질의와 문서간 유사도는 (문서 2)가 (문서 1)보다 높아질 것이다. 그러나, 명사를 포함한 내용어 전부를 색인할 경우는 이와 반대의 경우가 될 것이다.

본 논문에서는 색인어가 될 수 있는 대상을 명사만으로 국한시키지 않고, 모든 종류의 실질형태소를 색인어로 다룬다. 그리고 명사만을 색인어로 사용했을 때와 비교해서 실험해 보았다. 또한 단일어의 문헌빈도를 적절히 고려하여 의미 있는 구를 식별해 내고, 생성된 구에 구 구성요소간의 위치관계에 따라 적절한 가중치 할당을 하여 검색의 효율을 높이도록 하였다.

2. 관련연구

구 색인과 관련된 기존의 연구들을 살펴보기 이전에 우선 구에 관한 간략한 정의를 내려볼 필요가 있다. 기본적으로 구는 단어들의 집합이지만 학자들에 따라서 의견이 조금씩 다르다. 구가 단일어보다는 좀더 의미 있는 개념으로 문서를 표현한다는 점에서, 영어권의 경우에는 실제로 구를 검색에 활용하는 것이 상당히 일반적이다. 최근 TREC에 참여한 많은 시스템들이 적어도 하나 이상의 구 추출방법을 채택하고 있다[3].

이들이 사용한 구는 크게 두 가지이다. 하나는 통계적인 구(Statistical Phrases)인데, 이것은 문서 내에서 빈번히 인접하여 나타나는 임의의 단어들의 집합을 의미한다. 다른 하나는 구문적인 구(Syntactic phrases)로, 이것은 문장 내 특정한 구문구조를 이루고 있거나, 어떤 구문적인 관계성을 표현하는 단어들의 집합을 의미한다[6][7]. 구문적인 구가 복잡한 자연어처리기법을 요구하는 것과는 달리, 통계적인 구는 단어빈도나 공기정보만으로도 쉽게 추출이 가능하다.

두 가지의 방법 중 어느 것이 더 유용한지는 아직까지 명확하지 않다. Mitra[7]는 두 가지의 구를 각각 이용한 실험에서, 단일어를 사용하지 않은 비교실험에서는 구문적인 구가 성능이 높게 나타났지만, 단일어와 함께 사용한 실험에서는 그러한 현상이 사라졌었다. 또한 구가 높은 순위의 문서가 아닌 낮은 순위의 문서들에서 더 효과적으로 작용함을 보였다.

Fagan[5]은 통계적인 구를 이용한 실험에서, 단일어의 위치정보를 기반으로 5개의 문서집합에서 평균 10%정도의 성능을 향상시켰으나, 구를 구성하는 구성요소간 순서와 기타 관계성들을 고려하지 않아서 부적절한 구가 생성되었고 따라서 정확한 구문 분석 방법이 필요하다고 지적했다.

장명길[10]은 형태소 태그열 패턴으로 키팩트를 추출하고 이를 활용하여 명사구 내 연관된 단어의 쌍을 효과적으로 추출해내었다. 하지만 직관적이지 않은 가중치 계산과 부정확한 키팩트 과생성의 문제가 남아있었다.

원형석[11]은 복합명사 분할 및 명사구 합성 시 제한된 자연어처리기법을 이용하여 구문적인 구를 합성해 내고, 이를 조합한 16가지 실험에서 성능향상을 보였다. 하지만 대상 문서집합이 너무 작았고, 구를 구성하는 구성요소간의 거리를 가중치 계산에 고려하지 않았다.

3. 시스템 구성

정보검색시스템에서의 색인은 문서에 나타난 용어들의 빈도를 조사하고, 검색모델의 가중치 부여방법에 따라 각 용어들에 가중치가 부여되어, 찾기 쉬운 형태로 조직되고 저장되는 과정이다. 본 시스템에서는 색인어를 저장하는 자료구조로 가변차수 B트리를 이용한 역파일 구조를 사용하였다[13]. 하지만 색인어 위치관계를 활용한 가중치 부여과정을 색인시점이 아닌 검색시점으로 미루는 방법을 채택하여서, 실제로 색인은 용어빈도와 용어가 출현한 위치정보만이 저장되는 과정으로 제한하였다. 따라서 가중치 부여 방법에 따라 문서전체를 다시 색인해야하는 기존의 정보검색시스템들과는 달리, 여러 가지 가중치 부여방법을 검색과정에서 다각도로 수행할 수 있도록 하였다.

본 시스템은 크게 색인어 추출모듈, 색인어 저장모듈, 검색모듈의 세 부분으로 나눌 수 있다. 이 중에서 색인어 추출모듈에서는 형태소 분석기를 통해 형태소를 분석하고, 품사 부착시스템에 의해서 품사를 부착한다[14]. 그리고 품사가 부착된 형태소들 중에서 명사를 포함한 실질형태소를 추출하고 또, 그것이 존재하는 문장내의 위치정보를 추출한다. 영어권의 경우, 색인어 추출시스템은 일반적으로 어근추출(stemming)을 통하여 용어들을 정규화 시켜주고, 고빈도의 단어들을 불용어(stopwords)로 간주하여 색인어 추출에서 제거시켜주지만, 한국어를 다루는 대부분의 정보검색시스템들은 주로 명사추출기나 품사부착기에 의존해서 명사만을 색인어로 추출한다. 하지만 명사만을 색인어로 사용했을 경우에는 복합명사 이외의 유용한 구들을 색인할 수가 없기 때문에, 문장을 구성하는 단어들과의 관계성을 제대로 표현하기 힘들 것으로 생각된다.

본 시스템에서는 문장 내 의미가 없는 형식 형태소만

을 일종의 불용어로 간주하고 제외시켰을 뿐, 실질형태소 전체를 색인어로 간주하여 색인하였다. 그리고 실질형태소 중에서 명사만을 색인 했을 경우와 비교하여 실험해 보았다. 그리고 통계적인 구만을 추출하기 위해서 실질형태소 각각의 품사정보는 색인하지 않았다.

3.1 위치정보 색인 방안

통계적인 구는 기능어를 제외하고서 자주 인접해서 나타나는 임의의 단어집합이다. 만일 구 구성요소의 개수를 2개로 제한한다면 색인하는 시점에서 문장 내의 형식형태소를 제거한 뒤 실질형태소 바이그램을 추출하고 계산된 가중치 자체를 저장하여 색인하는 방법이 있을 수 있다. 하지만 본 연구에서는 구를 구성하는 구성요소간의 위치정보 즉, 거리와 순서를 어떻게 활용하느냐에 따른 성능변화를 측정하고자 했다. 그러기 위해서 색인하는 단계에서, 가중치 자체를 저장하지 않고 단지, 색인어의 빈도정보와 색인어가 출현한 문서 내 문장의 위치, 문장 내 색인어의 위치를 저장하였다. 그리하여 검색 단계에서 구 구성요소간의 거리와 순서를 어떻게 고려하느냐에 따라서 다음과 같은 4가지 형태의 구를 사용하였다.

[표 2] 구 구성요소의 위치관계에 따른 고려 사항

D1	거리와 순서를 무시한 구
D2	거리를 고려하지만 순서를 무시한 구
D3	거리를 무시하지만 순서를 고려한 구
D4	거리와 순서를 모두 고려한 구

구 구성요소간의 거리는 같은 문장 내에서 일정한 윈도우(window)[1] 이내에 존재하는 두 단일어로 이루어진 구 만을 대상으로 측정하였고, 윈도우를 넘어서 존재하는 구 구성요소간의 거리는 일정한 값으로 고정시켰다. 또한 윈도우 이내에서 존재하더라도 서로 다른 어절간에 존재하는 구 구성요소간의 거리는 형태소가 출현한 위치만으로 계산된 거리에 어절위치를 고려한 거리를 더하여 계산하였다. 예를 들어 '**남북 정상회담**'에서 생성될 수 있는 모든 구들과 구 구성요소의 거리는 다음과 같다.

남북_정상(2), 정상_남북(-2), 남북_회담(3), 회담_남북(-3), 정상_회담(1), 회담_정상(-1)

1) 본 논문에서 말하는 윈도우는 구 구성요소를 추출하는 영역이고, 문장 내 존재하는 일정수의 색인어로 정의된다.

위에서 괄호 안의 숫자는 구 구성요소간의 거리를 의미하며 거리가 음수가 되는 것은 두 번째 구성요소가 첫 번째 구성요소보다 실제로 앞서 나타난다는 의미이다. '남북_정상'의 경우에는 '남북'과 '정상'이 서로 다른 어절에서 나타났기 때문에 형태소간 거리 1에 어절간 거리 1을 더해서 2가 된다. 그리고 '정상_회담'의 경우에는 같은 어절 내에 존재하기 때문에 형태소간의 거리만을 고려하여 거리가 1이 된다. '정상_남북'과 같은 구는 거리가 -2로 음수가 된다. 만약 질의에는 '남북_정상'과 같이 순서가 똑바로 되어있고, 어떤 문서에 구성요소간의 순서가 거꾸로 나타난 구가 존재할 때, 이 구의 가중치에 일정한 벌점(penalty)을 줄 수 있게 하였다.

이들을 종합하여 본 논문에서는 문서 내 특정 형태소 a가 b와 공기하는 관계함수 $R(a,b)$을 다음과 같은 수식으로 정의하였다.

$$R(a,b)=\begin{cases}\min(Dist(a,b),Inf): If\ Dist(a,b)>0\\ \max(Dist(a,b),-Inf): If\ Dist(a,b)<0\end{cases}$$

$Dist(a,b)=offset(b)-offset(a)$ a,b가 같은 문장에 나타날 때

$offset(x)$: i번째 문장 내에서 색인어 x가 나타나는 위치

Inf : 거리가 적용되는 윈도우의 최대 크기

만일 구 구성요소 a,b 가 같은 문장에서 어느 정도의 거리를 두고 특정 윈도우 이내에서 나타난다면, $R(a,b)$는 $Dist(a,b)$만큼의 거리를 가지고 있을 것이다. 그렇지 않고 서로 다른 문장에서 나타나거나, 혹은 어느 정도 거리 이상에서 나타난다면, $R(a,b)$는 단순히 더 이상 거리를 고려하지 않고 Inf값(본 논문에서는 5)을 가지게 된다. 질의와 문서에서 어떤 단일어 a,b가 생성하는 관계함수를 각각 $R_q(a,b)$, $R_d(a,b)$라 놓자. 질의에서는 $R_q(a,b)$ = 1 또는 $R_q(a,b)$ = 2인 구만을 추출한다. 이때 $R_d(a,b)$의 절대값이 크면 클수록 $R_q(a,b)$의 값과 더욱 많은 차이가 생기게 된다. 본 논문에서는 문서 내에 존재하는 구의 가중치를 $R_d(a,b)$의 절대값에 반비례하도록 설정하였다.

3.2 구 생성방법

구가 색인어로서 가치가 있는 이유는 구가 가지고 있는 특정성이다. 구는 일반적으로 단일어보다 식별력(discriminating power)이 커서 문서를 보다 구체적으로 표현할 수 있기 때문에 검색 정확도를 향상시키는데 도움이 된다. 하지만 통계적인 구를 생성하는 과정에서 식별력이 떨어지는 구가 상당히 많이 발생한다. 예를 들어 문헌빈도의 관점에서, '~에 관한 내용', '~에 대한 문서'와 같은 구에서는 단일어 '관하', '내용', '대한', '문서'등은 전체 질의 중 약 80%정도의 질의 문장에서 항상 나타나며, 문헌빈도가 지나치게 크기 때문에 사실상 이들이 구로 사용되더라도 구 자체의 식별력이 떨어지게 된다.

일반적으로 식별력이 나쁜 색인어는 높은 문헌빈도를

가지고, 식별력이 좋은 색인어는 적당한 문헌빈도를 가지며, 식별력이 아주 없는 색인어는 극히 낮은 문헌빈도를 갖는 경향을 갖는다[1][5].

본 논문에서는 구 구성요소가 될 수 있는 조건으로 *df_threshold(문헌빈도 임계값)*라는 파라미터를 사용하였고, 각각의 구 구성요소는 다음과 같은 조건을 만족해야 한다.

$$df \le CollectionSize \times df\_threshold \qquad (1)$$

위에서 *CollectionSize*는 전체 문서집합에 존재하는 모든 색인어의 개수이고, 문헌빈도 임계값은 각각의 구 구성요소가 가지는 문헌빈도의 최대 비율이다. 만일 문헌빈도 임계값이 0.1이 될 경우 문헌빈도가 10% 이내인 단일어들만을 구 생성에 포함하겠다는 의도이며 이 값이 작을수록 구 생성에 포함시키는 단어의 수와 종류가 줄어들어 검색 시간이 단축되는 반면 값이 너무 작으면 의미 있는 구가 생성될 가능성이 줄어든다. 반면에, 이 값이 크면 구 생성에 포함시키는 단어의 수와 종류가 너무 늘어나서 검색 시간이 길어진다. 따라서 적당한 수준의 임계값을 실험을 통해 선택해서 사용하는 것이 좋다.

단일어의 경우에는 (2)에서와 같이 log를 사용한 역문헌빈도 파라미터에 의해서 가중치의 조절이 이루어지므로 문헌빈도 임계값을 사용할 필요가 없다.

3.3 구 가중치 부여방법

위와 같은 구가 생성이 되면 구의 중요도에 따라 그에 맞는 적절한 가중치를 부여해야 한다. 본 시스템에서 질의와 문서간의 유사도는 그 질의와 문서에 공통적으로 출현한 단일어들의 가중치와 구들의 가중치가 더해져서 계산된다. 따라서 단일어에 의해서 검색된 문서들이 어느 정도의 정확도를 가지고 있기 때문에, 구 가중치 부여방법에 따라 시스템의 성능이 좌우된다고도 할 수 있다.

우선 본 논문에 사용된 단일어의 가중치는 확률모델에 기반한 2-Poisson 모델의 가중치 부여방법을 확장한 TREC-8의 Okapi 시스템이 사용하였던 BM25방법에 의하여 가중치를 할당하였다[2].

$$sw_{ij} = \frac{tf_i}{k_1 \cdot \left((1-b) + b \cdot \frac{dl_j}{avdl}\right) + tf_i} \cdot \log \frac{N - df_i + 0.5}{df_i + 0.5}$$
$$where,\ k1 = 1.5, b = 0.5 \qquad (2)$$

식 (2)에서 *k1*값과 *b*값은 실험을 통하여 학습데이터로부터 얻어내야 할 파라미터이지만, 본 논문의 목적을 단일어 가중치 부여방법에 따른 성능향상에 두고 있지 않기 때문에 이 값들을 고정하고 사용하였다. 그리고 다음과 같은 두 가지의 구 가중치 부여방법을 사용하였다.

$$P1: pw_j(a,b) = \sum_{\forall a,b \in D_j} \frac{sw_j(a) + sw_j(b)}{2} \times \frac{1}{\sqrt{diff(a,b)}} \times \frac{1}{Penalty\_value}$$
$$P2: pw_j(a,b) = \sum_{\forall a,b \in D_j} C \times \frac{1}{\sqrt{diff(a,b)}} \times \frac{1}{Penalty\_value}$$
$$diff(a,b) = |R(a,b)| \qquad (3)$$

식 (3)에서, $pw_j(a,b)$는 *j*번째 문서에 공기하는 모든 단일어 a, b를 이용하여 생성되는 구의 가중치이고, $sw_j(a)$와 $sw_j(b)$는 각각 *j*번째 문서에 존재하는 단일어 a와 b의 가중치이다.

구의 가중치는 단일어의 가중치에 의존적이라고 가정하고 두 개의 단일어 a, b가 가지는 가중치의 평균을 사용한 *P1*과, 단일어의 가중치와는 상관없다고 가정하고 일정한 상수 *C*(본 논문에서는 5)를 사용한 *P2*를 사용한다. *diff(a,b)*는 문서 내에서 존재하는 구 구성요소간의 거리에 절대값을 취한 값으로, 질의에 존재하는 구 구성요소들이 서로 인접해서 나타났을 때, 문서 내에 존재하는 구들의 구성요소간 거리와의 차이를 반영한 값이다. *Penalty_value*(본 논문에서는 1.5)는 구 구성요소간의 순서를 고려하여 질의에서의 순서와 다를 경우 구의 가중치에 가해지는 벌점이다. *diff(a,b)*값을 직접 사용하지 않고 제곱근을 취한 값을 사용한 이유는 구 구성요소간의 거리가 비교적 적당히 떨어져있는 경우와, 다분히 많이 떨어져 있는 경우와의 차이를 줄이기 위함이다.

*P2*는 구 가중치 부여과정에 단일어의 가중치를 사용하는 것이 과연 효과가 있는지를 검증하기 위해서 사용한 방법이다. 단일어의 가중치를 고려하지 않고, 상수 값 C를 사용하는 이유는, 구가 생성되는 모든 시점에서 동일한 가중치를 부여하여서 의미 있는 구가 출현하는 횟수 자체가 검색성능에 어떠한 영향을 미치는지 알아보기 위한 의도이다.

식 (3)의 *diff(a,b)*와 *Penalty_value*를 다음과 같이 조합하면, [표 2]에서 제안한 4가지의 구를 사용할 수 있다.

$$D1: diff(a,b) = 1, Penalty\_value = 1$$
$$D2: diff(a,b) = |R(a,b)|, Penalty\_value = 1$$
$$D3: diff(a,b) = 1, Penalty\_value > 1$$
$$D4: diff(a,b) = |R(a,b)|, Penalty\_value > 1 \qquad (4)$$

식 (4)에서 *D4*는 거리와 순서를 모두 고려한 방법으로, 서로 인접해서 나타나는 구 만을 사용하기보다는 문서 내에 존재하는 모든 단어들의 쌍을 구로 고려하기 위한 것이다. 또한 구 구성요소간의 거리가 멀 경우와 구성요소간의 순서가 질의에서와 다를 경우에 구의 가중치를 줄이고자하는 의도이다.

구의 가중치가 할당이 되면 단일어의 가중치와 합산하여 다음과 같은 질의와 문서간의 유사도를 계산한다.

$$sim(d_j,q)=\sum_{t_i\in q} sw_{ij}\cdot qtf_i+\sum_{p_k\in q} pw_{kj}\cdot qpf_k \qquad (5)$$

식 (5)에서 sw_{ij}는 j번째 문서에 나타나는 i번째 단일어의 가중치이고, pw_{kj}는 j번째 문서에 나타나는 k번째 구의 가중치이다. qtf_i는 질의에 존재하는 i번째 단일어들의 빈도이고, qpf_k는 질의에 존재하는 k번째 구들의 빈도이다.

3.4 구 가중치 정규화 방법

단일어의 가중치는 Okapi 시스템이 BM25방법에서 사용했던 바와 같이, 문서길이에 의한 정규화가 이루어진다[2]. 일반적으로 문서의 길이가 길면 길수록 질의에 존재하는 동일한 형태의 구들이 나타날 확률이 높아진다. 따라서 구에서도 문서의 길이에 의한 가중치 부여방법을 다음과 같이 사용하였다.

$$pw_j(a,b)_{norm}=\frac{pw_j(a,b)}{(0.25\times avg\_dl+0.75\times dl_j)} \qquad (6)$$

위 방법은 Singhal[4]이 Pivoted Document Length Normalization에서 사용한 방법을 구 가중치 정규화에 적용한 것이다. 여기서 avg_dl은 평균 문서길이이고, dl_j은 j번째 문서의 길이이며, 문서길이는 문서 내에 존재하는 단일어의 개수로 계산된다.

4. 실험 및 평가

본 논문에서 구현한 시스템은 HANTEC 테스트 컬렉션에 적용하여 평가를 하였다. HANTEC 컬렉션은 약 12만 건의 문서집합으로 구성되어 있으며, 30개의 질의 집합으로 실험하였다. 정답문서 집합은 총 정답수가 365개인 L2를 사용하였으며, 질의 당 평균 정답 수는 약 12개이다.

[표 3] 색인어 추출 방법에 따른 문서길이

색인 방법	명사 색인	내용어 색인
총 문서길이	27709160	34935188
평균 문서길이	230.92	291.14

명사색인의 경우는 보통명사, 고유명사 외에 수사, 외국어 문자를 추출하였고, 내용어 색인의 경우는 명사 외에 용언과 관형어, 접사 등의 모든 실질 형태소를 추출하였다. [표 3]을 보면, 내용어를 색인 했을 경우가 명사를 색인 했을 경우에 비해서 1.3배 정도 많은 색인어를 가지고 있으며, 내용어 중의 약 21%정도가 명사를 제외한 기타 실질 형태소로 구성되어 있음을 알 수 있다.

HANTEC 컬렉션의 각 질의는 <title>, <desc>, <narr>, <query>의 네 필드로 구성되어 있다. 이중 <narr>필드는 적합문서를 판별하는 기준을 기술한 것으로 적합성 판단의 판정자가 검색된 문서집합 중에서 적절한 문서를 판별하는 기준을 제공하는 것을 주목적으로 하고 있으나 검색 시스템도 내부질의 생성에 사용할 수 있다. 이는 <title>과 <desc>필드만으로는 질의의 모호성 해소가 안 되는 경우가 많아 적합성 평가 시에 평가자 간의 일관성이 없을 수 있기 때문이다[12]. 본 논문에서는 <narr>필드의 질의만을 가지고 실험을 하였다.

[표 4] 질의어의 품사별 빈도 - <narr>필드

명사	동사	형용사	부사	관형사	보조용언	접미사	계
533	60	10	7	7	8	22	651

<narr>필드는 단문 형식이나 키워드 방식이기보다는 비교적 길고 완전한 형태의 문장질의이기 때문에 품사부착시 발생되는 오류가 다른 필드에 비해서 적고, 다른 필드에 비해서 명사 이외의 내용어를 좀 더 많이 포함하고 있다.

실험결과의 평가방법은 non-interpolated average precision을 사용하였다.

[표 5] 단일어 색인의 경우 평균 정확도

색인 방법	명사 색인	내용어 색인
평균 정확도	0.4499	0.4337

[표 5]에서와 같이 단일어만을 색인한 실험에서는 명사만 색인했을 경우가 내용어를 색인했을 경우에 비해서 더 나은 성능을 보였다. <narr>필드는 다른 필드에 비해 명사 이외의 내용어가 많지만, 그것들로는 질의의 내용을 효과적으로 표현하기가 어려움을 알 수 있었다. [표 4]를 보면, 동사가 명사를 제외한 내용어의 반 정도를 차지하고 있다. 동사의 약 80%정도를 차지하고 있는 단어들이 '있', '관하', '하' 등의 고빈도의 단어들이어서 이들은 실제로 구별력이 적기 때문에 검색효율을 저하시킨다. 따라서 단일어를 색인할 때는 명사만을 색인하는 것이 탐색시간 과 검색 정확도 면에서 효율적인 것으로 나타났다.

하지만 17번 질의 "ATM 망에서의 트래픽 막힘 제어 기법에 관한 연구"의 경우를 보면, '막힘'이라는 어절이 '막히다'라는 동사와 명사형 어말어미로 구성되어 있어서 '트래픽 막힘'이라는 의미 있는 구를 생성해 낼 수 있는 가능성을 보였고, 명사 색인의 경우보다 약 24%정도 높은 정확도를 보였다. 하지만 HANTEC 컬렉션의 질의 집합에는 명사가 전체 내용어중 82%에 달하고 있고, 의미 있는 구를 생성해 낼 수 있는 내용어들의 예가 부족했기 때문에 한국어에서 명사이외의 내용어들을 사용하는 것이 문서의 내용을 표현하는데 유리하다는 사실을 입증하지는 못했다.

[표 6] 위치관계에 따른 구 색인의 평균 정확도

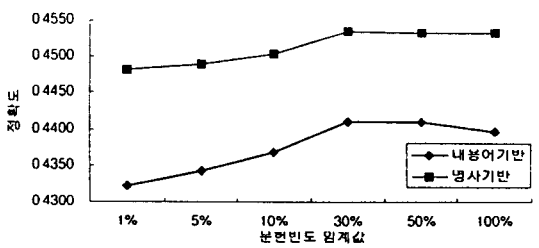
색인방법	명사 색인		내용어 색인	
가중치	단일어평균	상수	단일어평균	상수
baseline	0.4468	0.4230	0.4300	0.4153
거리	**0.4526**	0.4451	0.4395	0.4293
순서	0.4489	0.4330	0.4345	0.4229
거리+순서	0.4494	**0.4495**	**0.4403**	**0.4331**

[표 6]은 거리와 순서에 따른 구 색인의 성능을 분석한 실험결과이며, 사용된 정확도는 non-interpolated average precision이다. 문헌빈도 임계값으로 명사와 내용어 각각 0.15를 사용하였다.

우선 두 단일어의 평균을 구의 가중치에 사용한 실험(*P1*)에서, 거리와 순서를 고려하지 않은 방법은 단일어만을 사용했을 때보다 오히려 평균 정확도가 떨어짐을 알 수 있었다. 그러나 거리만을 고려한 경우(*D1*)에는 성능이 향상되었음을 알 수 있었다. 이것은 명사 색인이나 내용어 색인 모두의 경우에 마찬가지였다. 전반적으로 거리와 순서를 모두 고려한 방법의 결과가 높은 정확도를 보였지만, 거리만을 고려한 방법과 큰 차이를 보이지는 못했다. 이는 구 구성요소간의 순서가 검색 성능에 큰 영향을 주지 못했음을 말해준다.

가중치 부여방법별로 분석을 해 보면, 두 단일어의 평균을 사용한 경우(*P1*)가 구를 구성하는 구성요소의 가중치를 전혀 사용하지 않고 일정한 상수 값(본 실험에서는 5)을 사용한 경우(*P2*)에 비해서 더 나은 성능을 보였다. *P2*의 경우는 8개의 결과 모두가 단일어의 성능에도 미치지 못한 것으로 나타났는데, 이것은 상수 C를 변경하며 실험했을 때도 마찬가지였다. 이것은 생성된 구의 가중치가 구 구성요소의 가중치에 어느 정도 의존적이라는 것을 의미한다. 따라서 구의 가중치는 구 구성요소 각각의 단일어 가중치와, 구 구성요소 사이의 거리와 밀접한 관련이 있음을 알 수 있다.

색인어별 비교의 경우, 단일어와 마찬가지로 구를 사용했을 때도 명사만을 사용했을 때가 내용어를 사용했을 때에 비해서 더 나은 성능을 보였다. 이것도 단일어 실험에서와 마찬가지로 의미 있는 구를 생성해 내기 위해 필요한 명사이외의 내용어들이 부족했고, 이 경우 명사구를 생성하는 것 이외에 색인어 위치간의 관계성들을 추출하는데 어려움이 있는 것으로 보인다.



[그림 1] 문헌빈도 임계값과 평균 정확도

[그림 1]은 식별력이 있는 구를 생성하기 위해서, 문헌빈도 임계값을 조정하여 실험한 결과이다. 가중치 부여방법은 *P1*을 사용하였다. 명사구는 거리를 고려한 방법을, 내용어에 기반한 구는 거리와 순서를 모두 고려한 방법을 사용하였다. 임계값이 0.3 일 때 명사구에서는 0.4536, 내용어에 기반한 구에서는 0.4410으로 가장 좋았는데, 0.3이상 일 때는 더 이상 정확도가 올라가지 않음을 볼 수 있었다. 문헌빈도 임계값이 더 클수록 더 많은 단어들이 구 생성에 참여하므로, 검색 시간도 비례하여 증가하게 된다. 따라서 검색속도와 정확도를 모두 고려하는 문헌빈도 임계값의 선택이 필요함을 알 수 있었다.

5. 결론

본 논문에서는 단일어 문헌빈도와 문서 내 공기정보를 이용하여 통계적인 구를 생성해 내고, 구 구성요소간의 위치관계를 고려한 가중치 할당을 통하여, 영어권에서만 주로 사용되던 통계적인 구를 한국어 정보검색 시스템에 적용하여 효과적으로 사용할 수 있는 방안을 제안하였다. 구 구성요소의 위치관계를 고려한 실험을 통하여 구 구성요소간의 순서는 검색 성능과는 무관하였고, 거리를 적절히 활용하는 것이 필요함을 알 수 있었다. 단일어와 구를 사용한 각각의 실험에서 내용어를 색인했을때보다 명사를 색인했을때가 더 성능이 좋았다. 실험결과를 분석하면서 발견된 문제점은 통계적인 구에서는 다음과 같은 질의에서 의미 있는 구의 일부분만을 생성한다는 점이었다. '신기술의 현황과 전망'에서 '신기술_현황', '현황_전망'의 구만을 생성하기 때문에 '신기술_전망'과 같은 구를 생성하지 못했고, '당뇨병의 유형, 원인, 합병증'과 같은 예에서도 '당뇨병_원인', '당뇨병_합병증'과 같은 구를 생성할 수 없었다. 따라서 위와 같은 관계성을 가지는 질의어에 대해서는 색인어의 인접정보뿐만 아니라 구문적인 관계까지 고려하는 구 생성이 필요함을 알 수 있었다.

이번 연구에서는 구 구성요소간의 위치정보만을 이용한 통계적인 구 생성에 초점을 맞추었지만 앞으로는 구문적인 구에서 구성요소간 관계성을 추출하고, 이것과 위치관계에 의한 가중치할당을 접목시켜 통합적인 구 색인에 관한 연구를 할 계획이다.

6. 참고 문헌

[1] Gerard Salton, "Automatic Text Processing", Addison Wesley publishing company, 1988.

[2] Robertson, S.E. et al. "Okapi at TREC-8," In The Eighth Text REtrieval Conference (TREC-8), 2000.

[3] Donna K. Harman, editor. "The Sixth Text REtrieval Conference (TREC-6)," 1997.

[4] Amit Singhal, Chris Buckley, Mandar Mitra, "Pivot

ed Document Length Normalization," SIGIR, 1996.
[5] Joel L. Fagan, "The Effectiveness of a Non-Syntac tic approach to Automatic Phrase Indexing for Docume nt Retireval," JASIS, 1989.
[6] Joel L. Fagan, "Experiment in Automatic Phrase In dexing for Document Retireval: a Comparison of Synta ctic and Non-Syntactic Methods," Ph.D. thesis, Cornell University, 1987.
[7] Mandar Mitra, et al., "An Analysis of Statistical a nd Syntactic Phrase," RIAO'97 Computer-Assisted Info rmation Searching on Internet, pp. 200-214, 1997.
[8] 윤보현 외, "한국어 정보검색1.에서 구문적 용어 불일치 완화방안", 제 10회 한글 및 한국어 정보 처리 학술 발표 논문집, pp. 143-149, 1998.
[9] 김상범 외, "고려대학교 정보검색엔진 KUIR의 구조 및 특징," KOSTI 2000 한글 정보검색 학술발표 논문집, 2000.
[10] 장명길 외, "HANTEC 3.0에서의 키팩트 기반 텍스트 검색 방법에 관한 연구," KOSTI 2000 한글 정보검색 학술발표 논문집, 2000.
[11] 원형석 외. "복합명사 처리를 위한 통합 다단계 색인모델," HCI'99 학술발표대회 논문집, pp 80-87, 1999.
[12] 이석훈 외, "정보검색 평가체제 구축을 위한 HANTEC 테스트 컬렉션의 패키징," KOSTI 2000 한글 정보검색 학술발표 논문집, 2000.
[13] 이호, "언어 정보 획득을 위한 한국어 코퍼스 분석 도구", 고려대 석사학위 논문, 1994.
[14] 김진동 외, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델," 한국정보과학회 논문지, 24(12), 1997.