제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

단어 가중치 값을 이용한 복합명사 제한적 확장 및 검색 성능 개선

김현진*, 이충희*, 허정*, 장명길*
*한국전자통신연구원 휴먼정보검색연구팀
e-mail : jini@etri.re.kr

Improvement of retrieval system and generation of compound noun using word weight method

Hyun-Jin Kim*, Chung-Hee Lee*, Jeong Hur*, Myeong-Gil Jang*
*Dept. of Human Information Processing, ETRI

요 약

자동색인이나 정보검색 엔진에서는 효율적인 색인어 추출이 주요한 요인으로 작용한다. 특히 색인 집합의 많은 부분을 차지하는 복합명사의 경우에는 색인과 검색 두 분야 모두에 큰 문제로 여겨져 왔다. 본 논문에서는 복합명사를 이루는 단일 단어 중에 단어 가중치가 높은 것을 중심으로 복합명사를 확장하는 방식을 이용하여, 색인어를 추출하여, 복합명사가 제한적으로 확장되는 효과를 보여 주며, 검색에서는 질의문에 나타나는 명사들에 이러한 가중치 값을 적용하여 검색에 효과를 높여 주는 방식을 제안한다.

1. 서론

한국어에서 복합명사는 명사구를 표현하는 방식으로도 쓰이고, 사용자의 의도에 따라서 조합이 용이하여 정보검색이나 기계번역 등과 같은 언어 처리 응용 시스템에서 그 처리에 대한 방안이 많이 논의 되어왔다.. 특히 정보검색에서는 명사나 명사구가 색인어로서의 비중이 크고, 그 중에서도 단일어보다는 복합명사가 문서를 대표할 수 있는 특징적인 형태를 가지고 있다. 따라서 정보검색에서 이러한 복합명사를 적절히 다루게 되면 색인 추출에 따르는 비용의 절감과 함께 검색의 정확도를 높일 수 있는 결과를 낼 수 있다.

본 논문에서 다루는 복합명사의 경우는 3 단어 이상의 복합명사에서 의미있는 2 단어 이상의 복합명사들을 확장하는 것에 대한 것으로 기존에 코퍼스에서 공기정보와 단어의 위치 정보를 이용하여 확장하는 방식과는 달리 3 단어 이상의 복합명사를 이루는 단일 명사들의 각각의 가중치 값을 이용하여, 그 순위의 조합으로 명사를 확장하는 방식을 제안한다.

또한 이러한 명사의 가중치 값을 이용하여, 질의문에 나타나는 명사들의 가중치를 비교하여, 높은 가중치를 가진 색인어를 먼저 검색하게 하는 방식을 이용한 검색 효율을 높이는 방식도 제안하였다. 질의문에 이러한 가중치를 적용하면, 적합 문서 수를 미리 줄여서 검색 속도의 감소와 함께 정확도를 높이고 향후 질의 확장 등에 유용한 것으로 확인되었다.

2 장에서는 복합명사와 관련된 기존 연구들을 알아보고, 3 장에서는 단어 가중치 값을 적용한 복합명사의 제한적인 확장에 대한 방법론을 설명하고, 4 장에서는 이를 검색에서 질의문에 적용하는 방법을 보여주고, 5 장에서는 이와 관련한 각 실험에 대한 내용과 분석된 결과에 대해 설명한다. 6 장에서는 결론과 함께 향후 연구 방향에 대하여 논의한다.

2. 기존 연구

한국어 언어처리 분야에서는 그 동안 복합명사에 대한 연구가 활발하게 되어져 왔다. 정보검색 분야에서의 복합명사에 대한 연구는 주로 사용자의 의도에 따라서 명사끼리 조합이 자유로운 특징을 지닌 한국어 복합명사 처리에 대한 것으로 문장 내에서 띄어쓰

기를 하지 않은 복합명사를 단일 명사 조합으로 분리해 내는 방법[1][2][3][4]과 3 단어 이상의 복합명사에서 유효한 2 단어이상의 복합명사들을 분리하는 것에 대한 연구들[5][6]과 반대로 복합명사의 합성 규칙들을 이용하여, 명사구에서 복합명사로 추출해 내는 복합명사 생성[7]으로 크게 나눠 볼 수 있다.

여기서 3 단어 이상의 복합명사에서 유효한 복합명사들로 분리하는 데에는 다시 통계적인 방법을 이용하는 방법과 구구조를 활용하여 분리하는 방식 등의 연구가 이루어져 왔다. 통계적 정보를 이용하는 방식에서는 복합명사를 구성하는 단일어들의 통계적 행태 분석을 통하여 복합명사 개개의 어휘적 특성을 자동으로 획득하고 이를 이용하여 검색하는 방법을 제시하고 있고[6], 구구조를 활용하는 방법에서는 명사구 내의 단어들의 언어학적인 관계와 이를 기반으로 말뭉치에서 추출한 어휘간 공기관계를 바탕으로 복합명사의 구문구조를 분석하는 방식을 제시하고 있다[5].

3. 단어 가중치 값을 이용한 복합 명사의 제한적 확장

3.1. 주요 개념

단어 가중치 값을 이용한 복합명사의 확장 개념은 다음과 같다.

예를 들어, "청소년+흡연+규제+강화"라는 4 단어의 복합명사가 있을 때, 기존의 연구에서는 이러한 4 단어의 복합명사에서 3 단어와 2 단어의 복합명사로 구분해 낼 때, 기준으로 삼는 것이 그 조합된 단어가 실제로 쓰이는 유효한 단어이어야 하는 것이다. 즉, "청소년+흡연+규제+강화"라는 복합명사에서는 다음과 같은 단어의 조합이 추가로 확장될 수 있다.

청소년+흡연+규제+강화
청소년+흡연+규제
청소년+흡연+강화
청소년+규제+강화
흡연+규제+강화
.....
청소년+흡연
규제+강화
흡연+규제
.....

그런데, 여기서 보면 "청소년+흡연+규제+강화"라는 명사구에서는 "청소년+흡연"이 주요한 키(key)가 되는 단어 조합이라고 볼 수 있다. 검색의 측면에서 다시 보면, "흡연"에 관련된 문서 중에서도 "청소년+흡연"에 관한 것이 우선이고, 그러한 문헌 내에서도 "청소년+흡연"의 "규제"를 "강화"에 대한 문헌이 가장 적합한 문헌이라고 볼 수 있다. 이런 관점에서 보면 다음의 단어 조합만이 유효한 것이라고 볼 수 있다.

청소년+흡연+규제+강화
청소년+흡연+규제
청소년+흡연

"흡연+규제", "규제+강화" 등의 복합명사들은 기존의 관점에서는 유효한 복합명사의 확장이라고 볼 수 있으나, 검색의 효율적인 관점에서는 오히려 그 주요성이 떨어진다고 본다. 왜냐하면, "흡연+규제"는 "청소년+흡연+규제"로 한정되어지는 문헌의 대표성을 분산시킬 소지가 있고, "규제+강화" 또한 주제가 한정 되지 않는 일반적인 단어이므로 검색에서는 그 효용성이 떨어진다고 볼 수 있다.

따라서 본 논문에서는 이러한 관점에서 복합명사를 제한적으로 확장하는 데에 각 복합명사를 구성하는 구성명사들의 가중치 값을 이용하였다. 우선 복합명사를 구성하는 단일명사들이 색인 집합에서 가지는 각각의 가중치 값은 2-poisson model 을 이용하였다.

$$w_{ij} = \frac{tf_i}{k_1 \cdot \left((1-b) + b \cdot \frac{dl_j}{avdl}\right) + tf_i} \cdot \log \frac{N - df_i + 0.5}{df_i + 0.5}$$

가중치값을 구하는 집합으로는 Hantec2.0 12 만 문서셋으로 하였다. 총 40 만개의 단일명사의 개별 가중치 값을 구하고, 각각의 평균값을 저장하여서 복합명사를 이루는 단일명사의 가중치 값으로 계산하였다.

즉, "청소년+흡연+규제+강화"에 위의 가중치 값을 이용하여 분리해보면, 다음과 같다.

청소년(0.81) 흡연(1.19) 규제(0.65) 강화(0.47)

결과에서 보듯이, "흡연"이 중심 단어가 되고 "청소년"과 "규제"가 그 다음의 가중치 값을 나타내고 있다. 여기서 가중치 순위별 조합으로 3 단어 복합명사를 확장해 내면, "청소년+흡연+규제"가 되고, 순위별 조합으로 2 단어 복합명사를 확장해 내면 "청소년+흡연"으로 나타난다. 여기서 복합명사의 확장에서 구성명사의 위치가 바뀌어서는 안 된다는 제한이 있다.

3.2. 실험 및 분석

이러한 문헌에서 나타난 단어 가중치 값이 실제 복합명사의 구성명사로 나타난 명사의 가중치로 이용할 때, 어느 정도의 효율성을 가지는 지 알아보기 위해 다음과 같은 실험을 하였다.

단어 가중치 값은 색인어를 구할 때 쓰는 가중치 값을 그대로 이용하였다. 실험 대상으로 뽑은 복합명사는 국어정보베이스 1,2 조선일보 이규태코너에서 4 단어 복합명사 100 개와 3 단어 복합명사 100 개를 각각 추출하였다. 4 단어 복합명사의 경우에는 순위별 3 단어 복합명사 1 개와 2 단어 복합명사 2 가지를 확장

해서 총 3 종류의 복합명사로 확장하였고, 3 단어 복합명사의 경우에는 순위별 2 단어 복합명사를 2 개를 추출하였다. [표 1]은 각각 복합명사를 순위별로 조합한 결과의 일부이다.

[표 1] 복합명사 확장 실험 결과의 일부

복합명사 (순위)	3 단어 조합 (1,2,3 순위)	2 단어 조합 (1,2 순위)	2 단어 조합 (1, 3 순위)
현재(4)+국내(3)+컴퓨터(1)+환경(2)	국내+컴퓨터+환경	컴퓨터+환경	국내+컴퓨터
시민(1)+운동(2)+단체(3)+관계자(4)	시민+운동+단체	시민+운동	시민+단체
문학(1)+관련(4)+전문(2)+정보(3)	문학+전문+정보	문학+전문	문학+정보
운전(2)+면허(1)+시험(3)+기준(4)	운전+면허+시험	운전+면허	면허+시험
마약(2)+복용(1)+실태(3)+연구(4)	마약+복용+실태	마약+복용	복용+실태
문화(2)+예술(1)+정보(4)+검색(3)	문화+예술+검색	예술+검색	문화+검색
영어(2)+교사(1)+채용(3)+프로그램(4)	영어+교사+채용	영어+교사	교사+채용

[표 1]에서 보듯이, 4 단어 조합 복합명사에서 가중치가 가장 낮은 단어들은 주로 단어의 조합에서 큰 영향을 안주는 단어들로써, 3 조합과 2 조합에서 제거해 냄으로써 단어들의 대표성을 높이는 효과를 보여주고 있다. 다음의 [표 2]는 전체 200 개의 복합명사들을 가중치 값에 따라서 나눈 전체 결과이다.

[표 2] 코퍼스에서 나타난 복합명사 분리 실험 결과

4 단어 복합명사 분리 결과			
	3 단어 조합 (1,2,3 순위)	2 단어 조합 (1,2 순위)	2 단어 조합 (1,3 순위)
적합성	85%	71%	68%
3 단어 복합명사 분리 결과			
		2 단어 조합 (1,2 순위)	2 단어 조합 (1,3 순위)
적합성		75%	71%

각 실험군에서의 적합성은 미리 각 복합명사에서 의미적으로 유효한 복합명사를 구한 뒤, 자동으로 확장된 복합명사들과 비교해서 적합한지를 검사한 결과이다. 실험 결과 70%이상의 적합성을 보여 실제 활용에도 가능한 실험치를 나타냈다.

4. 단어 가중치 값을 이용한 검색 효율 향상

본 논문에서는 복합명사를 구성하는 단일 명사의 가중치 값이 복합명사를 제한적으로 확장하는데 활용해서 그 적합성이 높음을 보고, 이를 검색에서의 질의문장에도 적용해 보았다.

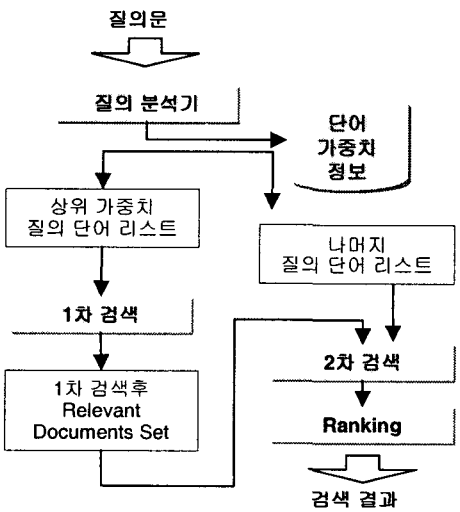
4.1. 주요 개념

복합명사 내에서도 단어의 가중치 값에 따라서 주요한 단어들이 구분되어 지듯이, 질의문장에서도 그 문장을 대표하는 단어들을 구분 지어서, 검색을 단계별로 하게 하면, 적합 문서의 수를 줄이고 검색 효율도 높일 수 있다.

1) 보스니아 내전의 의미와 영향 및 추후전망
2) 가정폭력의 실상과 사회에 미치는 영향
3) 유전자 조작식품에 대한 국제적인 논란
.....

위에 언급한 문장들은 Hantec2.0 에 들어 있는 30 개의 질의문 중 몇 개이다. 1)번 질의문을 보면, 우선은 "보스니아 내전"에 대한 내용이 언급된 문헌들이 적합 문서로 검색되어져야 함을 문장을 읽어 보면 알 수 있다. 그러나 일반적인 검색 엔진에서는 이러한 차이를 미리 알 수 없으므로, "의미", "영향", "추후전망"과 같은 단어들도 질의문 집합에 넣어서 적합 문서를 먼저 다 검색 한 후, 그 중에서 순위 알고리즘으로 상위의 적합한 문서를 사용자에게 제공하게 된다. 그렇게 되면 일반적인 단어인 경우에는 많은 문서에 빈번히 출현함으로써, 적합문서의 수가 많게 되므로 문서의 순위를 계산하는데도 방해를 하게 되므로 검색의 효율이 떨어지는 결과를 초래한다.

따라서 본 논문에서는 [그림 1]과 같은 절차로 검색을 시도하였다. 먼저 입력된 질의문에서 단어 가중치 정보를 구하여, 상위 가중치 값을 가진 단어들을 중심으로 1 차 검색을 하고, 1 차 검색에서 적합문서 집합으로 나온 문서 셋을 2 차 검색의 대상으로 넘긴다. 그 다음에 질의문에서 나머지 단어들을 중심으로 재검색을 시도하여, 순위 알고리즘을 통해서 문서의 순위를 계산한다.



[그림 1] 단어 가중치 값을 이용한 검색 과정

4.2. 실험 및 분석

실험으로는 Hantec2.0 의 12 만 문서 셋을 대상으로 하고, 30 개 질의문으로 검색을 시도하였다. 색인 가중치 값으로는 앞에서 언급한 2-poisson 모델을 이용하였고, 검색 모델은 vector-space 모델을 사용하였다. 실험 결과로는 Relevant document 수의 변화와 2 단계 검색의 검색 정확도의 변화를 측정하였다. 여기서 언급하는 Relevant document 수라고 하면, 질의문에서 분리된 명사들에 따라서 색인화일에서 검색을 할 때, 하나의 질의 명사라도 포함하고 있는 문서들을 모두 합친 수라고 보면 된다.

실험 결과는 다음과 같다.

[표 3] Hantec2.0 실험 결과

Hantec 30 개 질의	일반 검색	단어 가중치 값 적용	비고
질의별 Relevant Document 평균 개수	27,643 건	9,763 건	약 65% 감소
R-precision (G2)	0.48	0.52	0.04 증가

[표 3]에서 보여지듯이, 질의별 Relevant 문서 수는 평균 값을 비교해서 약 65% 감소의 결과가 있었다. 이것은 Vector space 모델의 경우 실제 Ranking 을 하기 위해 검색되어진 문서수의 감소를 의미하므로, 검색시에 속도 개선의 효과가 있고, Precision 이 중요한 관점이 되어진 현재의 검색 엔진의 요구 조건을 만족시킨다고 볼 수 있다.

또한 질의문에서 가중치 단위로 분리된 명사 집단을 중심으로 질의어 확장에도 활용 할 수 있다. 일반적으로 질의문을 확장할 때에는 질의문장에 있는 모든 명사를 대상으로 하기 때문에, 일반적인 단어의 경우에는 확장이 되면 오히려 더 Relevant 문서 수만 늘여서 Precision 이 떨어지게 되는 역효과가 나타날 수 있다. 예를 들어, "보스니아 내전의 의미와 영향 및 추후 전망"이라고 할 때, "의미", "영향", "추후전망"을 1 차 검색 시에 확장하게 되면, 이는 비교적 문서에서 많이 나타나며, 문서의 특성을 반영하는 역할이 낮은 단어이므로 별로 효과적인지 않다.

그러므로 제시한 방법에서처럼 1 차 검색에서는 "보스니아 내전"에 관련된 유사 단어만 확장하여, "보스니아 내전", "보스니아 내란", "보스니아 사태" 등과 같은 질의 확장을 시도하고, 1 차 검색후의 결과들 안에서 "의미", "영향" , "추후전망"와 같은 질의어들을 확장하면 효과적을 검색하는데 도움이 될 수 있다.

5. 결론

본 논문에서는 색인어 가중치 값으로 쓰이는 단어 가중치 알고리즘을 이용하여, 복합명사를 제한적으로 확장하는 방법과 실험 결과를 보이고, 이를 검색에 활용하여, 질의문 분석에 활용한 결과를 보였다. 결과들에서 보여지듯이, 여러 문서에 나타나는 단어들의 문서별 가중치 값이 전체적으로도 일치함을 보이고 있고, 이를 이용하면 문장 내에서의 단어별 가중치 값으로도 활용 가능함을 알 수 있었다.

향후에는 색인 집단의 단어 가중치 값 외에도 대량의 코퍼스에서 정련된 단어 가중치 값을 구해서 실험을 해 볼 필요가 있고, 현재는 전체 가중치 값의 평균만을 이용하여 순위로 분리를 했으나, collocation 정보를 활용해서 상대적인 우선 순위를 구해서 복합명사 확장 시에 시도할 계획이다.

참고문헌

[1] 윤보현, 조민정, 임해창, "통계정보와 선호규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지 제 24 권 제 8 호, 1997.

[2] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지, 제 25 권 제 1 호, 1998.

[3] 장동현, 맹성현, "효율적인 색인어 추출을 위한 복합명사 분석 방법", 제 8 회 한글 및 한국어 정보처리 학술발표논문집, 1996.

[4] 심광섭, "합성된 상호정보를 이용한 복합명사분리", 정보과학회논문지, 제 24 권 제 11 호, 1997.

[5] 윤준태, 정의석, 송만석, "명사간 어휘 정보를 이용한 한국어 복합명사 분석", 정보과학회 논문지, 제 25 권 11 호, 1998.

[6] 박영찬, 최기선, "통계적 정보를 이용한 복합명사 검색 모델", 한국과학기술원 박사학위논문

[7] 김미진, 박미성, 최재혁, 이상조, "효율적인 색인어 추출을 위한 합성명사 생성 방안에 대한 연구", 한국정보처리학회 논문지, 제 7 권 제 4 호, 2000.