한국어 정보검색에서의 복합명사 가중치 부여 방법 및 평가

김지영[0]　　맹성현
충남대학교　컴퓨터과학과
{jykim, shmyaeng}@cs.cnu.ac.kr

Weighting Methods and their Evaluations for Compound Nouns in Korean Text Retrieval

Ji-Young Kim[0]　Hyon-Myaeng Sung
Dept. of Computer Science, ChungNam University

요 약

한국어의 경우 띄어쓰기의 자유로움과 명사들이 비교적 자유롭게 결합하여 새로운 복합명사(compound noun)를 형성한다. 따라서, 정보검색에서 복합명사를 적절하게 처리하게 되면 검색 효율을 향상시킬 수 있다. 본 논문에서는 질의에 포함된 단일명사, 복합명사, 그리고 복합명사를 이루는 구성명사의 적절한 가중치 부여 방법에 대하여 기술한다. 일반적인 tf*idf가중치 방법은 문서 내 빈도수(tf)만을 강조하여 문서 내 발생빈도가 낮은 복합명사의 경우 낮은 가중치를 갖는다. 반대로, 역문헌 빈도수(idf)로 인해 복합명사가 단일명사보다 높은 가중치를 갖게 되면 단일명사의 가중치를 지나치게 떨어뜨려 검색 성능을 저하시킨다. 이런 문제를 해결하기 위해서 복합명사의 통계적인 특성을 고려하고, 복합명사를 이루는 구성명사의 적절한 가중치 사용과 tf*idf 변화 범위에 따른 파라메터를 이용하였다. 결과적으로 본 논문에서는 질의 색인어의 종류에 따라 가중치를 달리 부여함으로써 검색 성능을 향상시킬 수 있는 가중치 부여 방법을 제시하고 검증 실험을 통해 유효성을 제시했다는 점에서 그 의의가 있다고 하겠다.

1. 서론

최근 몇 십년 동안 디지털 정보가 기하급수적으로 증가함에 따라 사용자는 방대한 양의 텍스트 정보로부터 적합한 문서를 찾기 위해 효율적인 정보검색 시스템을 필요로 한다. 특히, 정보검색 시스템은 사용자의 요구에 만족하는 문서를 선택하고 순위를 결정하기 위해 사용자의 질의에 의존한다. 따라서, 질의에서 사용자의 요구를 나타낼 수 있는 정확한 색인어의 추출과 가중치 부여는 정보검색에서 매우 중요하다.

한국어의 경우 질의 색인어(query index term)는 대부분의 문서에서 개념적 중요도가 높은 명사 위주로 추출된다[10]. 그러나, 한국어는 명사들이 비교적 자유롭게 다른 명사들과 결합하여 새로운 복합명사를 이루는 경우가 많다. 이러한 한국어의 속성은 용어의 불일치를 발생하는데 이는 검색 성능을 향상하는 장애요인이 된다. 복합명사로 인해 발생하는 문제 중의 하나로 질의 색인어에 대한 가중치 부여 문제가 있다. 질의 색인어의 가중치는 색인어의 중요도에 따라 차별적으로 부여되며 검색 성능에 직접적인 영향을 준다. 그러나 지금까지의 연구는 문서에서 복합명사를 인식하고 분해,합성하는 방법에 집중되어 있었고 새로운 가중치 부여 방법에 대한 고려가 미흡하였기 때문에 단일명사(single noun)를 위해 고안된 기존의 가중치 부여 방법을 그대로 사용하는 것이 일반적이었다. 일반적인 통계에 기반한 가중치 방법은 문서 안의 빈도수를 사용하는데, 복합명사(compound noun)의 경우 문서 내 발생빈도가 단일명사보다 낮기 때문에 단일명사보다 낮은 가중치를 갖는다[5]. 또한, 복합명사의 가중치를 복합명사를 이루는 구성명사(component noun)의 가중치 함수로 가중치를 계산하는 방법도 사용하였으나 이 방법은 단조스러운 결과를 생성하지 못한다. 그러나, 복합명사와 복합명사를 이루는 구성명사는 단일명사와는 그 성질이 다르며 질의에서의 역할이 다르므로 각각의 중요도에 따른 새로운 가중치 방법이 필요하다.

본 논문의 구성은 다음과 같다. 2장에서 복합명사 처

리와 복합명사 가중치 부여와 관련된 연구를 살펴보고, 3장에서는 질의어의 통합 가중치 부여 방법에 대해 알아보고, 4장에서는 본 논문에서 제안된 가중치 부여 방법의 효용성을 알아보기 위한 다양한 실험과 실험결과에 대해 평가하며, 마지막으로 5장에서 결론을 내리고 향후연구에 대해 기술한다.

2. 관련연구

한국어의 경우 복합명사를 사용하여 검색의 신뢰도를 높이기 위한 연구는 색인어와 질의어간의 일치정도를 이용하여 복합명사의 가중치를 조정하는 방법[11], 통계적 명사패턴 분류를 이용하여 가중치를 조정하는 방법[9], 단어의 발생형태와 문맥연관 강도를 이용하여 가중치를 조정하는 방법[12],복합명사와 구성명사간의 부분정합을 통한 가중치 부여 방법[8] 등이 있다.

영어권의 경우 구와 그 구를 구성하는 성분단어를 함께 색인어로 사용하는 것이 일반적인 경향이다. 추출된 구를 대상으로 검색에 적용하는 연구는 단어와 구로 명확히 구분된 벡터를 이용하여 질의와 유사도 계산을 수행하는 방법[3], 추출된 구를 성분단어의 연결 리스트로 표현하고 질의와 비교하여 가중치를 조정하는 방법[6], 질의를 분석하여 구 단위로 만들어진 파스 트리를 문서의 단어들과 비교하는 방법[2], 구 단위로 만들어진 파스 트리를 복잡한 알고리즘을 이용하여 트리 매칭을 하는 방법[1]이 있다.

그러나, 위 연구들의 경우 질의 색인어의 어느 한 종류의 가중치 부여 방법에 대한 연구가 많았고, 가중치를 부여하기 위해 많은 시간이 소요되었다. 또한, 구조적 표현을 통한 가중치 부여 방식은 자연어의 구조적 정보를 이용하여 구와 구성명사의 가중치를 고려하지만, 기존의 벡터 표현에 기반을 둔 가중치 부여 방법을 사용할 수 없기 때문에 성능이 비교적 떨어지고 기존의 정보검색 시스템에 적용하기 어려운 단점이 있다.

3. 복합명사의 가중치 부여

본 장에서는 질의 색인어의 종류에 따른 가중치 부여 방법에 대해 몇 가지 실험을 통해 살펴보고, 전체적인 질의에 대한 가중치 부여 방법에 대해 기술하였다. 이때 복합명사의 가중치 방법을 결정하기 위해 복합명사가 다수 존재하는 질의를 통해 가중치 식을 결정해 나가며 관련되는 파라미터(parameter) 값에 대한 영향을 관찰한 후 값을 결정하는 방법을 사용하였다. 이때 구성되는 질의벡터는 복합명사와 단일어(구성명사+단일명사)로 구성이 된다. 단일어의 가중치는 구성명사 가중치와 단일명사 가중치의 합을 사용하여, 만약 문서가 단일명사만을 포함할 경우도 검색될 수 있도록 하였다.

본 논문에서는 tf, idf를 이용하되 복합명사와 구성명사의 특성을 고려하여 변형된 식을 사용하였고, 복합명사, 구성명사, 그리고 단일명사의 가중치 부여를 위해 다음과 같은 점을 고려하였다.

- 복합명사의 중요도에 따라 가중치를 부여하기 위해 구성명사의 가중치 성분을 이용하여 가중치를 높여 주었다.
- 복합명사에서 분해되어 생성된 구성명사는 독립적으로 하나의 개념을 나타내기 위해 사용된 것이 아니라, 복합명사 안에서 일정한 역할을 위해 사용된 것이다. 따라서, 단일명사와는 가중치 면에서 달리 취급되어야 한다. 본 논문에서는 구성명사가 질의에서 복합명사의 구성명사로 사용된 빈도수를 가중치 계산에 적용하였다.
- 독립적으로 하나의 개념을 표현하는 단일명사는 질의 안에서의 빈도수를 고려하여 가중치 계산에 사용하였다.

다음은 본 논문에서 제안한 복합명사 가중치 계산식이다.

- 복합명사의 가중치(W_c)
  - 복합명사 자체가 갖는 가중치(X)

$$X=\left(\frac{idf_c}{\max idf}*\beta+(1-\beta)\right)*\left(\frac{tf_c}{\max tf}*\gamma+(1-\gamma)\right)$$

  - 구성명사를 통해 갖는 가중치(Y)

$$Y_A=0,\quad Y_B=\left(\frac{\sum_{i=1}^{n} idf_{pi}}{n*\max\ idf}\right)$$

$$Y_C=\log\left(\frac{n*\max idf}{\sum_{i=1}^{n} idf_{pi}}\right),\quad Y_D=\left(\frac{n*\min idf}{\sum_{i=1}^{n} idf_{pi}}\right)$$

$\beta,\gamma : tf, idf$의 변화 범위 결정

idf_c : 복합명사역문서빈도

tf_c : 복합명사질의내발생빈도수

idf_{pi} : 복합명사의구성명사역문서빈도

  - 복합명사의 가중치(W_c)

A. $W_C=\alpha X+(1-\alpha)*Y_A$

B. $W_C=\alpha X+(1-\alpha)*Y_B$

C. $W_C=\alpha X+(1-\alpha)*Y_C$

D. $W_C=\alpha X+(1-\alpha)*Y_D$

- 구성명사의 가중치(W_{pi})

$$W_{pi}=\left(\frac{idf_{pi}}{\max idf}*\beta+(1-\beta)\right)*\left(\frac{tf_{pi}}{\max f}*\gamma+(1-\gamma)\right)$$

idf_{pi} :구성명사 역문서 빈도

tf_{pi} :복합명사의 구성명사로 질의내 발생하는 빈도수

■ 단일명사의 가중치(W_s)

1. A,B,C 방법에 사용.

$$W_s=\left(\frac{idf_s}{\max idf}*\beta+(1-\beta)\right)*\left(\frac{tf_s}{\max f}*\gamma+(1-\gamma)\right)$$

2. D 방법에 사용

$$W_s=\left(\frac{idf_s}{\max idf}*\beta+(1-\beta)\right)*\left(\frac{tf_{all}}{\max f}*\gamma+(1-\gamma)\right)$$

idf_s :단일명사 역문서 빈도

tf_s :단일명사 질의내 발생 빈도수

tf_{all} :단일어(구성명사+단일명사)질의내 발생빈도수 (tf_s+tf_{pi})

여기서, α값은 복합명사의 가중치가 과도하게 부여되는 것을 막기 위하여 조절해주는 상수이며, β,γ는 가중치 계산에 사용되는 tf, idf의 변화범위를 결정하는 상수이다. 즉, 상수 α,β,γ는 복합명사의 가중치를 결정하는 데 이용되는 파라미터로서 각각의 값의 변화에 따라 평균 정확도에 어떤 영향을 주는지를 나타낸다.

위 식을 살펴보면, 복합명사의 가중치는 복합명사 자체가 갖는 가중치와 구성명사를 통해 갖는 가중치로 나누어 볼 수 있는데, 이때 구성명사의 idf를 이용하였다. 각 복합명사 가중치 방법들을 살펴보면, 방법 A는 일반적인 통계적 방법인 tf, idf를 이용한 방법이며, B,C,D는 복합명사 자체가 갖는 가중치에 복합명사를 이루는 구성명사의 가중치 성분을 더하여 질의에서의 복합명사의 중요도를 높여주었다. 복합명사를 이루는 구성명사는 단일명사와 차별을 두기 위해 질의에서 복합명사의 구성명사로 발생하는 빈도수를 구하여 tf로 대신 사용하였다. 단일명사의 경우, 방법 A,B,C의 경우 질의 내에서 단일명사만으로 사용되는 빈도수를 이용한 것이고, 방법 D의 경우 단일명사와 구성명사 구분 없이 질의 내에서 출현하는 전체 빈도수를 사용하였다.

위 식의 검색의 효용성을 알아보기 위해 HANTEC 테스트 컬렉션[7]을 이용하여 몇 가지 실험을 하였고, 이때 질의는 HANTEC 질의 50개 중에서 복합명사를 많이 포함한 질의 5개를 골라 사용하였다. 이때 평가 방법은 11점 평균 정확도(11-point average precision)을 사용하였다.

다음은 본 실험에서 사용된 질의의 예이다.

<num> 08
<title> 유통시장
<desc> 국내 유통시장에 대한 국내외 대기업 진출 현황
<narr> 국내 유통시장에 진출한 국내 대기업의 사업 내용이나 국내 유통시장 진출을 위하여 기업합병 및 인수를 시도하고 있는 외국 기업들의 사업내용에 관한 문서는?
<quer> 유통시장 대기업 외국기업 사업 진출 마케팅 인수 합병

[그림3-2] 실험에 사용된 질의의 예

먼저, 위 식을 이용하여 검색하여 각 질의에 대한 평균 정확도를 살펴보았다. 이때 α={0.3, 0.5, 0.7, 0.9}, β=0.5, γ=0.5를 사용하였다.

실험결과를 보면, 복합명사의 가중치 방법들 모두 α값에 따라 평균 정확도가 달랐고, α값이 0.5일 때 비교적 좋은 평균 정확도를 갖는 것으로 보아 복합명사 자체가 갖는 가중치와 구성명사를 통해 갖는 가중치의 적당한 조절을 통해 가중치를 부여할 필요가 있음을 알 수 있다. [표3-1]은 각 방법별로 가장 좋은 평균 정확도를 갖는 α값과 각 질의의 평균 정확도를 정리한 것이다.

[표3-1] 복합명사 가중치 부여 방법 실험 1

방법	α	8	9	11	19	47	전체
A	1	0.1557	0.3755	0.3448	0.2963	0.5242	0.3397
B	0.5	0.1509	0.3709	0.3517	0.3333	0.5242	0.3462
C	0.9	0.1509	0.3775	0.3448	0.3148	0.5242	0.3425
D	0.5	0.1509	0.3775	0.2966	0.3333	0.5403	0.3397

위 [표3-1]을 보면 방법 B가 α값이 0.5일 때 0.3462로 가장 좋은 성능을 보였다. 각 방법들은 α값에 따라 평균 정확도가 달라지는데, 이것으로 보아 적절한 α값의 사용은 검색 성능에 좋은 영향을 주는 것을 알 수 있다.

두번째로, 복합명사에 대한 각 방법별 가중치 계산을 복합명사 자체가 갖는 가중치와 구성명사를 이용한 가중치의 관계를 이용하였는데, 첫번째 실험의 경우 복합명사의 가중치 증가보다는 평균치를 사용하였다면 이 실험에서는 구성명사의 가중치를 사용하여 복합명사의

가중치를 높여 주었다. 그리고, *tf*, *idf* 변화 범위를 달리하여 가중치를 부여하였다.

다음은 두번째 실험에서 복합명사의 가중치 계산을 위해 사용된 식이다.

■ 복합명사의 가중치(W_c)

A. $W_C = X + (1-X) * Y_A$

B. $W_C = X + (1-X) * Y_B$

C. $W_C = X + (1-X) * Y_C$

D. $W_C = X + (1-X) * Y_D$

질의 색인어의 *if*, *idf*의 변화 범위에 따른 평균 정확도를 살펴보기 위해 다음과 같이 네 가지 세부 실험을 하였다.

1. β=0.5, γ=0.5을 사용
2. β=0.3, γ=0.7을 사용
3. β=0.5, γ=0.7을 사용
4. β=0.7, γ=0.7을 사용

[표3-2]는 각 방법이 네 가지 세부 실험에서 가장 좋은 평균 정확도를 갖는 경우를 나타내며, 방법 A가 β=0.5, γ=0.7을 사용하여 검색했을 때 0.3495로 가장 좋은 검색 성능을 보였다.

[표3-2] 복합명사 가중치 부여 방법 실험 2

방법	β·γ	8	9	11	19	47	전체
A	β=0.5,γ=0.7	0.1651	0.3907	0.3448	0.3148	0.5323	0.3495
B	β=0.3,γ=0.5	0.1698	0.3841	0.3310	0.2963	0.5161	0.3395
C	β=0.3,γ=0.7	0.1604	0.3841	0.3103	0.2963	0.5242	0.3351
D	β=0.5,γ=0.5	0.1415	0.3642	0.2966	0.2963	0.5323	0.3262

두번째 실험결과를 살펴보면, 두번째 실험은 첫번째 실험보다 평균 정확도가 좋지 않았고, 복합명사 자체가 갖는 가중치만을 사용한 방법 A가 복합명사 자체가 갖는 가중치와 구성명사의 가중치 성분을 더한 다른 방법 B,C,D보다 좋은 평균 정확도를 보였는데, 이는 복합명사의 가중치가 과도하게 부여되어 오히려 검색 성능 향상에 악영향을 준 것으로 보인다. 또한, *tf*, *idf*의 범위 변화에 따른 각 방법들의 평균 정확도를 살펴보면, γ=0.7일때, β값이 0.3, 0.5, 0.7로 늘어남에 따라 평균 정확도가 떨어졌고, β=0.5일 때 γ값이 0.5에서 0.7로 늘어남에 따라 역시 평균 정확도가 떨어졌다. 즉, β,γ 값은 복합명사의 가중치를 조절하여 검색에 영향을 주는 것을 알 수 있다.

이 실험을 통해서 복합명사의 가중치를 적절하게 조절해주는 α값이 필요하며, *if*, *idf*의 변화 범위 또한 검색 성능에 영향을 주는 것을 알 수 있었다.

세번째는, 위 두 실험을 통합한 가중치 부여 방법으로 두번째 실험에 α값을 적용하여 복합명사의 가중치를 조절해 주었다. 다음은 복합명사 가중치 부여에 사용된 식이며, 그 외 단일명사 및 구성명사의 가중치는 두번째 실험과 동일하다.

■ 복합명사 가중치(W_c)

A. $W_C = \alpha X + (1-\alpha) * (1-X) * Y_A$

B. $W_C = \alpha X + (1-\alpha) * (1-X) * Y_B$

C. $W_C = \alpha X + (1-\alpha) * (1-X) * Y_C$

D. $W_C = \alpha X + (1-\alpha) * (1-X) * Y_D$

세 번째 실험에서도 두번째 실험과 마찬가지로 α, β, γ 값에 따라 세가지 세부 실험을 하였다.

1. α={0.3, 0.5, 0.7}, β=0.5, γ=0.5을 사용
2. α={0.3, 0.5, 0.7}, β=0.3, γ=0.7을 사용
3. α={0.3, 0.5, 0.7}, β=0.5, γ=0.7을 사용

다음 [표3-3]은 실험3에 대한 세가지 세부 실험에서 가장 좋은 평균 정확도를 갖는 경우를 나타낸다.

[표3-3] 복합명사 가중치 부여 방법 실험 3

방법	β·γ	α	8	9	11	19	47	전체
B	β=0.3, γ=0.7	0.5	0.1698	0.3709	0.3586	0.3519	0.5403	0.3583
C	β=0.3, γ=0.7	0.5	0.1698	0.3775	0.3310	0.3519	0.5403	0.3541
D	β=0.5, γ=0.5	0.7	0.1462	0.3775	0.3310	0.3333	0.5403	0.3457

실험3은 전체적으로 앞 두 실험보다 좋은 평균 정확도를 보였고, 특히, β=0.3, γ=0.7 그리고 α=0.5를 갖는 방법 B가 0.3583으로 가장 좋은 평균 정확도를 보였다.
즉, 복합명사 가중치 부여에 있어서 복합명사의 가중치를 적절하게 조절하기 위한 α값과 *tf*, *idf* 값의 범위 변화가 검색 성능의 향상에 좋은 영향을 주는 것을 알 수 있다.

세가지 실험을 분석해보면, 방법 B가 가장 좋은 성능을 보이며, 방법 C와 D의 경우, 논리적으로 봤을 때 구성명사가 일반적(general)이면 가중치는 올라가고 특정단어(specific word)이면 가중치가 떨어뜨리는 효과를 보인다는 면에서는 같으나 평균 정확도에서 차이를

보였다. 이는 단일명사의 가중치의 차이라고 할 수 있는데, 방법 D의 경우 단일명사의 빈도수를 단일명사뿐만 아니라 구성명사의 빈도수까지 더해줌으로써 단일명사의 가중치가 높아져 검색 성능을 떨어뜨리는 효과를 가져온 것으로 보인다. 결과적으로 단일명사, 복합명사, 그리고 구성명사 각각의 가중치 부여 방법이 필요하며, 이를 위해 복합명사의 가중치를 적절하게 조절해 줄 수 있는 α값과 *if*, *idf*의 범위 변화를 결정하는 β,γ 값의 적절한 사용이 검색 성능에 영향을 주는 것을 알 수 있다.

4. 실험 및 평가

본 논문에서 제안한 질의어의 통합 가중치 부여 방법의 유효성을 검증하기 위해 3장에서 관찰한 가중치 부여 방법을 실험을 통해 통계적으로 유효한 결과를 도출하였다.

4.1 실험 및 평가 방법

본 실험을 위해 벡터공간 모델(Vector Space Model)을 검색 모델로 한 정보검색 시스템을 사용하였다. 질의색인어의 가중치는 본 논문에서 제안한 통합 가중치 부여방법을 이용하였고, 색인어의 가중치는 2-Possion Model[4]를 사용하였다.
다음은 색인어의 가중치 부여를 위해 사용된 식이다.

$$Wi=\frac{tfi}{k\cdot((1-b)+b\cdot\frac{DocLength}{AvgDocLength})+tfi}\cdot(\log\frac{N-df+0.5}{df+0.5})$$

k = 2.0, b = 0.75, N = 총 문서수,

df = 문서빈도수, tf_i = 키워드 빈도수

문서 벡터와 질의 벡터 간의 유사도 계수는 다음과 같은 내적(inner product)이 사용되었다.

$$Sim(Q,D_i)=\sum_{j=1}^{n} w_{qi}\times w_{dj}$$

w_q : 질의 가중치, w_d : 문서 가중치

실험에 사용된 실험 데이터는 한국어 정보검색의 성능 평가를 위한 한글 테스트 컬렉션 HANTEC2.0[7]을 사용하였다. HANTEC 테스트 컬렉션은 12000만 건의 문서집합, 50개의 질의집합, 그리고 각 질의에 대한 적합 문서로 구성된 국내 최대의 한국어 정보검색 테스트 컬렉션이다. 본 실험에서는 복합명사를 포함한 30개의 질의를 골라 사용하였다.

각 실험에 대한 평가 방법은 11점의 재현율 각각에 해당하는 정확율을 평균한 11점 평균 정확도(11 point average precision)를 사용하였다.

4.2 복합명사 가중치 부여 방법에 대한 평가

앞 장에서의 세 가지 실험을 통해 정의된 가중치 부여 방법의 유효성을 제시하기 위해 복합명사를 포함한 질의 30개를 대상으로 실험하였다. 본 실험은 앞 장에서의 가중치 실험이 질의 30개에 적용했을 때도 같은 결과가 나오는지 알아보기 위한 것으로, 앞 장에서 같이 세 가지 실험을 실시하였고 그에 따른 α, β, γ 에 따른 평균 정확도의 변화를 살펴보았다.

다음은 각 실험을 간단하게 요약한 것이다.

- 실험1: 복합명사 가중치 계산식 = $\alpha X+(1-\alpha)*Y$
  α={0.3, 0.5, 0.7}, β=0.5, γ=0.5
- 실험2: 복합명사 가중치 계산식 = $X+(1-X)*Y$
  β={0.3, 0.5, 0.7}, γ={0.5, 0.7}
- 실험3: 복합명사 가중치 계산식 = $\alpha X+(1-\alpha)*(1-X)*Y$
  α={0.3, 0.5, 0.7}, β={0.3, 0.5,0.7},
  γ={0.5, 0.7}

아래의 각 표들은 세가지 실험 결과를 나타낸다.

[표4-1] 실험결과1 : 각 가중치 방법에 대한 11점 평균 정확도 비교

방법 \ α	0.3	0.5	0.7
A	0.3274		
B	0.3292	0.3295	0.3291
C	0.3294	0.3278	0.3285

[표4-2] 실험결과2 : 각 가중치 방법에 대한 11점 평균 정확도 비교

β, γ	A	B	C
β=0.5,γ=0.5	0.3197	0.3187	0.3185
β=0.3,γ=0.7	0.3463	0.3410	0.3317
β=0.5,γ=0.7	0.3497	0.3398	0.3311

[표4-3] 실험결과3 : 각 가중치 방법에 대한 11점 평균 정확도 비교

β, γ	방법 α	0.3	0.5	0.7
β=0.5	B	0.3358	0.3322	0.3296
γ=0.5	C	0.3341	0.3321	0.3303
β=0.3	B	0.3518	**0.3550**	0.3508
γ=0.7	C	0.3514	0.3543	0.3505
β=0.5	B	0.3509	0.3505	0.3486
γ=0.7	C	0.3483	0.3473	0.3464

각 실험의 결과를 살펴보면 앞 장에서의 결과와 같은 결과를 보이고 있다. 즉, 실험1의 경우 α값에 따라 다른 평균 정확도를 가지면 복합명사 자체가 갖는 가중치를 사용한 A보다 구성명사의 가중치 성분을 이용한 방법 B,C,D가 더 좋은 성능을 보였다. 반대로 실험2에서는 복합명사의 가중치가 과도하게 부과된 방법 B,C,D가 방법 A보다 평균 정확도가 낮았다. 복합명사의 가중치 조절을 위해 α값과 *tf, idf* 의 범위에 변화를 준 실험3은 전체적으로 가장 높은 평균 정확도를 갖으며, α값이 0.5이고 *tf, idf* 변화 범위가 β=0.3, γ=0.7을 갖는 방법 B가 0.3550로 가장 좋은 결과를 보였다. 즉, 적절한 복합명사의 가중치 부여를 위한 α값과 *tf, idf* 변화 범위의 사용이 검색의 성능 향상에 좋은 영향을 줌을 알 수 있었다.

5. 결론 및 향후연구

한국어의 경우 띄어쓰기의 자유로움과 명사들이 비교적 자유롭게 결합하여 새로운 복합명사를 형성한다. 따라서, 정보검색에서 복합명사를 적절하게 처리하게 되면 시스템의 검색 효율을 향상시킬 수 있다. 본 논문에서는 질의에 포함된 단일명사, 복합명사 그리고 구성명사의 적절한 가중치 부여 방법을 여러 가지 실험을 통해 살펴보고 이를 통해 검색의 성능을 향상시킬 수 있는 방법에 대하여 기술하였다.

tf, idf 가중치 방법은 문서 내 빈도수만을 강조하여 문서 내 발생빈도가 낮은 복합명사의 경우 낮은 가중치를 갖는다. 또한, 복합명사는 개념적 중요도에 따라 단일명사보다 높은 가중치를 갖지만, 과도하게 부여되면 오히려 검색 성능을 저하시킨다. 이런 문제점을 해결하기 위해 여러 가지 실험을 통해 복합명사의 가중치 방법을 도출하였다. 먼저, 복합명사 자체가 갖는 가중치에 구성명사의 가중치 성분을 더하여 복합명사의 가중치를 높여주었고, 복합명사가 적절한 가중치를 갖도록 하기위해 구성명사가 갖는 가중치와 구성명사를 통해 갖는 가중치를 적절하게 조절하기 위한 α값의 사용과 *tf, idf* 변화 범위에 따른 평균 정확도를 알아보았다. 결과적으로 본 연구는 질의 색인어의 종류에 따라 가중치를 달리 부여함으로써 검색 성능을 향상시킬 수 있는 가중치 부여 방법을 제시하고 검증 실험을 통해 유효성을 제시했다는 점에서 그 의의가 있다고 하겠다.

향후 연구로는 본 논문에서 제안한 가중치 부여 방법들과 기존 연구와의 실험적 비교가 필요하며, 복합명사 가중치 부여 방법에 대한 개선과 질의뿐만 아니라 문서 색인어에 적용이 요구된다. 또한, 명사구의 합성을 통해 생성된 복합명사에 대한 가중치 부여 방법에 대한 연구도 요구되며, 질의 특성에 따른 동적인 가중치 부여 방법에 대한 연구와 실험을 통한 새로운 가중치 부여 방법이 필요할 것으로 보인다.

6.참고문헌

[1] Alan Smeaton, Ruairi O'Donnell, Fergus Kelledy, "Indexing Structures Derived from Syntax in TREC-3: System Description," TREC-3 Report, 1994.

[2] Giulia Galbiati, "A Phrase-Based Matching Function," Journal of the American Society for Information Science, Vol. 42, No. 1, pp36-48, 1991.

[3] Joel Fagan, "Automatic Phrase Indexing for Document Retrieval : An Examination of Syntactic and Non-Syntactic Methods," In Proc. of 10th ACM SIGIR Conf., pp91-101, 1987.

[4] S. E. Rovertson, S. Walker, "On Relevance Weights with Little Relevance Information," Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1997.

[5] Tomek Strzalkowski, "Natural Language Information Retrieval," Information Processing & Management, Vol.31, No3, pp337-417, 1995.

[6] Y. Ogawa, A. Bessho, M. Hirose, "Simple String as Compound Keywords: An Indexing and Ranking Method for Japanese Texts," In Prof. of ACM SIGIR, pittsburg, PA. USA, pp227-236, 1993.

[7] 김지영, 장동현, 맹성현, 이석훈, 서정현, 김현, "한국어 테스트 컬렉션 HANTEC의 확장 및 보완," 제 12회 한글 및 한국어 정보처리 학술대회, pp210-215, 2000.

[8] 김평, "정보검색에서의 복합명사 부분정합 기법 및 지원 저장구조 개발," , 충남대, 석사학위논문, 1999.

[9] 박영찬, 최기선, "통계적 명사패턴 분류를 이용한 복합명사 검색 모델," 제 8회 한글 및 한국어 정보처리 학술대회, pp21-31, 1996.

[10] 신동욱, "복합명사의 통계적 처리에 대한 평가," 제8회 한글 및 한국어 정보처리 학술대회, pp36-41, 1996.

[11] 윤보현, 김상범, 임해창, "한국어정보검색에서 구문적 용어불일치 완화방안," 제 10회 한글 및 한국어 정보처리 학술대회, pp143-149, 1998.

[12] 최대선, "구색인에서 성분단어의 가중치 부여 방법에 대한 연구," 포항공대, 석사학위논문, 1997.