용어가중치 결합이 검색 효율성에 미치는 영향 연구

최성환[U] 정영미
연세대학교 문헌정보학과
csh-libinfo@hanmail.net[0], ymchung@yonsei.ac.kr

The Impact of Combining Term Wights on Retrieval Effectiveness

Sung-Hwan Choi[U] Young-Mee Chung
Dept. of Library and Information Science, Yonsei University

요 약

본 논문에서는 데이터 결합 영역에서 문서값을 정규화 하는 기법과 결합함수에 따라 용어가중치 결합이 검색성능에 어떤 영향을 미치는가를 분석하였으며, 특히 용어가중치 결합이 실질적으로 효율적인가를 성능 향상률 측면과 검색시스템의 효율성 측면에서 검증하고, 성능이 향상된 용어가중치 결합의 특징을 분석하였다. 실험결과 대부분의 용어가중치 결합은 문서값 정규화 기법과 실험집단에 관계없이 높은 성능 향상률을 보이지 않았다. 특히 단일가중치로 높은 검색성능을 보였던 상위 가중치 알고리즘들은 다른 가중치 알고리즘과 결합할 경우 두드러진 성능 향상률을 보이지 않았다. 검색시스템의 효율성 측면에서 용어가중치 결합을 평가한 결과 문헌 내 단어빈도를 최대단어빈도로 정규화한 가중치 알고리즘이 코사인 정규화 기법을 적용한 가중치 알고리즘들과 결합될 때 5개 실험집단에서 최적 단일가중치 보다 2% 이상 높은 성능을 보였다. 이는 서로 다른 특성을 지니는 용어가중치 알고리즘들이 장단점을 보완하여 검색성능을 향상시킬 수 있다는 것을 의미한다. 그러나 용어가중치 결합의 효율성은 컬렉션과 가중치 알고리즘의 특성에 의존적이었으며, 비록 각 용어가중치 결합의 성능이 높게 나타날지라도 최적의 성능을 보인 단일가중치와 비교하면 그 성능 차이가 미미하거나 낮아서 대부분의 용어가중치 결합이 실질적으로 효과적이지 못하였다.

1. 서 론

데이터 결합이란 질의나 문헌에 대해 여러 가지 표현 방법 또는 여러 검색기법들을 결합하는 것이라 할 수 있다[1]. 이들 데이터 결합 연구의 배경은 기본적으로 정보검색의 복잡성과 불확실성으로 인해 특정 단일 표현이나 기법들로는 일부 적합문헌만을 효과적으로 검색할 수 있기 때문에 질의와 문헌간의 여러 적합 증거들을 결합하여 검색효율을 향상시키려는 시도에서 출발한다. 본 논문에서는 데이터 결합 영역에서 7개 컬렉션과 17개의 가중치 알고리즘을 대상으로 용어가중치 결합이 실질적으로 효율적인가를 성능 향상률 측면과 검색시스템의 효율성 측면에서 검증하고, 성능이 향상된 용어가중치 결합의 특징을 분석하였다.

2. 실험 설계

검색모형으로는 벡터공간모델을 채택하였고 용어가중치 결합 기법의 효율성을 검증하기 위해 다양한 가중치 알고리즘을 적용하였다. 벡터공간모델은 대표적으로 코사인 유사계수를 이용하지만 많은 연구자들이 내적 유사도(inner product similarity)나 다른 유사도 함수를 이용하여 왔다. 이는 유사계수에 따라 검색성능이 달라지기 때문에 주제영역이나 가중치에 따라 다양한 유사계수를 적용해 온 것이다. 본 논문에서는 계산복잡도와 효율성 측면에서 문헌 d_j와 질의 q_k사이의 유사도는 다음과 같이 내적유사도로 계산하였다.

$$Sim(d_j, q_k) = \sum_{i=1}^{n} (td_{ij} \times tq_{ik})$$

문헌 색인은 한글 실험집단의 경우 형태소 분석기 HAM으로 색인 하였고, 영문 실험집단은 HAM과 Porter 스테밍 알고리즘으로 색인 하였다. 질의어 처리도 문헌 색인 방식과 동일하게 자동으로 처리하였다. 문헌 d_j의 용어가중치 td_{ij}는 <표 1>과 같이 각 가중치 알고리즘 공식에 따라 계산하였으며, 질의어 가중치 tq_{ik}는 $\langle \log tf+1 \rangle$을 공통적으로 사용하였다. 문헌 d_j와 질의 q_k사이의 유사도는 두 벡터에서 일치되는 색인어의 가중치에 의존하기 때문에, 용어가중치 부여 기법은 벡터공간모델뿐만 아니라 대부분의 검색모델에서 검색효율에 영향을 미치는 중요한 요소이다.

성능평가는 고정된 11개의 재현율 0.0, 0.1, 0.2, … , 1.0에 대해 정확률을 산출하고 이에 대한 평균값으로 검색성능을 측정하는 11-포인트 평균정확률을 사용하였다. 용어가중치 결합시의 검색효율 평가는 각 가중치 기법을 각각 실행하여 검색된 문헌의 문서값들을 정규화시키고 결합함수를 이용하여 통합한뒤 상위 1000건에 대하여 11-포인트 평균정확률로 성능을 평가하였다. 그리고 용어가중치 결합의 성능 향상률은 결합된 2개의 단일가중치 중에서 높은 검색성능을 제공하는 것과 비교하였다.

본 논문에서 사용한 7개 실험집단은 한글 실험집단 2개(KT95, KRIST)와 영문 실험집단 5개(CACM, CISI, CRAN, LISA, MED)이다. 실험대상 가중치 용어가중치 알고리즘은 문헌 내 단어빈도와 역문헌빈도의 조합(*ntn, htn, atn, dtn, stn*) 수준과 코사인 정규화(*lnc, ntc, ltc, anc, atc*), 피벗 문헌길이 정규화(*dnb, dtu, Lnu, ltu*), *Okapi* 문헌길이 정규화(*otb, onb, otu*)에 따른 용어가중치 알고리즘으로 구분할 수 있다. 이들 용어가중치 알고리즘은 단어빈도, 역문헌빈도, 정규화의 세가지 요소들을 조합하여 만들어진 가중치 알고리즘들로서 실험에 사용한 용어가중치 공식은 <표 1>과 같다.

<표 1> 실험대상 용어가중치 알고리즘

(1) $tf \cdot \log\frac{N}{n}$	*ntn*	(2) $0.5+0.5\frac{tf}{\max tf}\times\log\frac{N}{n}$	*atn*
(3) $1+\log(1+\log tf)\cdot\frac{N+1}{n}$	*dtn*	(4) $\frac{1+\log(tf)}{1+\log(total\ tf)}\times\frac{\log(\frac{N}{n})}{\max idf}$	*stn*
(5) $\frac{\log(tf+1)\times(\log\frac{N}{n})}{\log(unique\ terms)}$	*htn*	(6) $\frac{\log tf+1.0}{\sqrt{\sum(\log tf+1.0)^2}}$	*lnc*
(7) $\frac{tf\cdot\log\frac{N}{n}}{\sqrt{\sum(tf\cdot\log\frac{N}{n})^2}}$	*ntc*	(8) $\frac{(\log tf+1.0)\times\log(\frac{N}{n})}{\sqrt{\sum[(\log tf+1.0)\times\log(\frac{N}{n})]^2}}$	*ltc*
(9) $\frac{0.5+0.5\frac{tf}{\max tf}}{\sqrt{\sum(0.5+0.5\frac{tf}{\max tf})^2}}$	*anc*	(10) $\frac{1+\log(1+\log tf)}{0.8+0.2\cdot\frac{dl(in\ byte)}{avdl(in\ byte)}}$	*dnl*
(11) $\frac{0.5+0.5\frac{tf}{\max tf}\times\log\frac{N}{n}}{\sqrt{\sum(0.5+0.5\frac{tf}{\max tf})^2\times(\log\frac{N}{n})^2}}$			*atc*
(12) $\frac{[1+\log(1+\log tf)]\times(\log\frac{N}{n}+1)}{(1.0-slope)\times pivot+slope\cdot unique\ terms}$			*dtu*
(13) $\frac{(1.0+\log tf)\cdot(\log\frac{N}{n})}{(1.0-slope)\times pivot+slope\cdot unique\ terms}$			*ltu*
(14) $\frac{\frac{1.0+\log tf}{1.0+\log(avgtf)}}{(1.0-slope)\times pivot+slope\cdot unique\ terms}$			*Lnu*
(15) $\frac{tf}{2\times(1-0.75+0.75\times\frac{byte}{average\ byte})+tf}$			*ont*
(16) $0.4+0.6\frac{tf}{tf+0.5+1.5\frac{dl(in\ unique\ terms)}{avdl(in\ unique\ terms)}}\times\frac{\log(\frac{N+0.5}{n})}{\log(N+1.0)}$			*otu*
(17) $\frac{tf}{2\times(1-0.75+0.75\times\frac{byte}{average\ byte})+tf}\times\frac{\log(\frac{N}{n})}{\max idf}$			*otb*

tf : 문헌 내 단어 *k*의 출현빈도, *total tf* : 문헌 내 단어들의 출현빈도 합, *avgtf* 문헌 내 단어들의 출현빈도 평균: *N* : 총문헌수, *n* : 단어 *k*가 출현하는 문헌수(*df*), *dl* : 문헌길이, *avdl* : 평균 문헌길이, *slope*: 0.2, *pivot*: *avdl*

서로 다른 용어가중치 기법에 의해 검색된 문헌들의 문서값 $SIM(d_j, q_k)$은 상이한 분포를 가지기 때문에 정규화 과정은 필수적으로 이루어져야 한다. 본 논문에서는 문서값 정규화 기법에 따른 검색성능을 분석하기 위해 다음과 같이 5가지 정규화 기법을 사용하였다.

(1) 단순정규화(max)

각 검색결과의 문서값을 최대값으로 정규화 한다.

$$SIM_{\max} = \frac{Individual\ Sim}{Max_{sim}}$$

(2) 사인정규화(sin)

사인함수 그래프의 특성을 이용하여 조정하는 것으로 단순정규화 방법에 비해 큰 문서값을 부여한다.

$$SIM_{\sin} = \sin(\frac{\pi}{2}\times SIM_{\max})$$

(3) 코사인정규화(cos)

코사인 함수 그래프 특성을 이용하여 조정하는 것으로 단순정규화 방법에 비해 작은 문서값을 부여한다.

$$SIM_{\cos} = 1-\cos(\frac{\pi}{2}\times SIM_{\max})$$

(4) 최대-최소정규화(mmx)

최대 문서값만을 이용하는 것이 아니라 최소값도 동시에 이용하여 문서값의 범위를 조정하는 정규화 방법이다.

$$SIM_{mmx} = \frac{Individual\ Sim - Min_{sim}}{Max_{sim} - Min_{sim}}$$

(5) 시그모이드정규화(*sig*)

단극시그모이드 함수로 각 문서값을 정규화 시킨다.

$$SIM_{sig} = \frac{1}{1+\exp(-\alpha(Individual\ Sim)+\beta)}$$
$$(\alpha=0.25,\ \beta=0)$$

정규화된 문서값을 어떻게 결합시키느냐에 따라 실제로 검색성능은 상당한 편차를 보일 수 있다. 따라서 검색결과를 결합하는 함수에 따라 용어가중치 결합이 검색성능에 어떤 영향을 미치는지를 분석하기 위해 본 논문에서는 합계(SUM), 최대값 선택(MAX), 최소값 선택(MIN)의 3가지의 문서값 결합함수를 사용하였다[2].

3. 실험결과 및 분석

본 논문에서 사용한 실험집단 7개에서 각 가중치의 범주별로 최상위 가중치 알고리즘들은 보정된 단어빈도와 역문헌빈도의 조합 수준이 3개(*atn, htn, stn*), 코사인 정규화 기법을 적용한 가중치 알고리즘이 2개(*atc, ltc*), 피벗 문헌길이 정규화 기법을 적용한 가중치 알고리즘이 2개(*ltu, dtu*), *Okapi* 문헌길이 정규화 기법을 적용한 가중치 알고리즘이 2개(*otb, otu*)로 총 9개 가중치로 나타났다. 최상위 가중치 알고리즘들 가운데 ltc 가중치는 코사인 정규화 기법을 사용한 가중치 알고리즘에서, *otb* 가중치는 *Okapi* 문헌길이 정규화 기법을 적용한 가중치 알고리즘에서 각각 가장 좋은 평균 성능을 보였다. 이들 *ltc, otb* 가중치 알고리즘들은 5개 실험집단(CACM, CISI, CRAN, LISA, MED)에서 가장 높은 검색효율을 보이는 안정적인 성능을 보였다. 피벗 문헌길이 정규화는 정규화 인자로 고유단어수를 이용한 가중치 알고리즘들이 좋은 성능을 보였으며, 문헌 내 단어빈도와 역문헌빈도 조합 수준에서는 단어출현빈도를 보정시키기 위해 문헌길이를 반영한 가중치 알고리즘들이 좋은 성능을 보이는 특징을 보였다.

<표 2>는 실험집단수와 향상률에 따라 성능이 향상된 용어가중치 결합의 수와 비율을 나타낸다. 4개 실험집단 이상에서 78개(57.35%), 5개 실험집단 이상에서 53개(38.97%), 6개 실험집단 이상에서 19개(13.97%), 7개 실험집단에서 2개(1.7%)의 가중치 결합쌍이 성능이 향상되었다. 그러나 실험집단 1개 이상에서 5% 이상 성능이 향상된 가중치 결합쌍은 46개였으나 4개 실험집단 이상에서 5% 이상 향상된 가중치 결합쌍은 9개, 6개 실험집단에서 5% 이상 성능이 향상된 경우는 anc-ntn 가중치 결합쌍 뿐이다. 이것은 대부분의 용어가중치 결합이 실험집단의 특성에 따른 영향을 많이 받고 있으며 성능 향상이 높다 하더라도 1-2개 실험집단에서만 효과적이라는 것을 의미한다.

<표 2> 검색성능이 향상된 용어가중치 결합의 수(비율)

향상률 / 실험집단수	0% 이상	5% 이상
4개 이상	78개(57.35%)	9개(6.62%)
5개 이상	53개(38.97%)	1개(0.74%)
6개 이상	19개(13.97%)	1개(0.74%)
7개 이상	2개(1.7%)	-

검색성능이 5% 이상 향상된 용어가중치 결합쌍을 가중치 유형별로 분석한 결과, 단순조합 가중치 알고리즘과 코사인 정규화 가중치 알고리즘들이 결합하였을 때 높은 성능 향상률을 보였다. 특히 *ntn-anc* 결합쌍은 검색성능이 5% 이상 향상된 실험집단 비율이 0.86(6개 실험집단)으로 가장 높았다. 그리고 단순조합인 *dtn* 가중치가

코사인 정규화 기법을 적용한 *lnc, ntc, atc, anc* 가중치나 피벗 코사인 정규화 기법을 적용한 *Lnu, dnb* 가중치, 그리고 *Okapi* 문헌길이 정규화 기법을 적용한 *onb* 가중치와 결합했을 때 실험집단 4개 이상에서 5% 이상 성능이 향상되었다. 따라서 높은 향상률을 보이는 가중치 결합쌍들은 단순조합 *dtn* 가중치와 다른 가중치 알고리즘들이 결합했을 때 실험집단에 크게 영향을 받지 않고 높은 검색성능을 보이고 있다는 것을 알 수 있다. 그러나 어떤 실험집단에서도 5% 이상의 성능 향상을 보이지 못하는 가중치 결합쌍이 많았다. 특히 단일가중치로 높은 평균 성능을 보였던 *htn* 가중치, 피벗 코사인 정규화 기법을 적용한 *dtu* 가중치나 *Okapi* 문헌길이 정규화 기법을 적용한 *otb* 가중치들은 다른 어떤 가중치와 결합해도 성능 향상이 5% 이상 되지 않았다. 또한 *atn*과 *htn, stn* 가중치 알고리즘 역시 다른 가중치 알고리즘과 결합되었을 때 대부분 5% 이상 향상되지 못했다. 그리고 특정 용어가중치가 다른 가중치와 결합될 때 검색성능 향상에 미치는 영향력은 코사인 정규화 기법을 적용한 가중치 알고리즘, 단어빈도와 역문헌빈도 조합의 가중치 알고리즘 순으로 나타났다.

용어가중치 결합이 검색성능에 미치는 영향을 분석한 결과 성능이 향상되는 결합들 대부분의 경우가 단일가중치 알고리즘들의 검색효율이 높지 않은 가중치 결합쌍이었다. 다시 말해서 원래 검색성능이 높지 않은 가중치 알고리즘이 다른 가중치 알고리즘과 결합하여 성능이 높게 향상되었다는 것인데, 문제는 가장 좋은 혹은 높은 검색효율을 보이는 단일가중치 보다 낮은 검색성능을 보이거나 거의 차이를 보이지 못하는 경우에는 데이터 결합 기법의 효율성에 의문을 제기해 볼 필요가 있다. 예를 들어 *anc* 가중치와 *ntn* 가중치를 결합시키는 경우 최대-최소정규화(mmx)로 문서값을 정규화시키고 합계함수로 통합할 때 검색성능이 6개 실험집단에서 5% 이상의 향상을 보였다. *anc*와 *ntn*의 용어가중치 결합시 KRIST(1.7%)을 제외하고 KT95(6.7%), CACM(16.8%), CISI(5.5%), CRAN(10%), LISA(17%), MED(5.1%) 실험집단들에서 5% 이상 성능이 향상되었으며, 3개 실험집단에서는 10% 이상 향상되었다. 그러나 *anc* 가중치와 *ntn* 가중치가 결합될 때 결합되는 두 단일가중치 중에서 높은 검색성능을 보이는 것과 비교하면 모든 실험집단에서 성능 향상이 되었으나 각 실험집단별로 가장 높은 검색성능을 보이는 단일가중치와 비교하면 모두 낮은 성능을 보였으며, CRAN과 MED 실험집단을 제외한 5개 실험집단에서는 10% 이상 낮은 검색성능을 보였다. 이것은 하나의 단적인 일례이지만 7개의 실험집단에서 검색성능이 향상된 대부분의 용어가중치 결합이 이런 현상을 나타냈다. 따라서 용어가중치 결합시 높은 성능 향상을 보이고 있다 할지라도 최적의 단일가중치 알고리즘과 비교해 검색성능이 낮거나 거의 차이가 없다면 그런 용어가중치 결합은 실질적으로 효율적이라 할 수 없다.

비록 각 용어가중치 결합이 성능 향상을 가져오더라도 상술한 바와 같이 최대의 성능을 보이는 단일가중치 보다 더 낮은 검색성능을 보인다면, 용어가중치 결합은 검색시스템의 효율성 측면에서 효과적이지 못하다고 할 수 있다. 따라서 <표 3>와 같이 각 실험집단별로 최대의 성능을 보이는 가중치 결합쌍을 분석하였다. <표 3>에서 알수 있듯이 단일 최적 가중치보다 가중치 결합의 검색성능이 5% 이상 차이를 보이는 실험집단은 KRIST뿐이며, 컬렉션간에 동일 가중치 결합이 하나도 없다. 이는 용어가중치 결합이 컬렉션과 가중치 알고리즘에 의존적이다는 것을 의미한다. 특히 실험집단 7개 전체에서 특정 용어가중치가 서로 다른 16개의 가중치와 결합될 때 최적의 단일가중치 보다 2% 이상 높은 성능을 보인 평균실험집단수를 분석하면 *atn*(0.75), *ntc*(0.44), *htn*(0.31), *otu*(0.31), *atc*(0.25), *otb*(0.25), *ltc*(0.19), *dtn*(0.13), ltu(0.13), *stn*(0.06), *dnb*(0.06), *onb*(0.06) 순으로 나타났다. 이를 통해 대부분의 가중치 알고리즘이 다른 가중치 알고리즘과 결합했을 때 최적의 단일가중치 보다 2% 이상의 높은 성능을 보이지 않는다는 것을 알 수 있다. 특히 단순조합인 *ntn*, 코사인 정규화 기법을 적용한 *anc*, 피벗 코사인 정규화인 *Lnu, dtu* 가중치들이 다른 가중치와 결합될 때는 모든 실험집단에서 최적의 단일가중치 보다 2% 이상 높은 성능을 보이지 않았다.

<표 3> 최대의 검색효율을 보이는 가중치 결합쌍

컬렉션	KT95	KRIST	CACM	CISI	CRAN	LISA	MED
가중치결합쌍	atn-dtn	atc-atn	ntc-otu	atn-ltc	otb-ltu	otb-dtu	htn-ltc
(11p-avg)	(0.4375)	(0.3503)	(0.4049)	(0.2447)	(0.4481)	(0.3997)	(0.5461)
성능 향상률	3.1%	5.0%	2.9%	4.1%	0.5%	-0.0	2.5%
문서값정규화	max	sig	mmx	sig	sig	sin	sig

여기서 한가지 주목할만한 실험결과는 문헌 내 단어빈도를 최대 단어빈도로 정규화한 *atn* 가중치가 코사인 정규화 기법을 적용한 가중치 알고리즘들과 결합하였을 때 5개 실험집단(KT95, KRIST, CACM, CISI, MED)에서 최적 단일가중치와 비교해 2% 이상의 향상률을 보였다. 이는 상이한 특성을 지니는 용어가중치 알고리즘들이 장단점을 보완하여 검색성능을 향상시킬 수 있다는 것을 의미한다.

한편, 실험집단 4개 이상에서 3% 이상의 성능 향상을 보인 용어가중치 결합의 개수를 결합함수별로 비교해 보면 합계함수(SUM)가 3% 이상에서 10개, 5% 이상에서 5개로 최소값 선택함수(MIN)나 최대값 선택함수(MAX) 보다 성능이 향상된 용어가중치 결합 개수가 많았으며, 최대값 선택함수(MAX)로 문서값을 결합할 경우 4개 실험집단 이상에서 성능이 향상된 용어가중치 결합은 하나도 없었다. 그리고 용어가중치 결합에서 문서값 정규화와 검색성능간의 상관관계를 분석한 결과 문서값 정규화 과정은 필수적이나 문서값 정규화 기법들간에는 현저한 성능 차이를 보이지 않았으며, 실험집단과 가중치 알고리즘의 특성이 검색성능에 더 많은 영향을 미쳤다.

4. 결 론

본 논문에서는 용어가중치 결합이 검색성능에 미치는 영향을 분석하기 위해 다양한 실험집단을 대상으로 여러 가중치 알고리즘들을 결합한 결과, 용어가중치 결합의 효율성은 실험집단과 가중치 특성에 의존적이었다. 그리고 SMART 시스템에서 사용하는 전통적인 가중치 기법들간의 결합은 높은 성능 향상률을 보였으나 단순조합 수준에서 문헌길이를 반영한 가중치 알고리즘, 피벗 코사인 정규화, 그리고 Okapi 문헌길이 정규화를 적용한 가중치 알고리즘들 보다 5% 이상 높은 검색성능을 보이지는 못했다. 비록 각 용어가중치 결합의 성능이 높게 나타날지라도 최적의 성능을 보인 단일가중치와 비교하면 그 검색성능의 차이가 미미하거나 낮아서 대부분의 용어가중치 결합이 실질적으로 효과적이지 못하였다. 이런 문제들을 해결하기 위한 향후 연구과제로는 컬렉션과 질의 특성에 따른 적응형 용어가중치 결합 연구가 필요할 것이다.

참고 문헌

[1] 이기호. 1999. "적합성 피드백 방법을 이용한 검색효율의 향상", 충남대학교 컴퓨터 공학과, 박사학위논문.

[2] Belkin, N. J., Colleen Cool, W. Bruce Croft, James P. Callan. 1993. "The effect of multiple query representations on information retrieval performance", Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339-346.

[3] Fox, E.A., and J.A. Shaw. 1993. "Combination of multiple searches", <http://trec.nist.gov/>

[4] Greengrass, Ed. 2000. "Information Retrieval: A survey", <http://www.csee.umbc.edu/cadip/readings/>