문헌 자동분류에서 용어가중치 기법에 대한 연구

Comparative Evaluation of Term Weighting Methods in Automatic Document Classification

이재윤, 최보영, 정영미 (연세대학교 대학원 문헌정보학과)

Jae-Yun Lee, Bo-Young Choi, Young-Mee Chung
Department of Library and Information Science, Yonsei University

정보검색 시스템의 성능을 향상시키기 위해서 다양한 용어가중치 공식이 제안되어왔다. 용어가중치는 질의와 문헌을 비교하는 검색의 경우 뿐만 아니라 문헌과 문헌을 비교하는 자동분류에서도 성능에 영향을 미칠 수가 있다. 본 논문에서는 다양한 용어가중치 공식에 대해서 살펴보고, 문헌 자동분류 성능에 미치는 영향을 문헌 클러스터링 실험과 범주화 실험을 통해 확인해 보았다.

1 서론

정보검색을 위해서 문헌을 표현하는 색인어에 적절한 가중치를 부여하는 문제는 오래동안 연구되어온 과제이다(Salton and Buckley 1988). 특히 최근에는 본문검색이 활발해지면서 가중치 공식의 중요성이 더욱 강조되어, 좋은 검색 성능을 보이는 공식이 몇 가지 보고되어 있다.

한편, 문헌 자동분류 시스템에서도 문헌과 문헌을 비교하기 위해서 색인어, 즉 분류자질에 적절한 가중치를 부여할 필요가 있다. 그런데 전체 색인어 중에서 질의어와 부합되는 일부에 부여한 가중치만 활용하는 검색 시스템과는 달리, 자동분류 시스템에서는 거의 모든 색인어에 부여한 가중치를 활용한다는 특징이 있다. 따라서 검색시스템에서 좋은 성능을 보이는 가중치 기법이 반드시 분류시스템에서도 좋은 성능을 보인다고 보장할 수는 없다.

본 연구에서는 검색 환경에서 제안된 기존의 가중치 공식을 검토하여, 이들이 자동분류 환경에서 어떤 성능을 보이는지를 살펴보고자 하였다.

2 용어가중치 공식

용어가중치(TW)를 구성하는 요소는 단어빈도(TF), 역문헌빈도(IDF), 문헌길이정규화(DL)의 세 가지이다. 이중에서 역문헌빈도에 대해서도 몇 가지 공식이 제안되어 있지만, 본 연구에서는 다음의 공식을 역문헌빈도로 사용하였다.

$$IDF = \frac{N}{df}$$

실험에서는 역문헌빈도를 제외한 두 공식의 적용을 두 자리 기호로 표시하였다. 본 연구에서 사용한 단어빈도 가중치 공식과 문헌길이정규화 공식을 아래에서 살펴본다.

2.1 단어빈도 가중치

단어빈도는 문헌 내 출현여부만을 반영하는 이진값이나 출현빈도 자체를 가중치로 사용할 수도 있으며, 이외에 다양한 공식이 제안되어 있다. <표 1>에 본 연구에서 적용한 단어빈도 가중치 공식을 정리하였다.

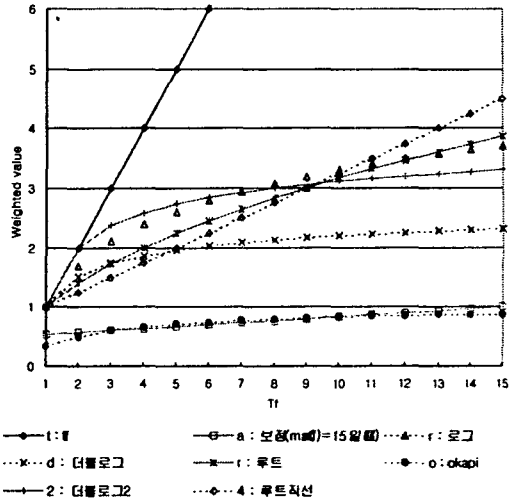
(1)이진 TF : 출현한 경우는 모두 1로 지정한다.

<표 1> 단어빈도 가중치 공식

이름	공식	기호
이진	$TF = 1(\text{if } tf>0),\ 0$	b
단순	$TF = tf$	t
로그	$TF = 1+\log(tf)$	l
더블로그	$TF = 1+\log(1+\log(tf))$	d
루트	$TF = \sqrt{tf}$	r
보정	$TF = (1-w)+w\times\frac{tf}{max\_tf}$	a
Okapi	$TF = \frac{tf}{2+tf}$	o
더블로그2	$TF = 1+\log_2(1+\log_2(tf))$	2
루트직선	$TF = \frac{tf+3}{4}$	4

(2)단순 TF : 단어빈도 tf를 그대로 사용한다.
(3)로그 TF : tf가 1인 단어의 지나치게 낮은 영향력을 보충하고, tf가 높은 단어의 지나친 영향력을 낮추기 위해 TREC-1에서 SMART 팀이 제안한 공식이다.
(4)더블로그 TF : 질의가 소수의 질의어로 구성된 경우에는 로그 TF로도 tf가 높은 단어의 지나친 영향력을 낮추는 것이 불충분하다고 판단하여 Singhal, et al. (1998)은 TREC-7에서 로그를 두 번 취하는 공식을 제안하였다.
(5)루트 TF : 로그 TF와 마찬가지 효과를 가지지만 tf가 높은 경우의 의미를 로그 TF보다는 덜 축소하는 공식이다.
(6)보정(augmented) TF : 가중치를 일정 범위로 한정시켜서 최소 빈도의 단어라도 일정 값 이상이 되도록 하면서 동시에 최대 값도 제한하는 공식이다. SMART 시스템에서는 w값을 0.5로, INQUERY 시스템에서는 w값을 0.6으로 사용하고 있다.
(7)Okapi TF : 2-포아송 모형을 적용하는 Okapi 시스템에서 사용하는 공식이다.
(8)더블로그2 TF : tf가 0, 1, 2일때는 가중치도 0, 1, 2가 되고 tf가 3 이상일 때에는 가중치가 로그곡선을 따르도록 공식을 고안하였다.
(9)루트직선 TF : 예비 실험에서 좋은 성능을 보인 루트 TF 공식과 가중치 선의 기울기가 유사하도록, tf가 1일때와 9일때의 값이 루트 TF와 같은 직선공식을 고안하였다.

이상의 여러 가지 단어빈도 가중치 공식은 결국 <그림 1>과 같이 tf가 낮은 경우와 높은 경우 사이의 가중치 차이를 어떻게 줄 것인가



<그림 1> tf에 따른 각 공식값의 변화

를 판단하는 문제이다. 가중치 선의 기울기는 이진 TF가 0으로 가장 낮고 단순 TF가 1로 가장 높다. 다른 공식들은 모두 이진 TF와 단순 TF 사이에서 적절한 기울기를 선택한 결과로 볼 수 있다.

2.2 문헌길이 정규화

전문 검색 시스템에서는 길이가 긴 문헌일수록 각 단어의 출현빈도가 높고 출현하는 단어의 종류가 많다는 두 가지 원인 때문에, 짧은 문헌에 비해서 검색될 확률이 높다는 문제가 발생한다. 이는 클러스터링에서도 마찬가지여서 길이가 긴 문헌일수록 다른 문헌과의 유사도가 상대적으로 높게 될 여지가 있으므로 정규화를 해야 한다. <표 2>에 본 연구에서 적용한 문헌길이정규화 공식을 정리하였다.
(1)코싸인 정규화 : SMART 시스템에서 사용되는 방법으로서 TF와 IDF의 조합인 TF·IDF 값인 w를 해당 문헌 내 모든 단어의 w값의 제곱의 합의 제곱근으로 나눈다.
(2)최대 tf 정규화 : 보정 TF 공식을 사용하는 경우이다. 이 방식은 긴 문헌에서 단어의 출현빈도가 높다는 문제는 어느 정도 해결하지만, 출현단어의 종류가 많다는 문제는 해결하지 못한다.
(3)바이트길이 정규화 : Okapi 시스템에서 채택한 기법으로서 Okapi의 TF 공식과 결합해서

<표 2> 문헌길이 정규화 공식 * 정규화하지 않은 경우의 기호는 n

명칭	공식	기호	조합
코사인	$TW = \frac{w}{\sqrt{\sum w_i^2}}$	c	tc
최대tf	$TW = (1-w) + w \times \frac{tf}{max\_tf}$	a	ta(=an)
바이트길이	$TW = \frac{tf}{2 \times (1-b+b \times \frac{document\ length}{average\ document\ length}) + tf}$	o	to
피벗유니크	$TW = \frac{\frac{1+\log(tf)}{1+\log(average\ tf)}}{0.8+0.2 \times \frac{unique\ tf}{average\ unique\ tf}}$	u	uu
피벗바이트길이	$TW = \frac{tf}{0.8+0.2 \times \frac{length\ of\ document\ (in\ bytes)}{average\ length\ of\ document\ (in\ bytes)}}$	b	tb lb db
로그	$TW = \frac{1+\log(tf)}{1+\log(total\ tf)}$	l	ll
단어수	$TW = \frac{tf}{0.5+1.5 \times \frac{total\ tf}{average\ total\ tf} + tf}$	w	tw
피벗단어수	$TW = \frac{tf}{0.8+0.2 \times \frac{total\ tf}{average\ total\ tf}}$	x	tx lx dx
루트	$TW = \sqrt{\frac{tf}{\sum tf}}$	r	rr

쓰인다. 이 공식에서 b는 보통 0.75가 되며 문헌의 길이는 byte로 나타낸다.

(4)피벗유니크 정규화 : 기존 정규화 기법의 단점을 보완하면서 긴 문헌이 검색될 확률이 높은 두 가지 원인을 모두 해소하도록 고안된 공식이다(Singhal, et al. 1996).

(5)피벗바이트길이 정규화 : Singhal, et al. (1998)이 복합어구를 색인할 경우에는 색인어의 길이가 다양하기 때문에 색인어의 수를 이용하는 기존의 피벗 유니크 정규화가 불리할 수 있다고 보고 상대적으로 색인어의 길이에 영향을 덜 받는 바이트길이 정규화를 채택하여 약간 변형한 공식이다.

(6)로그 정규화 : 각 단어의 tf를 문헌 내 색인어의 총 빈도(tf의 합)로 나누어서 상대적인 빈도를 구하는 공식이다. 로그를 써서 지나친 차이를 누그러뜨리게 하였다.

(7)단어 수 정규화 : 바이트 길이 정규화 공식을 사용하되 문헌의 길이를 byte가 아닌 색인어의 총 빈도로 구하였다.

(8)피벗 단어 수 정규화 : 피벗 바이트 길이 정규화 공식을 사용하되 문헌의 길이를 byte가 아닌 색인 단어의 총 빈도로 구하였다.

(9)루트 정규화 : 로그 정규화와 같은 형태지만 루트 대신 로그를 썼다.

3 자동분류 실험

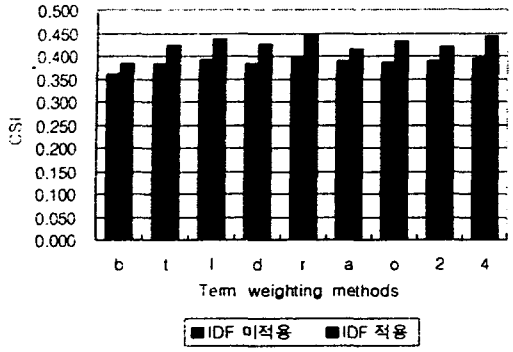
3.1 문헌 클러스터링 실험

클러스터링 실험에 사용한 실험집단은 연세대학교 문헌정보학과에서 구축한 KFCM-CL 1,020건이다. KFCM-CL은 경제 및 국제분야 신문기사 모음인 KFCM 중에서 일부를 추출하여 분류 정보를 추가한 것이다.

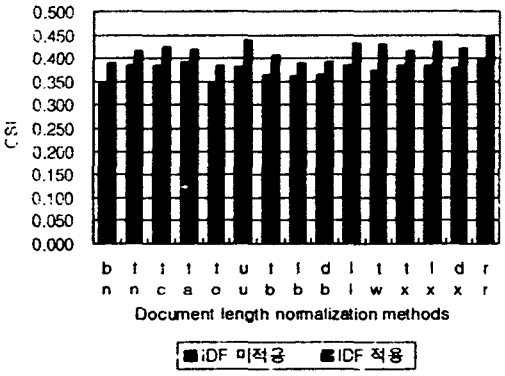
클러스터링 기법으로는 계층적 기법 중에서 완전연결을 채택하였고 유사계수로는 단어빈도가중치 실험에서는 코싸인 계수를, 문헌길이정규화 실험에서는 자카드 계수를 사용하였다.

계층적 클러스터링에서는 결합이 진행됨에 따라서 클러스터의 조직이 계속 변하기 때문에 유사도 상위 문헌쌍 1,000개에서부터 1,000쌍씩 추가하면서 50,000쌍까지 50단계에서의 성능을 클러스터 일치계수 CSIM(정영미 1999)으로 판정하여 평균을 구하여 평가하였다.

단어빈도가중치 공식과 문헌길이정규화 공식을 달리하며 실험한 결과를 각각 <그림 2>와 <그림 3>에 나타내었다.



<그림 2> 단어빈도가중치 공식별 클러스터링 성능



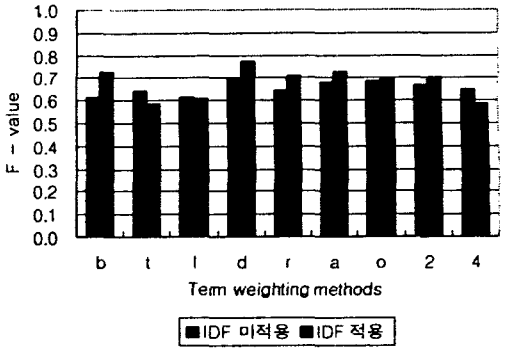
<그림 3> 문헌길이정규화 공식별 클러스터링 성능

3.2 문헌 범주화 실험

문헌범주화 실험에 사용한 실험집단은 KTSET 중 정보과학회 논문 880건이었고, 이중 50건을 검증집단으로, 나머지를 학습집단으로 사용하였다. 범주화 기법으로는 KNN에서 K를 50으로, 유사계수는 코싸인 계수, 최적범주 결정은 범주별 유사도의 합을 기준으로 하였다. KNN분류기의 성능은 범주별 분류정확률과 분류재현율을 평균한 것의 F값을 구하여 판정하였다.

4 결론

전체적으로 IDF를 적용하면 성능이 향상되었지만, 범주화에서는 그렇지 못한 경우도 나타났다. 그리고 단어빈도가중치 실험에서 다른 공식에 비해 중간 이하의 성능을 보이는 b, d, a 공식이 범주화 실험에서는 IDF를 곱했을 때



<그림 4> 단어빈도가중치 공식별 KNN 분류 성능

1, 2, 3위의 성능을 보여서 클러스터링과 범주화에서 좋은 성능을 보이는 공식이 상이한 것으로 나타났다. <그림 1>을 보면 이 세 공식은 모두 기울기가 낮은 공식임을 알 수가 있다.

문헌길이정규화 공식의 경우에는 바이트길이를 이용한 공식 o, b보다 단어수를 이용한 공식 w, x가 클러스터링에서 좋은 성능을 보였다. IDF를 곱한 경우 1~4위의 조합 공식은 rr, uu, lx, ll인데 이들은 모두 단어빈도에 로그나 루트를 쓴 공식이어서 문헌길이정규화 공식을 구성하는 단어빈도 공식의 중요성을 확인할 수 있었다. 또한 기존의 공식 대신 제안한 루트정규화 공식이 IDF의 적용 여부에 상관없이 가장 좋은 클러스터링 성능을 보였다.

이와 같은 분류실험 결과 기존의 정보검색 시스템에서와는 다소 다른 결과를 얻은 것에 대해서 향후 이론적인 심층분석과 함께 다양한 실험집합에 대한 검증이 필요할 것이다.

참고문헌

정영미. 1999. 지식 자동분류를 위한 유사성 척도의 평가. 제2회 디지털도서관 컨퍼런스 논문집 (데이터베이스진흥센터), 87-97.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5): 513-523.

Singhal, A., et al. 1996. Pivoted document length normalization. Proceedings of the SIGIR'96: 21-29.

Singhal, A., et al. 1998. AT&T at TREC-7. Proceedings of the Seventh Text REtrieval Conference (TREC 7): 239-251.