# 피벗 역문헌빈도 가중치 기법에 대한 연구

## A Study on the Pivoted Inverse Document Frequency Weighting Method

이 재 윤(Jae-Yun Lee)*

초 록

역문헌빈도 가중치 기법은 문헌 집단에서 출현빈도가 낮을수록 색인어의 중요도가 높다는 가정에 근거하고 있다. 그런데 이는 중간빈도어를 중요하게 여기는 여타 이론과는 일치하지 않는 것이다. 이 연구에서는 저빈도어보다 중간빈도어가 더 중요하다는 가정에 근거하여 역문헌빈도 가중치 공식을 수정한 피벗 역문헌빈도 가중치 기법을 제안하였다. 제안된 기법을 검증하기 위해서 세 실험집단을 대상으로 검색실험을 수행한 결과, 피벗 역문헌빈도 가중치 기법이 역문헌빈도 가중치 기법에 비해서 특히 검색결과 상위에서의 성능을 향상시키는 것으로 나타났다.

## ABSTRACT

The Inverse Document Frequency (IDF) weighting method is based on the hypothesis that in the document collection the lower the frequency of a term is, the more important the term is as a subject word. This well-known hypothesis is, however, somewhat questionable because some low frequency terms turn out to be insufficient subject words. This study suggests the pivoted IDF weighting method for better retrieval effectiveness, on the assumption that medium frequency terms are more important than low frequency terms. We thoroughly evaluated this method on three test collections and it showed performance improvements especially at high ranks.

키워드 : 역문헌빈도, 정보검색, 용어가중치, Information Retrieval, Inverse Document Frequency, Term Weights

* 연세대학교 문헌정보학과 강사(memexlee@lis.yonsei.ac.kr)

■ 논문 접수일 : 2003. 11. 25
■ 게재 확정일 : 2003. 12. 4

# 1 서 론

역문헌빈도 가중치에 대한 연구는 초기인 1970년대를 제외하면 최근에는 그다지 많지 않다. 가중치 공식의 종류도 용어가중치를 구성하는 세 요소인 용어빈도 가중치, 역문헌빈도 가중치, 문헌길이 정규화 중에서는 가장 적은 편이라고 할 수 있다.

그중에서도 역문헌빈도 가중치 IDF는 적합성 기반 검색에 있어서 가장 성공적인 파라미터로 평가된다는 지적이 있을 만큼(Roelleke 2003) 정보검색 분야에서 오랫동안 변화 없이 광범위하게 사용되어왔다.

역문헌빈도 가중치는 출현빈도가 낮을수록 색인어의 중요도가 높다는 전제에 근거한 공식이다. 그런데, 이와 같은 가정이 반드시 직관과 일치하는 것은 아니다. 물론 문헌빈도가 1,000인 용어와 1인 용어를 비교해보면 1인 용어가 중요하다고 생각할 수 있다. 그러나 검색대상 문헌들 중에서 딱 한 문헌만 골라주는 용어와, 열 개 내지 스무 개를 골라주는 용어 중에서 어느 것이 검색에 더 도움이 되는가를 생각해보면 반드시 빈도가 낮다고 더 중요한 것은 아니라고 할 수 있다.

자동색인의 효시를 이룬 Luhn(1958)의 연구에서는 어느 수준 이상의 고빈도어나 지나친 저빈도어는 주제어로서 의미가 없으므로 문헌식별력이 높은 중간빈도어를 색인어로 선정해야 한다고 주장하였다. 또한 문헌분리가 이론에서도 중간빈도어를 중요한 색인어로 간주하고 있다(Salton, Yang, & Yu 1975).

이와 같이 역문헌빈도 가중치 공식은 어떤 면에서 직관과 불일치하며, 다른 이론과 상충하는 면도 있다. 따라서 저빈도어보다 어느 정도 수준의 중간빈도어의 가중치를 더 높게 산출하도록 전통적인 역문헌빈도 가중치 공식을 수정하는 것이 검색성능의 개선에 도움이 될 수 있을 것이다.

이 연구에서는 이런 점을 고려하여 중간빈도어의 가중치가 높아지도록 역문헌빈도 가중치 공식을 간단하게 수정하는 방안을 연구하였다. 그리고 검색성능의 비교실험을 통해 제안하는 방식의 성능을 검증해보았다.

# 2 역문헌빈도 가중치에 대한 선행 연구

Sparck Jones(1972)는 Zipf의 법칙을 고려하여 용어의 가중치가 문헌빈도에 반비례하도록 해야 한다면서 문헌빈도가 n인 용어에 대해서 다음과 같은 지수함수 형태의 공식을 제시하였다.

$$f(n) = m, \ where \ 2^{m-1} < n \leq 2^m$$

(m은 조건을 만족하는 정수)

그리고 문헌집단에서는 다음과 같이 전체 문헌 수 N을 고려해야 한다고 하였다.

$$f(N)-f(n)+1$$

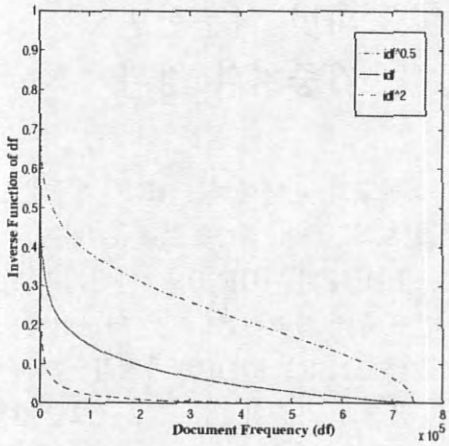
이에 대해서 Robertson(1972)이 Sparck Jones의 제안을 로그함수로 해석하고 지나친 고빈도어는 가중치가 0이 되도록 1을 더한 항을 제거하면서 현재 널리 사용되는 IDF 공식이 탄생하였다.

$IDF=\log_2\frac{N}{df}$ , 여기서 $df$는 문헌빈도

몇 년 지나지 않아 이 두 사람은 확률검색모형(Robertson & Sparck Jones 1976)을 제안하였는데, 이때의 가중치 공식에서 적합성 정보를 나타내는 항을 제거한 아래의 공식이 2-Poisson 검색모형에서 현재 사용되고 있다.

$$w=\log\frac{N-df+0.5}{df+0.5}$$

한편 Singhal(1997)은 역문헌빈도 가중치 차이의 반영 정도에 대해서 분석하면서 역문헌빈도 가중치를 그대로 쓰지 않고 특정 값으로 제곱해서 쓰는 방안을 실험하였다. 예를 들어 〈그림 1〉에서와 같이 0.5 제곱을 하면 역문헌빈도 가중치의 차이를 덜 반영하고 제곱을 하면 더 반영하는 셈이 된다.




〈그림 1〉 역문헌빈도 제곱 함수의 가중치 곡선
(Singhal 1997, 60에서 인용)

Singhal은 0.75제곱, 1.5제곱, 2.0제곱 등의 여러 경우를 실험하였는데, 최종적으로 가중치 함수의 기울기가 조금 더 가파르도록 1.5제곱을 취한 아래의 공식이 가장 좋은 성능을 보였다고 보고하였다.

$$idf^{1.5}=(\log\frac{N}{df})^{1.5}$$

다만 이 방식은 검색결과 중에서 재현율이 높은 지점, 즉 낮은 순위의 정확률을 뚜렷하게 향상시키는 반면에, 재현율이 낮은 검색결과 상위에서의 정확률은 기존의 방식과 비슷하거나 오히려 저하되는 문제가 있는 것으로 나타났다.

이 연구에서는 이후 전통적인 역문헌빈도를 IDF, 확률검색모형에서의 역문헌빈도를 IDF-P, Singhal의 IDF 1.5제곱을 IDF-S로 표기하였다.

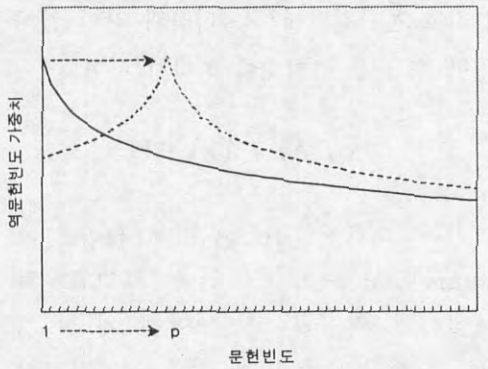
## 3 피벗 역문헌빈도 가중치의 정의

문헌빈도와 반비례하는 IDF 가중치는 문헌빈도가 1일 때 최고값을 가진다. 이는 IDF-P나 IDF-S도 마찬가지다. 이 연구에서 제안하는 피벗 역문헌빈도(Pivoted IDF; PIDF) 가중치 공식은 IDF 공식의 가중치 최고점을 〈그림 2〉와 같이 문헌빈도가 p인 지점(중심점; pivot point)으로 이동시키도록 간단하게 수정한 것이다. 제안한 가중치 공식 PIDF는 아래와 같다.

$$PIDF = \log_2 \frac{N}{|df - p| + 1}$$

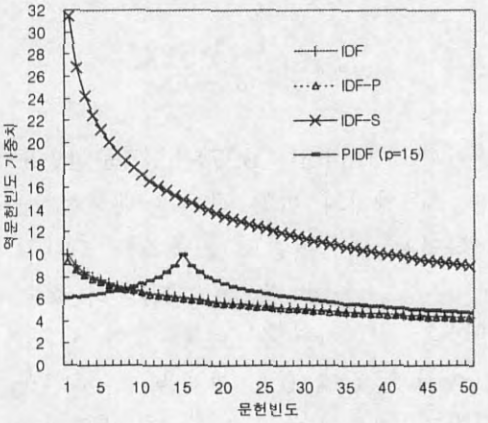
여기서 N은 전체문헌 수, df는 출현한 문헌 수, p는 상수이다. df가 p와 같을 때 분모가 가장 작아지도록 하기 위해서 문헌빈도인 df와 상수 p와의 차이에 절대값을 취했으며, 1을 더함으로써 분모가 0이 되지 않도록 하였다. 따라서 주어진 용어의 문헌빈도가 p와 같으면 차이가 가장 작아서 최고 가중치를 가지고, 문헌빈도가 p보다 크거나 작으면 차이가 커져서 가중치는 낮아지게 된다.

PIDF 공식을 적용할 때에는 문헌빈도가 p인 용어의 가중치가 IDF 공식에서의 문헌빈도 1인 용어의 가중치와 같게 된다. 중심점 p를 기준으로 문헌빈도가




〈그림 2〉 PIDF 공식에 의한 역문헌빈도 최고 지점의 이동

커지거나 작아지면 가중치도 낮아지므로 문헌빈도가 1인 용어의 PIDF 가중치는 문헌빈도가 2p-1인 용어의 가중치와 같게 된다. PIDF 공식에서 p가 1일 때에는 IDF 가중치 공식이 된다. 전체 문헌 수 N=1,000이고 p를 15로 할 때의 PIDF 가중치를 다른 역문헌빈도 공식의 가중치와 함께 나타내면 〈그림 3〉과 같다.




〈그림 3〉 문헌빈도에 따른 각 공식별 가중치 (N=1,000일 때)

〈그림 3〉에서 볼 수 있듯이 IDF-P는 IDF와 거의 같은 값을 가지게 되며, IDF-S는 훨씬 큰 값이면서 저빈도어의 가중치가 높아지는 정도가 IDF에 비해서 더 심하다.

# 4 성능 비교 실험

## 4.1 실험 환경 및 설계

문헌빈도 가중치 공식으로서 PIDF를 사용한 검색 성능을 다른 공식과 비교하여 검증하는 실험을 수행하였다. 이를 위해서 우선 다양한 p값에 따른 PIDF가중치의 성능을 IDF 가중치의 성능과 비교해본 다음, IDF-P 및 IDF-S 공식과도 비교하였다.

실험집단으로는 〈표 1〉과 같이 규모와 주제, 언어가 다양한 3개 실험집단을 이용하였다. 이중에서 CACM 실험집단은 원래 질의문이 64건이지만 적합문헌이 1건 이하이거나 서지사항을 질의로 사용한 경우 등을 제외하고 45건의 질의만을 채택하였다. 그리고 HANTEC-S 실험집단은 원래 12만 건으로 구성된 HANTEC 2.0 실험집단(김지영 외 2000) 중에서 과학기술 분야만을 대상으로 한 것이다.

검색 시스템은 벡터공간모형에 근거해서 Windows XP를 운영체제로 하는 PC 상에서 Visual FoxPro 소프트웨어를 사용하여 구현하였다.

실험은 역문헌빈도 가중치를 문헌 색인어에 부여하는 경우와 질의어에 부여하는 경우로 나누어서 진행하였다. 기존의 역문헌빈도 가중치에 대한 연구는 두 경우 중 한 쪽을 대상으로 수행된 것이 대부분이다. 검색 성능 면에서는 IDF 가중치의 경우에 질의어에 적용하는 것이 더 좋다고 알려져 있다(Buckley, Allan, & Salton 1993). Singhal (1997)의 IDF-S도 질의어 가중치에 적용하는 상황에서 연구된 것이다.

역문헌빈도 가중치를 색인어에 부여하는 경우에 색인어의 가중치 $DW(t_i)$와 질의어의 가중치 $QW(t_i)$는 다음의 공식으로 산출하였다.

$$DW(t_i) = (1 + \log tf(t_i)) \times \text{역문헌빈도가중치}$$

$$QW(t_i) = 1 + \log tf(t_i)$$

〈표 1〉 실험집단의 특성

| | Medline | CACM | HANTEC-S |
|---|---|---|---|
| 언 어 | 영 어 | 영 어 | 한 국 어 |
| 분 야 | 의 학 | 전산학 | 과학기술 전반 |
| 성 격 | 초 록 | 초 록 | 초록 또는 전문 |
| 문헌 수 | 1,033 | 3,204 | 39,838 |
| 질의문 수 | 30 | 45 | 30 |

질의어에 역문헌빈도 가중치를 부여하는 경우의 DW($t_i$)와 QW($t_i$)를 구하는 공식은 다음과 같다.

$$DW(t_i) = 1 + \log tf(t_i)$$

$$QW(t_i) = 1 + \log tf(t_i) \times \text{역문헌빈도가중치}$$

문헌 D와 질의 Q간의 유사도 sim(D,Q)는 다음과 같이 코사인 계수로 산출하였다.

$$sim(D,Q) = \frac{\sum DW(t_i)QW(t_i)}{\sqrt{\sum DW(t_i)^2 \times \sum QW(t_i)^2}}$$

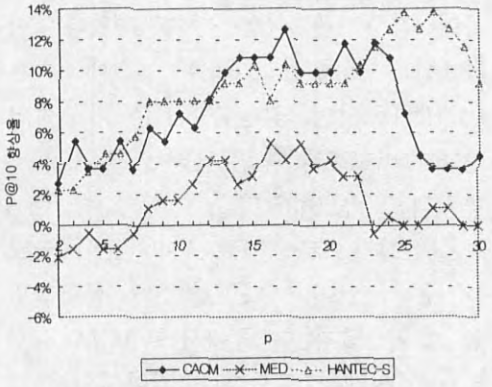
검색 결과의 평가척도로는 10위내 정확률(P@10), 30위내 정확률(P@30), 10위내 순위정확률(RP@10), 30위내 순위정확률(RP@30), 11지점 평균 정확률(11-AP)의 다섯 가지를 적용하였다. 이와 같이 다양한 평가척도를 적용한 이유는 역문헌빈도 가중치 공식의 차이가 검색성능에 미치는 영향을 다양한 각도에서 살피려고 했기 때문이다.

P@10과 P@30은 각각 검색결과의 10위와 30위 이내에 적합문헌의 비율이 얼마인가를 알려준다. RP@10과 RP@30은 검색결과의 10위와 30위 이내 문헌의 순위가 어느 정도로 적절한가를 알려준다. 10위까지 살펴보는 두 척도는 재현율이 낮은 최상위 순위에서의 성능을, 30위까지 살펴보는 두 척도는 중상위 순위에서의 성능을 반영한다. 반면에 11지점 평균 정확률은 전체적인 순위의 적절함을 반영한다.

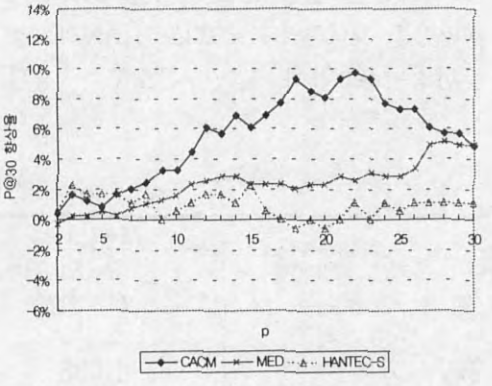
## 4.2 색인어에 가중치를 부여하는 경우

### 4.2.1 p값의 변화에 따른 PIDF 가중치 공식의 성능

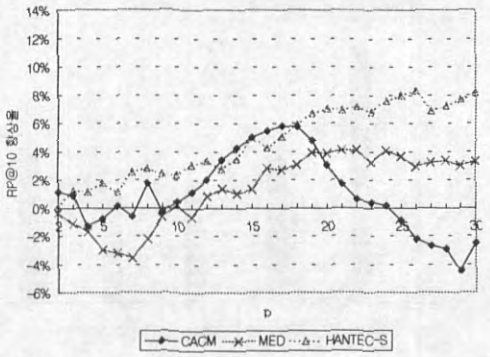
세 실험집단을 대상으로 문헌 색인어에 PIDF 가중치를 부여하되 p를 2에서 30까지 증가시키면서 검색한 경우의 성능이 기존의 IDF 가중치를 부여한 경우의 성능(PIDF 방식에서 p=1일 때)에 비해서 달라진 비율을 각 척도별로 〈그림 4〉에 제시하였다.



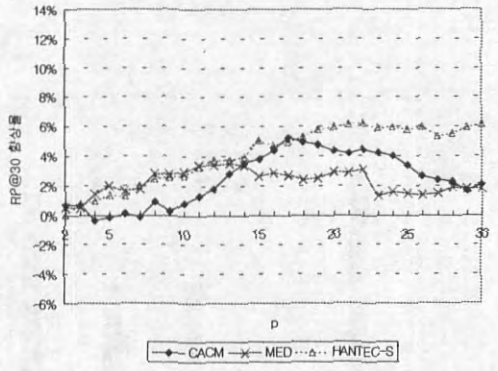

(a) 10위내 정확률의 향상율



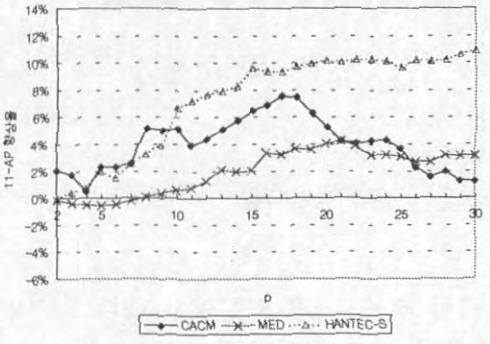

(b) 30위내 정확률의 향상율




(c) 10위내 순위정확률의 향상율




(d) 30위내 순위정확률의 향상율




(e) 11지점 평균 정확률의 향상율

〈그림 4〉 p값에 따른 PIDF 가중치 공식의 검색 성능 (색인어에 적용)

실험 결과는 실험집단에 따라서 다소 차이가 있는 것을 볼 수가 있다. CACM 실험집단에 대해서는 p가 10에서 20사이인 경우에 모든 평가척도에서 성능이 향상되었다. 또한 30위내 정확률을 제외하면 p가 17일 때 가장 좋은 성능을 보여서 10위내 정확률은 IDF 대비 12.61%, 11지점 평균정확률은 IDF 대비 7.56% 향상되었다. p가 25보다 커지게 되면 성능이 급격히 저하되어 10위내 순위정확률 척도에서는 오히려 IDF보다 성능이 떨어졌다. 이는 p값을 지나치게 높이면 비록 전반적인 성능은 향상되더라도 최상위 10위 이내의 문헌 순위는 오히려 부적절해짐을 의미한다.

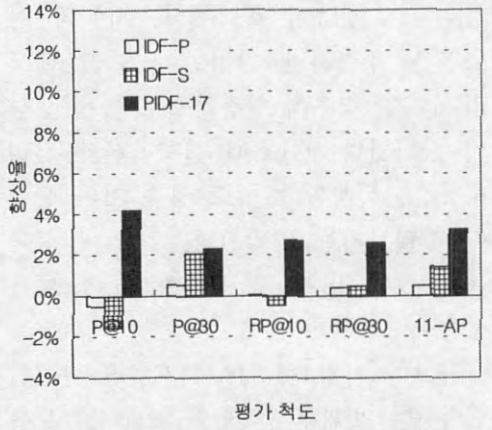
Medline 실험집단에 대해서는 성능 향상율이 CACM 실험집단의 경우에 미치지 못했으나, 역시 p가 10 이상인 경우에는 성능이 좋아졌다. 성능의 최고점은 뚜렷하지 않지만 p가 15에서 20사이인 경우에 안정적으로 높은 성능이 나타났다. CACM 실험집단과 달리 p가 25 이상일 때에도 성능 저하 현상이 뚜렷하지는 않았지만, 10위내 정확률은 p가 23일 때 IDF보다 더 낮은 성능을 보였다.

규모가 가장 큰 HANTEC-S 실험집단에서는 30위내 정확률을 제외하고 모든 척도에서 p를 20 정도까지 높일 수록 성능도 꾸준히 향상되었으며, 이보다 더 p를 크게 하더라도 30까지는 더 높은 성능이 유지되는 것으로 나타났다.

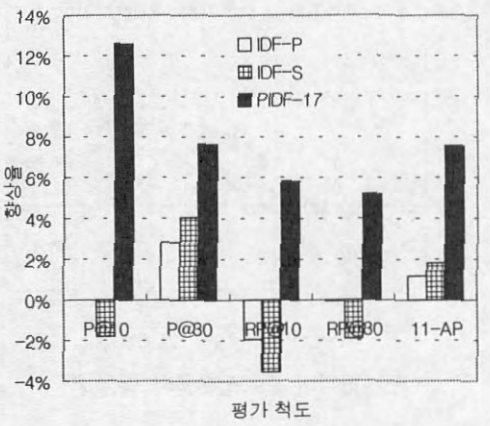
세 실험집단에서 공통적으로 살펴볼 수 있는 것은 다음과 같다. 첫째, p가 10에서 25 사이인 경우에 PIDF가 IDF에 비해서 향상된 성능을 보인다. 둘째, p가 15에서 20 사이이면 IDF 대비 성능 향상율이 최고에 가깝다. 셋째, 기존의 성능평가 척도로 가장 많이 사용되는 11지점 평균정확률과 10위내 정확률 면에서는 확실한 성능 향상 효과를 얻을 수 있다.

### 4.2.2 역문헌빈도 공식의 성능 비교

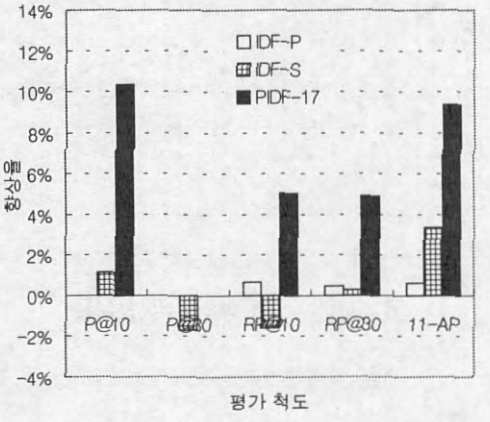
IDF 가중치 공식 대비 성능을 기준으로 앞의 분석 결과에 근거하여 p가 17일 때의 PIDF-17 공식과 IDF-P, IDF-S 공식의 성능을 〈그림 5〉의 (a)~(c)에서 비교하였다.




(a) Medline




(b) CACM



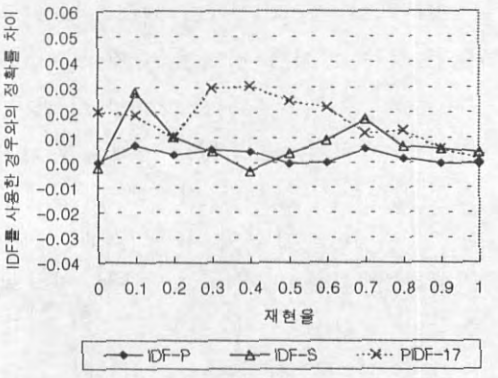

(c) HANTEC-S

〈그림 5〉 IDF 공식의 성능 대비 역문헌빈도 공식별 향상율 비교 (색인어에 적용)

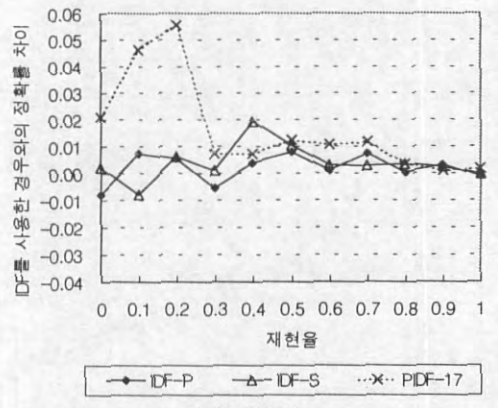
비교 결과 세 실험집단 모두에서 PIDF-17이 평가척도를 불문하고 다른 가중치 공식을 월등하게 앞서는 것으로 나타났다. 반면에 IDF-P와 IDF-S는 P@10과 RP@10 척도에서 IDF와 비슷하거나 낮은 성능을 보였다. 이는 이 두 공

식이 검색결과 순위의 최상부에서는 IDF 공식보다 성능을 향상시키지 못함을 시사한다.

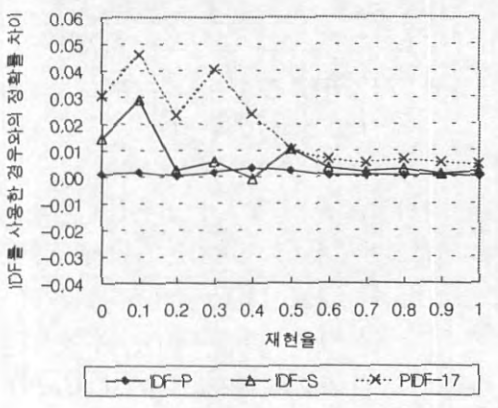
이를 확인하기 위해서 재현율의 변화에 따른 정확률의 변화를 나타내는 재현율-정확률 곡선을 그려보았으나 IDF와 IDF-P의 경우처럼 성능 차이가 크지 않은 경우에 식별하기가 곤란하였다. 이에 따라 재현율 11 지점의 정확률을 비교하되 각 공식의 정확률에서 IDF 공식의 정확률을 뺀 값을 〈그림 6〉에 표시하였다. 〈그림 6〉을 보면 검색결과의 상위를 나타내는 낮은 재현율 지점에서 PIDF-17과 IDF의 정확률 차이가 뚜렷하게 나타난다. 구체적으로 PIDF-17의 정확률이 IDF의 정확률에 비해서 2% 포인트 이상 향상된 경우는 Medline 집단에서 재현율 0.6 이하일 때(0.2 제외), CACM 집단에서 재현율 0.2 이하일 때, HANTEC-S 집단에서 재현율 0.4 이하일 때이다. 그러나 IDF-S 공식의 경우에는 재현율이 0.5 이하일 때 IDF보다 낮은 정확률을 보이는 경우가 일부 나타난다(Medline 집단에서 재현율 0, 0.4 지점, CACM 집단에서 재현율 0.1 지점, HANTEC-S 집단에서 재현율 0.4 지점). IDF-P 공식과 IDF 공식의 차이는 거의 없이 0% 포인트 전후로 나타난다.




(a) Medline



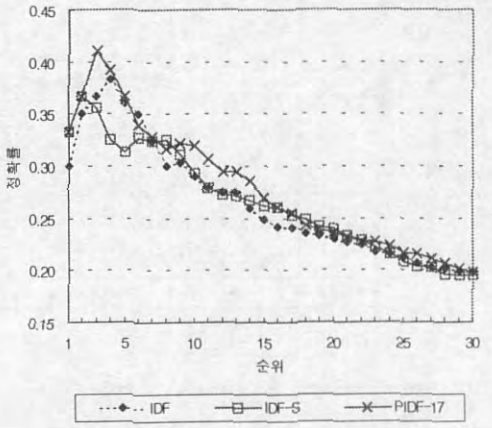

(b) CACM




(c) HANTEC-S

〈그림 6〉 IDF 공식의 성능 대비 11지점 정확률 비교 (색인어에 적용)

앞에서와 같이 11지점 재현율을 기준으로 성능 차이를 살펴보는 것 대신 실제 검색 결과의 순위를 기준으로 성능 차이를 비교할 수도 있다. "재현율 0.2 지점"이라는 표현보다 이를테면 "검색결과 10위"라는 표현이 훨씬 직관적으로 인식하기 쉽다는 점에 이 분석 방식의 장점이 있다. 세 실험집단 중 가장 규모가 큰 HANTEC-S에서 1위부터 30위까지 각 순위마다의 정확률을 〈그림 7〉에 제시하였다.




〈그림 7〉 HANTEC-S에서 각 순위별 정확률 비교 (색인어에 적용)

〈그림 7〉을 보면 6~8위인 지점을 제외하고 15위 정도까지는 PDIF-17의 성능이 뚜렷하게 다른 공식보다 앞서는 것을 볼 수 있다. 반면에 IDF-S 공식은 5위를 전후한 지점에서 IDF 보다 정확률이 현저히 낮게 나타난다. 따라서 색인어 가중치에 적용할 때 PIDF-17은 IDF나 IDF-S에 비해서 검색결과 10위 내지 15위 이내에 적합문헌을 평균적으로 더 많이 포함시킨다고 말할 수 있다. 반면에 IDF-S 공식은 검색결과 5내지 6위 이내에 적합문헌을 IDF 공식보다 평균적으로 적게 포함시킨다고 말할 수 있다.

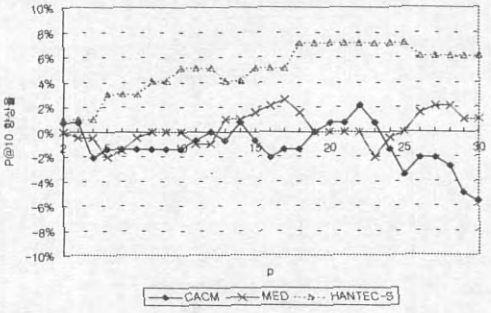
### 4.3 질의어에 가중치를 부여하는 경우

역문헌빈도 가중치를 색인어가 아닌 질의어에 부여하는 것은 모든 용어가 아닌 질의에 나타난 용어에만 적용하는 셈이다. 만약 문헌간 유사도를 내적유사도 공식으로 계산하고 문헌길이 정규화를 적용하지 않는다면 역문헌빈도 가중치를 문헌과 질의 중 어느 한 쪽에 적용하는 것은 동일한 결과를 낳는다. 어차피 유사도를 계산할 때 문헌과 질의간 일치하는 용어의 가중치를 곱해서 합산하기 때문이다. 그러나 코사인 유사계수를 사용하거나 다른 문헌길이 정규화 방법을 적용한다면 역문헌빈도 가중치를 어느 쪽에 적용하는가에 따라서 결과가 달라진다.

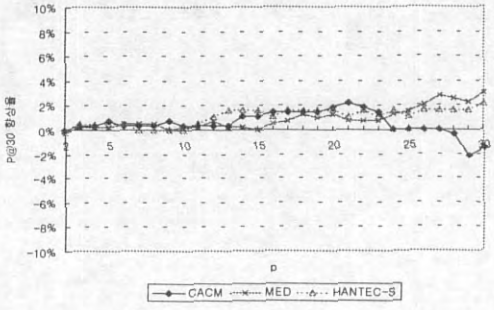
Singhal(1997)의 IDF-S 공식도 질의어에 가중치를 부여해서 성능 향상을 검증한 경우이므로, 앞의 색인어 가중치 실험에서 좋지 못한 성능을 보였다고 해서 폄하할 수는 없다. 먼저 언급한 바와 같이 검색 성능은 일반적으로 역문헌빈도 가중치를 질의어에 적용하는 경우가 더 높으므로 질의어에 대한 실험이 더 의미가 크다고 볼 수 있다.

#### 4.3.1 p값의 변화에 따른 PIDF 가중치 공식의 성능

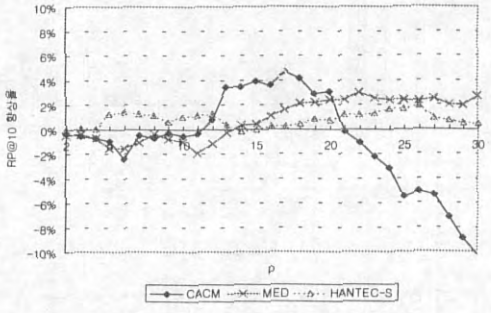
색인어에 대한 실험과 마찬가지로 질의어에 PIDF 가중치를 부여한 경우의 성능이 IDF 가중치를 부여한 경우의 성능에 비해서 달라진 비율을 각 평가 척도별로 〈그림 8〉에 제시하였다.



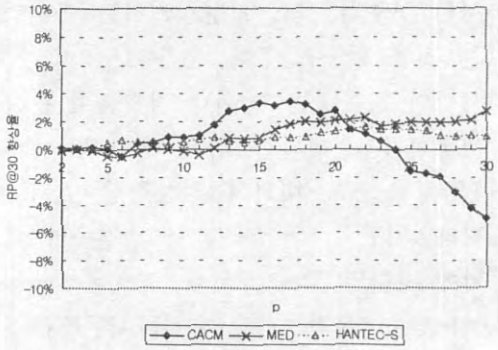

(a) 10위내 정확률의 향상율



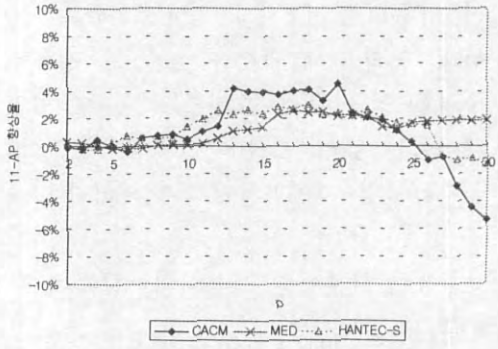

(b) 30위내 정확률의 향상율




(c) 10위내 순위정확률의 향상율




(d) 30위내 순위정확률의 향상율




(e) 11지점 평균 정확률의 향상율

〈그림 8〉 p값에 따른 PIDF 가중치 공식의 검색 성능 (질의어에 적용)

색인어에 대한 경우와 달리 질의어에 적용한 경우에는 IDF 대비 성능향상 효과가 크지 않게 나타났다. p가 10에서 25 사이인 경우에 대부분의 척도에서 성능이 향상되긴 하지만 대개 2% 내외에 그쳤으며, 최상위 순위의 성능을 나타내는 10위내 정확률 면에서는 CACM 집단의 경우에 약간 저하되는 현상도 나타났다. 따라서 질의어에 문헌빈도를 부여할 경우에는 PIDF 가중치가 IDF 가중치에 비해 상위 순위에

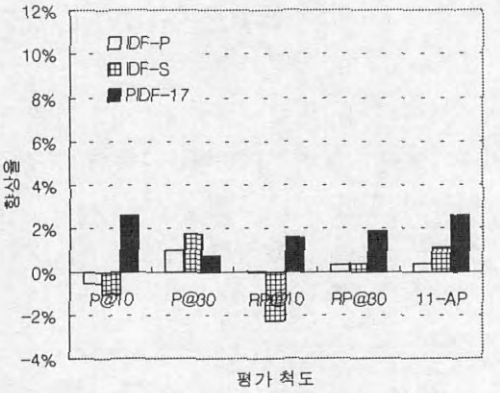
서의 성능을 크게 향상시키지는 못한다는 것을 알 수 있다. p가 15에서 20 사이인 경우에는 10위내 정확률이나 순위정확률을 제외한 다른 평가척도에서 1%에서 6% 내외의 성능향상 효과가 나타났지만, 색인어에 적용한 경우의 차이에 비하면 크지 않았다.

다만 규모가 가장 큰 HANTEC-S 실험집단의 경우에 성능평가 척도로 널리 사용되는 11지점 평균정확률과 10위내 정확률 면에서는 2~7% 정도로 뚜렷한 성능 향상 효과를 얻을 수 있었다. 이는 앞서 색인어 가중치로 적용했을 때의 경우와 마찬가지 결과이다.

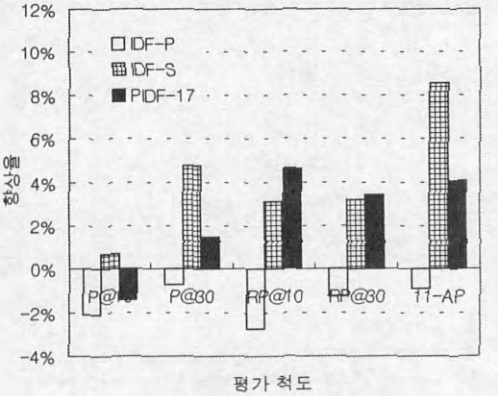
### 4.3.2 역문헌빈도 공식의 성능 비교

질의어에 대해서 IDF 가중치 공식을 적용했을 때의 성능을 기준으로 p가 17일 때의 PIDF-17 공식과 IDF-P, IDF-S 공식의 성능을 〈그림 9〉에서 비교하였다.

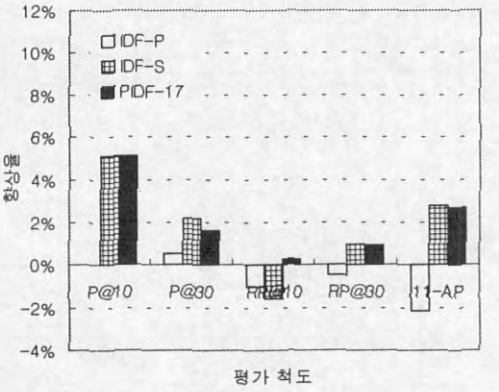
전반적으로 Medline 집단에서는 PIDF-17 공식의 성능이, CACM과 HANTEC-S에서는 IDF-S 공식의 성능이 가장 좋았다. 그러나 최상위 순위에서의 성능을 반영하는 10위내 정확률이나 10위내 순위정확률은 CACM 집단에서 10위내 정확률을 제외하면 PIDF-17의 성능이 IDF-S와 비슷하거나 더 좋은 것으로 나타났다. IDF-S 공식의 최상위 순위에서의 성능은 Medline 집단이나 HANTEC-S 집단에서는 IDF 공식과 비슷하거나 오히려 낮게 나타났다.




(a) Medline



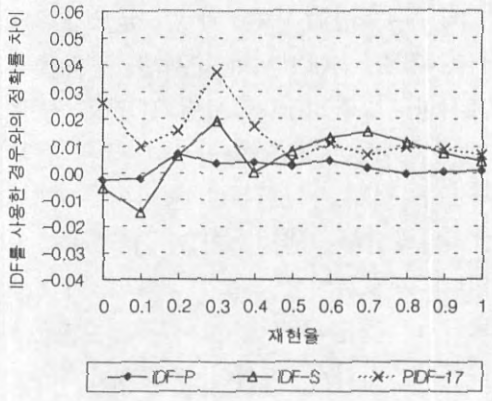

(b) CACM



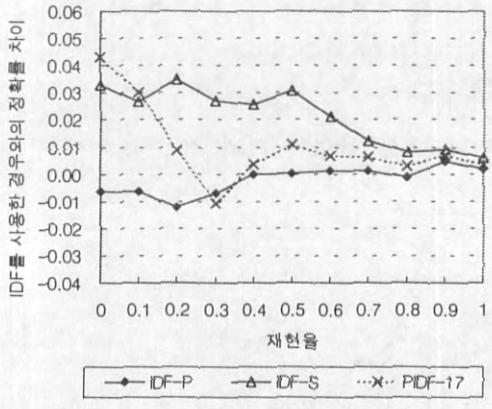

(c) HANTEC-S

〈그림 9〉 IDF 공식의 성능 대비 역문헌빈도 공식별 향상율 비교 (질의어에 적용)

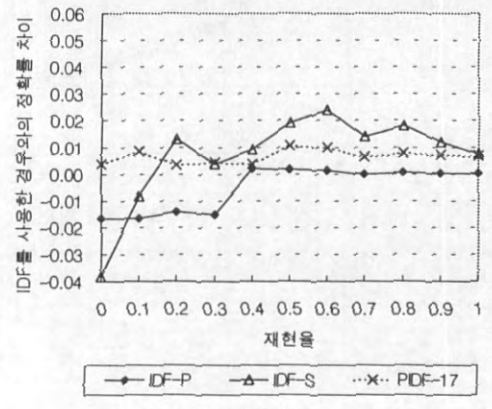
색인어 가중치로 적용했을 때의 〈그림 6〉과 마찬가지로 재현율 11 지점의 각 공식의 정확률에서 IDF 공식의 정확률을 뺀 값을 〈그림 10〉에 표시하였다. 이를 보면 검색결과의 상위를 나타내는 낮은 재현율 지점에서 IDF-S 공식은 IDF 공식보다 낮은 정확률을 보이는 경우가 나타난다(Medline 집단에서 재현율 0.2 이하일 때, HANTEC-S 집단에서 재현율 0.1 이하일 때). 반면에 PIDF-17 공식은 Medline 집단에서 재현율 0.4 이하일 때, CACM 집단에서 재현율 0.1 이하일 때, HANTEC-S 집단에서 재현율 0.1 이하일 때 각각 가장 높은 정확률을 보였다.




(a) Medline



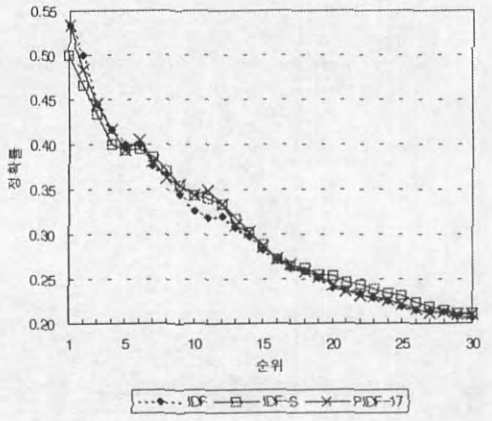

(b) CACM




(c) HANTEC-S

〈그림 10〉 IDF 공식의 성능 대비 11지점 정확률 비교 (질의어에 적용)

세 실험집단 중 가장 규모가 큰 HANTEC-S에서 각 순위별 정확률을 표시한 〈그림 11〉을 보면 IDF-S 공식은 6위까지는 IDF 공식보다 낮은 정확률을 보이다가 7위부터 높아지는 것으로 나타난다. 반면 PIDF-17 공식은 20위 정도까지는 IDF 공식과 비슷하거나 더 높은 성능을 보이는 것으로 나타난다. 따라서 질의어 가중치에 적용할 때 PIDF-17은 IDF나 IDF-S에 비해서 검색결과 10위 내지 20위 이내에 적합문헌을 평균적으로 더 많이 포함시킨다고 말할 수 있다. 이는 색인어 가중치에 대한 실험과 거의 같은 결과이다.




〈그림 11〉 HANTEC-S에서 각 순위별 정확률 비교 (질의어에 적용)

## 5 결 론

전통적인 역문헌빈도 가중치 기법은 저빈도어일 수록 높은 가중치를 부여하는 방식이다. 그러나 문헌분리가를 비롯한 다른 이론을 고려하면 저빈도어보다 중간빈도어의 가중치를 높게 하는 방안을 검토해볼 여지가 있었다. 문헌빈도가 1인 용어에 가장 높은 가중치를 부여하는 역문헌빈도 가중치 공식을 수정하여 중간빈도어(문헌빈도 p인 용어)의 가중치를 최고로 하는 피벗 역문헌빈도 (PIDF) 가중치 공식을 제안하였다. 제안된 PIDF 방식은 문헌빈도 중심점 p를 파라미터로 하여 역문헌빈도 가중치 공식의 분모가 문헌빈도와 p의 차이가 되도록 한 것이다. 따라서 주어진 용어의 문헌빈도가 p와 같으면 차이가 가장 작아서 최고 가중치를 가지고, 문헌빈도가 p보다 크거나 작으면 차이가 커져서 가중치는 낮아지게 된다.

제안된 피벗 역문헌빈도 가중치를 이용한 검색실험 결과 색인어에 적용하는 경우에는 p를 10에서 25 사이로 설정했을 때 다른 역문헌빈도 가중치 공식에 비해서 우월한 성능을 얻었다. 그러나 질의어 가중치로 사용할 경우에는 성능 향상 효과가 크지 않았다. 다만 p를 17로 설정한 PIDF-17 공식을 이용하면 검색결과 최상위 문헌들의 순위는 여러 실험집단에 걸쳐서 일관되게 개선되는 것으로 나타났다. 이런 차이는 색인어 가중치로 사용할 때에는 역문헌빈도 가중치를 모든 용어에 적용하지만, 질의어 가중치로 사용할 때에는 그렇지 않기 때문으로 짐작된다.

역문헌빈도 가중치의 반영 정도를 높인 Singhal의 IDF 1.5 제곱 방식은 높은 검색순위에서의 성능 향상에 실패했었고, 색인어 가중치로 사용할 경우에는 성능 향상 효과도 미미하게 나타났다. 이와 달리 피벗 역문헌빈도 가중치는 전반적인 성능과 함께 검색결과 상위에서의 성능도 향상시킬 수 있었으며, 색인어 가중치로 사용할 경우에 더 큰 성능 향상 효과를 얻을 수 있었다.

향후에는 피벗 역문헌빈도 가중치를 이 연구에서 사용한 검색모델 이외에 다른 모델에 대해 적용해봐야 할 것이다. 벡터공간 검색모형 중에서도 문헌길이 정규화를 코사인 정규화가 아닌 다른 방식, 예를 들면 피벗 유니크 정규화와 같은 방식을 사용하는 경우에의 성능 차이도 살펴봐야 한다. 또한 벡터공간 검색모형이 아닌 확률검색 모형이나 연관검색, 피드백 검색 등에 적용해보는 것도 중요한 과제이다.

## 참 고 문 헌


김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완. 『제12회 한글 및 한국어 정보처리 학술대회 논문집』, 210-215.

Buckley, C. J. Allan, and G. Salton. 1993. "Automatic routing and ad-hoc retrieval using SMART: TREC 2." *Proceedings of the Second Text REtrieval Conference (TREC 2)*: 45-55.

Luhn, H. P. 1958. "The automatic creation of literature abstracts." *IBM Journal of Research and Development*, 2(2): 159-165. 재인용: 정영미, 『정보검색론』. (개정판) 서울: 구미무역(주) 출판부, 1993.

Robertson, S. E. 1972. "Term specificity." *Journal of Documentation*, 28(2): 164.

Robertson, S. E., and Karen Sparck Jones. 1976. "Relevance weighting of search terms." *Journal of the American Society for Information Science*, 27, 129-146.

Roelleke, T. 2003. "A frequency-based and a Poisson-based definition of the probability of being informative." *Proceedings of the 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 227-234.

Salton, G., C. S. Yang, and C. T. Yu. 1975. "A theory of term importance in automatic

text analysis." *Journal of the American Society for Information Science*, 26(1): 33-44.

Singhal, Amit. 1997. *Term Weighting Revisited*. Ph.D. Thesis, Department of Computer Science, Cornell University.

Sparck Jones, Karen. 1972. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation*, 28(1): 11-21.