역문헌빈도 가중치의 재검토

Inverse Document Frequency Weighting Revisited

이재윤, 연세대학교 문헌정보학과 강사

Jae-Yun Lee, Yonsei University

역문헌빈도 가중치는 문헌 집단에서 출현빈도가 낮을수록 색인어의 중요도가 높다는 가정에 근거하고 있다. 이 연구에서는 역문헌빈도 가중치의 가정에 의문을 제기하고, 이를 보완하는 새로운 문헌빈도 가중치 공식을 제안하였다. 제안한 가중치 공식은 저빈도어가 아닌 중간빈도어가 더 중요하다는 가정에 근거한 것으로서 역시 문헌빈도를 이용한 함수이다. 문헌빈도에 의한 가중치를 문헌의 색인어에 부여하는 경우와 질의어에 부여하는 경우로 나누어서 실험을 수행하고, 두 경우의 차이점을 논하였다.

1 서 론

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

이중에서도 역문헌빈도 가중치 공식 IDF는 적합성 기반 검색에 있어서 가장 성공적인 파라미터로 평가된다(Roelleke 2003)는 지적이 있을 만큼 정보검색 분야에서 오랫동안 변화 없이 광범위하게 사용되어 왔다. IDF에 대한 최근 연구는 확률이론이나 정보이론에 근거해서 재해석하는 시도가 있을 뿐이다.

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 검색 대상 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. 실제로 문헌분리가 이론에서는 저빈도어나 고빈도어가 아닌 중간빈도어를 중요한 색인어로 간주하고 있다.

이와 같이 역문헌빈도 가중치 공식은 어떤 면에서 직관과 불일치하며, 다른 이론과 상충하는 면도 있다. 이 연구에서는 이런 점을 고려하여 중간빈도어의 가중치가 높아지도록 역문헌빈도 가중치 공식을 간단하게 수정한 형태를 제시하고 실험을 수행

하였다.

2 문헌빈도 가중치

Sparck Jones(1972)는 Zipf의 법칙을 고려하여 용어의 가중치가 문헌빈도에 반비례하도록 해야 한다면서 지수함수의 형태를 제시하였다. 이에 대해서 Robertson(1972)이 Sparck Jones의 제안을 로그함수로 해석한 결과로 현재 널리 사용되는 IDF 공식이 탄생하였다.

$$IDF = \log_2 \frac{N}{df}$$

이후 몇 년 지나지 않아 확률검색모형(Robertson and Sparck Jones 1976)이 제안되었는데, 이때의 가중치 공식에서 적합성 정보를 나타내는 항인 R과 r을 제거한 아래의 공식이 2-Poisson 검색모형에서 현재 사용되고 있다.

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

한편 Singhal(1997)은 역문헌빈도 가중치에 대한 분석 결과, 가중치 함수의 기울기가 더 가파르도록 IDF 가중치의 1.5제곱을 취한 아래의 공식이 더 바람직하다고 주장하였다.

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

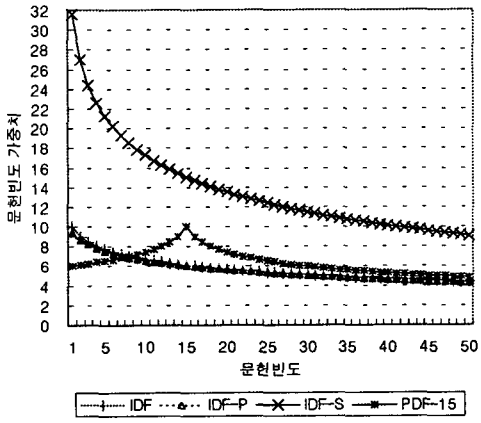
이 연구에서는 이후 전통적인 역문헌빈도를 IDF, 확률검색모형에서의 IDF를 IDF-P, Singhal의 IDF 1.5제곱을 IDF-S로 표기하였다.

3 역문헌빈도 가중치 공식의 수정

문헌빈도와 반비례하는 IDF 가중치는 문헌빈도가 1일 때 최고 값을 가진다. 이는 IDF-P나 IDF-S도 마찬가지다.

이 연구에서 제안하는 문헌빈도 가중치 공식은 IDF 공식의 가중치 최고점을 문헌빈도가 p인 지점으로 이동시키도록 간단하게 고안하였다. 제안한 가중치 공식 PDF는 아래와 같다.

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



〈그림 1〉 문헌빈도에 따른 각 공식별 가중치

PDF 공식을 적용할 때에는 문헌빈도가 p인 용어의 가중치가 IDF에서의 문헌빈도

1인 용어의 가중치와 같게 된다. 문헌빈도가 1인 용어의 PDF 가중치는 문헌빈도가 2p-1인 용어의 가중치와 같다. p가 15일 때의 가중치를 다른 문헌빈도 가중치 공식과 함께 나타내면 〈그림 1〉과 같다.

〈그림 1〉에서 볼 수 있듯이 IDF-P는 IDF와 거의 같은 값을 가지게 되며, IDF-S는 훨씬 큰 값이면서 저빈도어의 가중치가 높아지는 정도가 IDF에 비해서 더 강하다.

4 검색 실험

4.1 실험 설계

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

실험문헌집단으로는 문헌 수 1,033건인 Medline 실험집단과 문헌 수 3,204건인 CACM 실험집단을 사용하였다. CACM 실험집단은 원래 질의문이 64건이지만 적합문헌이 1건 이하이거나 서지사항을 질의로 사용한 경우 등을 제외하고 45건의 질의만을 채택하였다.

검색 실험은 문헌빈도 가중치를 문헌 색인어에 부여하는 경우와 질의어에 부여하는 경우로 나누어서 진행하였다. 용어빈도 가중치는 로그 TF 공식을 사용하였고 문헌간 유사도는 코사인 유사계수로 비교하였다.

검색 결과의 평가척도로는 10위내 정확률(P@10), 30위내 정확률(P@30), 10위내 순위정확률(RP@10), 30위내 순위정확률(RP@30), 11지점 평균 정확률, 3지점 평균 정확률의 여섯 가지를 적용하였다. 이와 같이 다양한 평가척도를 적용한 이유는 문헌빈도 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 정확률과 3지점 평균 정확률은 전체적인 순위의 적절함을 반영한다. 그 중에서도 11지점 평균 정확률은 최상과 최하 순위에서의 성능까지 감안하지만, 3지점 평균 정확률은 상대적으로 중간 순위(여기서는 재현율 0.2, 0.5, 0.8의 세 지점)의 성능만 반영하는 점이 다르다.

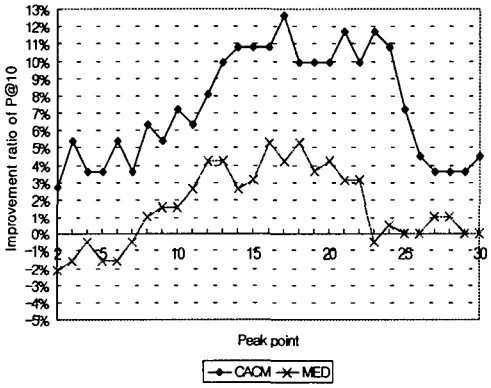
4.2 문헌 색인어에 가중치를 부여하는 경우

1) p값의 변화에 따른 PDF 가중치 공식의 성능 문헌 색인어에 PDF 가중치를 부여한 경우의 성능이 IDF 가중치를 부여

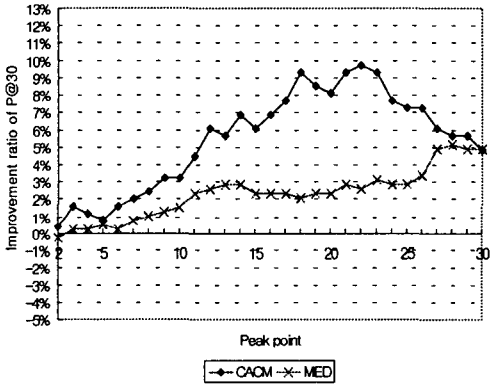
한 경우의 성능에 비해서 달라진 비율을 각 척도별로 〈그림 2〉~〈그림 7〉에 제시하였다.

실험 결과는 실험집단에 따라서 다소 차이가 있는 것을 볼 수가 있다. CACM 실험집단에 대해서는 p가 10에서 20사이인 경우에 모든 평가척도에서 성능이 향상되었다. 또한 P@30을 제외하면 p가 17일 때 가장 좋은 성능을 보여서 P@10은 IDF 대비 12.61%, 3지점 평균 정확률은 IDF 대비 12.59% 향상되었다. p가 25보다 커지게 되면 성능이 급격히 저하되어 RP@10 척도에서는 오히려 IDF보다 성능이 떨어졌다. 이는 전반적인 성능은 향상되더라도 최상위 10위 이내의 문헌순위는 오히려 부적절해짐을 의미한다.

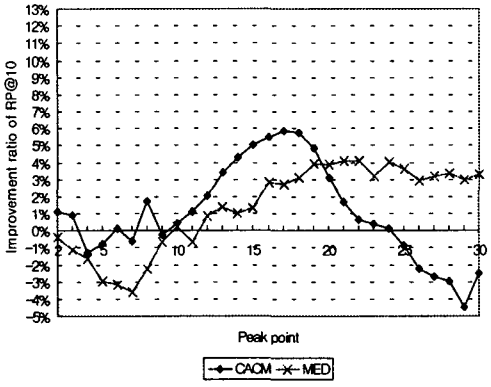
한편 Medline 실험집단에 대해서는 성능 향상율이 CACM 실험집단의 경우에 미치지 못했으나, 역시 p가 10 이상인 경우에는 성능이 좋아졌다. 성능의 최고점은 뚜렷하지 않지만 p가 15에서 20사이인 경우에 안정적으로 높은 성능이 나타났다. CACM



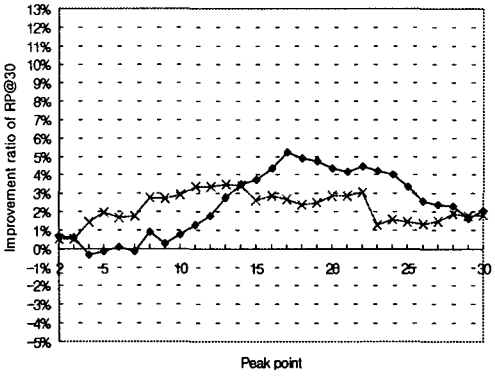
〈그림 2〉 색인어 가중치에 PDF 적용 - P@10



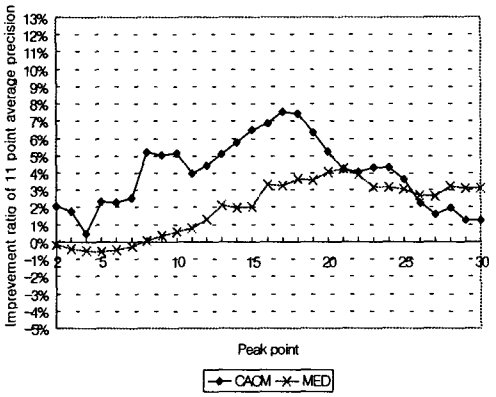
〈그림 3〉 색인어 가중치에 PDF 적용 - P@30



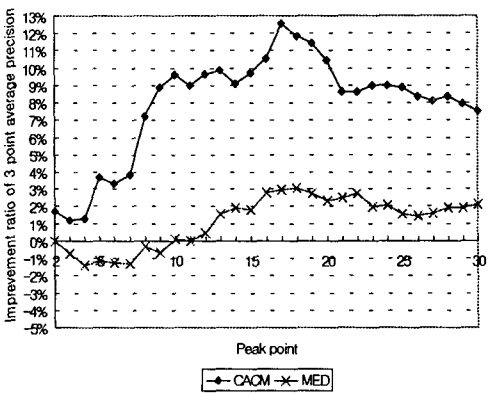
〈그림 4〉 색인어 가중치에 PDF 적용 - RP@10



〈그림 5〉 색인어 가중치에 PDF 적용 - RP@30



〈그림 6〉 색인어 가중치에 PDF 적용 - 11지점 평균 정확률



〈그림 7〉 색인어 가중치에 PDF 적용 - 3지점 평균 정확률

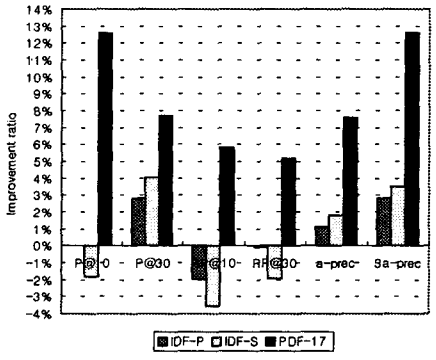
실험집단과 달리 p가 25 이상일 때에도 성능 저하 현상이 뚜렷하지는 않았지만, P@10 척도에서는 *p*가 23일 때 IDF보다 나쁜 성능을 보였다.

두 실험집단에서 공통적으로 살펴볼 수 있는 것은 다음과 같다. 첫째, *p*가 10 이하이거나 25 이상이면 IDF와 성능 차이가 없거나 오히려 나쁘다. 둘째, *p*가 15에서 20 사이이면 IDF 대비 성능 향상율이 최고에 가깝다.

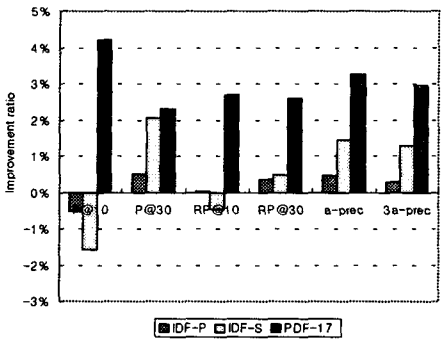
2) 문헌빈도 공식의 성능 비교

IDF 공식 대비 성능을 기준으로 *p*가 17일 때의 PDF-17 공식과 IDF-P, IDF-S 공식을 비교해서 〈그림 8〉과 〈그림 9〉에 제시하였다.

비교 결과 CACM과 Medline 실험집단에서 PDF-17이 평가척도를 불문하고 다른 공식을 월등하게 앞서는 것으로 나타났다. 반면에 IDF-P와 IDF-S는 P@10과 RP@10



〈그림 8〉 색인어에 적용한 문헌빈도 공식의 성능 비교 - CACM



〈그림 9〉 색인어에 적용한 문헌빈도 공식의 성능 비교 - Medline

척도에서 IDF와 비슷하거나 낮은 성능을 보였다. 이는 이 두 문헌빈도 공식이 검색 결과 순위의 최상부에서는 IDF 공식보다 성능을 향상시키지 못함을 뜻한다.

4.3 질의어에 가중치를 부여하는 경우

문헌빈도 가중치를 탐색어가 아닌 질의어에 부여하는 것은 모든 용어가 아닌 질의에 나타난 용어에만 적용하는 셈이다. 만약 문헌간 유사도를 내적유사도 공식으로 계산하고 문헌길이 정규화를 적용하지 않는다면 문헌빈도 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 문헌빈도 가중치를 어느 쪽에 적용하는가에 따라서 결과가 달라진다.

Singhal(1997)의 IDF-S 공식도 질의어에 가중치를 부여해서 성능향상을 검증한 경우이므로, 앞의 색인어 가중치 실험에서 좋지 못한 성능을 보였다고 해서 폄하할 수는 없다. 질의어 가중치에 대한 문헌빈도 가중치 실험은 검색 시스템에 있어서는 색인어 가중치의 경우보다 더 중요한 의미를 가진다.

1) *p*값의 변화에 따른 PDF 가중치 공식의 성능

질의어에 PDF 가중치를 부여한 경우의 성능이 IDF 가중치를 부여한 경우의 성능에 비해서 달라진 비율을 각 평가척도별로 〈그림 10〉~〈그림 15〉에 제시하였다.

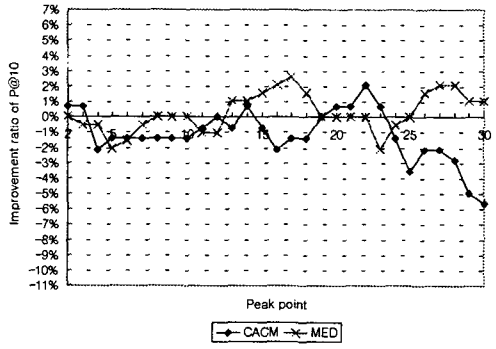
색인어에 부여한 경우와는 달리 IDF 대비 성능향상 효과가 미미하게 나타났다. *p*가 10에서 25 사이인 경우에 대부분의 척도에서 성능이 향상되긴 하지만, 최상위 순위의 성능을 나타내는 P@10이나 RP@10에서는 그렇지 못한 경우도 있었다. 따라서 질의어에 문헌빈도를 부여할 경우에는 PDF 가중치가 IDF 가중치에 비해 상위 순위에서의 성능을 그다지 향상시키지 못한다는 것을 알 수 있다. *p*가 15에서 20 사이인 경우에는 P@10을 제외하면 모든 평가척도에서 1%에서 6% 내외의 성능향상효과가 나타났지만, 색인어에 적용한 경우의 차이에 비하면 크지 않았다.

2) 문헌빈도 공식의 성능 비교

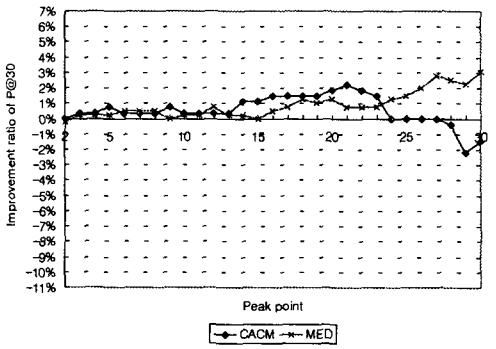
IDF 공식 대비 성능을 기준으로 *p*가 17일 때의 PDF-17 공식과 IDF-P, IDF-S 공식을 비교해서 〈그림 16〉과 〈그림 17〉에 제시하였다.

CACM 실험집단에서는 색인어 가중치의 경우와 달리 IDF-S의 성능이 대체적으로 가장 높게 나타났다. 그러나 RP@10과 RP@30 척도로는 PDF-17이 약간 더 좋은 것으로 나타났다. 이는 PDF-17이 만들어 내는 최상위 문헌의 순위가 더 적합함을 뜻한다.

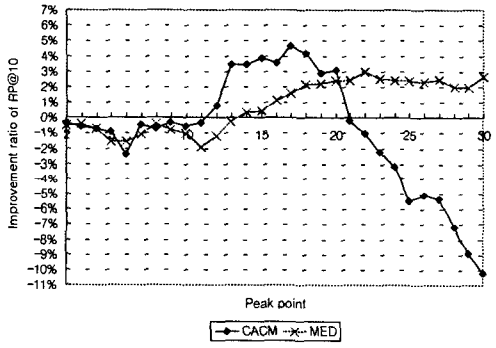
Medline 실험집단에서는 색인어 가중치의 경우와 마찬가지로 PDF-17의 성능이



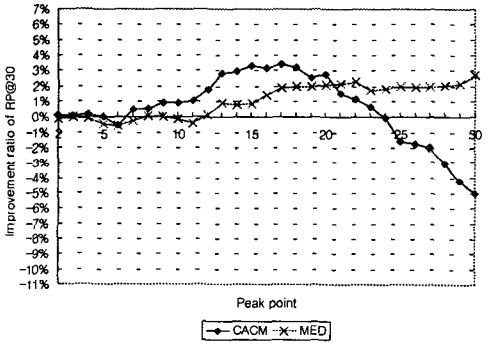
〈그림 10〉 질의어 가중치에 PDF 적용 - P@10



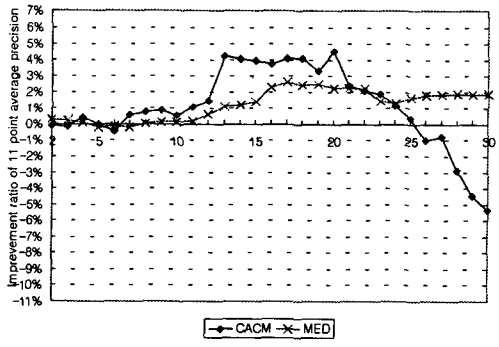
〈그림 11〉 질의어 가중치에 PDF 적용 - P@30



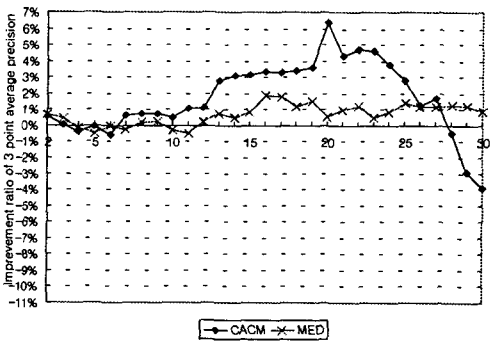
〈그림 12〉 질의어 가중치에 PDF 적용 - RP@10



〈그림 13〉 질의어 가중치에 PDF 적용 - RP@30

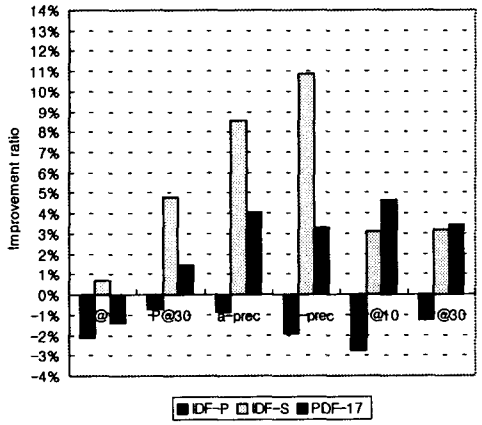


〈그림 14〉 질의어 가중치에 PDF 적용 - 11지점 평균 정확률

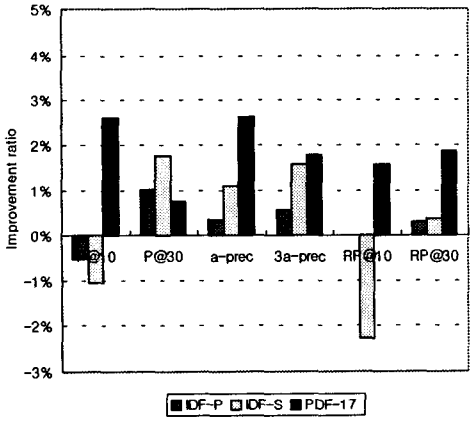


〈그림 15〉 질의어 가중치에 PDF 적용 - 3지점 평균 정확률

가장 좋았지만, 차이는 크지 않았다.



〈그림 16〉 질의어에 적용한 문헌빈도 공식의 성능 비교 - CACM



〈그림 17〉 질의어에 적용한 문헌빈도 공식의 성능 비교 - Medline

5 결 론

문헌빈도가 1이 아닌 p인 경우의 가중치가 가장 높도록 PDF 가중치를 제안하고 실험을 통해 성능을 검증하였다.

실험 결과 색인어 가중치로 사용하는 경우에는 p를 10에서 25 사이로 설정했을 때 다른 문헌빈도 가중치 공식에 비해서 월등한 성능을 얻었다. 그러나 질의어 가중치로 사용할 경우에는 성능 향상 효과가 크지 않았다. 다만 검색결과 최상위 문헌들의 순위는 일관되게 개선되는 것으로 나타났다. 이런 차이는 색인어 가중치로 사용할 때에는 문헌빈도 가중치 공식을 모든 용어에 적용하지만, 질의어 가중치로 사용할 때에는 그렇지 않기 때문으로 짐작된다.

이 연구에서 이용한 실험문헌집단은 수천 건에 불과하므로, 앞으로 대규모 문헌집단에 대한 실험과 다양한 상황에의 적용을 통해 PDF 가중치 공식의 유용성을 검증해야 할 것이다.

참 고 문 헌

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

Robertson, S. E., and Karen. Sparck Jones. 1974. "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.

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.