P-norm 검색의 문헌 순위화 기법에 관한 실험적 연구

A Study of Document Ranking Algorithms in a P-norm Retrieval System

고미영(Mi-Young Ko)* 정영미(Young-Mee Chung)*

목 차

초 록

본 연구의 목적은 문헌의 구조에 근거한 비통계적 용어 가중치 기법을 사용함으로써 기존의 불논리 검색 시스템에 용이하게 적용될 수 있는 P-norm 검색의 효과적인 문헌 순위화 기법을 찾아내는 데 있다. 또한 용어 가중치를 사용하여 검색된 문헌들을 대상으로 상위문헌 몇 개와 유사도가 높은 문헌의 순위를 높여주는 순위 조정 과정을 추가하여 검색성능을 더욱 향상시킬 수 있도록 하였다. 비통계적 가중치 기법으로는 필드 가중치와 근접거리 가중치를 사용하였고, 통계적 기법을 이용한 검색도 실시하여 검색성능을 비교하였다. 순위 조정 실험에서는 문헌간의 유사도 측정의 기준에 되는 상위문헌 수를 1건으로 사용하는 경우부터 5건으로 사용하는 경우까지 5번에 걸친 실험을 실시하였다. 실험결과 비통계적 가중치 기법은 통계적 기법보다 더욱 효과가 있었고, 순위 조정 과정은 전반적으로 검색효율이 크게 향상되는 것으로 밝혀졌다.

ABSTRACT

This study is to develop effective document ranking algorithms in the P-norm retrieval system which can be implemented to the Boolean retrieval system without major difficulties by using non-statistical term weights based on document structure. Also, it is to enhance the performance by introducing the rank adjustment process which rearranges the ranks of retrieved documents according to the similarity between the top ranked documents and the rest of them. Of the non-statistical term weight algorithms, this study uses field weight and term pair distance weight. In the rank adjustment process, five retrieval experiments were performed, ranging between the case of using one record for the similarity measurement and the case of using first five records. It is proved that non-statistical term weights are highly effective and the rank adjustment process enhance the performance further.

키워드 : 문헌 순위화 기법, P-norm 검색, 필드 가중치, 근접거리 가중치, 순위 조정

* 연세대학교 문헌정보학과

■ 논문 접수일 : 1998년 12월 4일

1 서 론

P-norm 검색은 현재 온라인 검색시스템에서 가장 많이 채택되고 있는 불논리 검색의 장점은 그대로 유지하면서, 한편으로는 질문과 부분적으로 일치하는 문헌을 검색할 수 없으며 검색된 문헌들에 순위를 부여할 수 없다는 단점을 해결하기 위해 불논리 검색을 개선한 확장 불논리 검색이다. 로버트슨(1977)은 문헌 순위화의 필요성과 그 방법을 설명하면서 문헌 순위화는 다음과 같이 두 가지 가정을 기초로 하고 있다고 지적하였다. 첫째, 검색시스템은 최적의 성능을 위하여 검색된 문헌들을 순위화하여 탐색자로 하여금 그가 원하는 수준에까지 도달할 수 있도록 해야 하며, 둘째, 문헌들을 어떤 방법으로든 순위화해야 하는 분명한 이유는 일반적으로 '적합'하다는 것은 이원적이 아니라 연속적인 속성을 지니기 때문에 이용자에게는 가장 적합한 문헌들이 덜 적합한 문헌들 보다 앞서 제공되어야 한다는 것이다.

P-norm 검색은 불논리를 이용하여 질문을 비교적 정확하게 표현할 수 있으면서, 탐색식과 부분적으로 일치하는 문헌도 검색할 수 있고 또한 검색된 문헌을 순위화할 수 있는 장점 때문에 현재로서는 불논리 검색이 지니는 여러 가지 문제점을 해결할 수 있는 가장 강력한 모형으로 여겨지고 있다. P-norm 검색모형의 성능은 불논리 연산식을 조절할 수 있도록 하는 p 파라미터 값, 탐색식, 그리고 용어 가중치에 의해 영향을 받는데, 파라미터 값과 탐색식에 대한 연구는 다수 발표되었으나, 용어가중치에 대한 연구는 상대적으로 적어 주로 단어의 출현빈도에 근거한 통계적 기법이 사용되고 있다.

본 연구에서는 P-norm 검색시스템에서 용어의 가중치를 이용하여 문헌을 순위화할 때 새로운 접근 방법을 사용할 경우, 즉 문헌의 구조에 근거한 용어의 가중치를 사용할 경우나, 두 용어가 문헌에서 인접하여 출현하는 거리를 가중치로 이용하는 경우 등과 같이 단어의 출현위치에 따라 가중치를 부여하는 비통계적 기법을 사용하여 문헌을 순위화했을 때가 통계적인 기법들을 이용하여 문헌을 순위화했을 때에 비해 검색성능을 향상시킬 수 있는지의 가능성을 살펴보고자 한다. 또한 용어의 가중치를 이용한 1차 검색 후 출력된 상위문헌 몇 개와 나머지 문헌과의 유사도를 측정하여 유사도의 크기 순으로 문헌의 순위를 조정하는 과정을 추가하여 검색성능을 더욱 향상시킬 수 있는 방안을 제시해 보고자 한다.

2 P-norm 검색 모형 및 문헌 순위화기법

P-norm 검색 모형은 불논리 검색, 퍼지 검색, 벡터공간 모형들을 결합하여 일반화시킨 검색모형으로, 이 검색에서는 탐색어에 가중치가 부여된 질문과 색인어에 가중치가 부여된 문헌과의 거리로써 유사도를 측정하여 문헌의 순위를 산출한다. 문헌 D와 질문Q간의 유사도는 다음과 같이 정의된다.

이 유사도 공식은 탐색어와 색인어에 부여하는 용어 가중치와 p 파라미터 값에 영향을 받으므로 최적의 파라미터 값을 구하는 것이 많은 실험연구의 대상이 되었다. 높은 p 값은 불논리 연산 규칙을 엄격하게 준수하게 되고 낮은 p 값은 좀더 느슨하게 적용하여 1에 가까울수록 벡터 공간 모형과 같이 처리된다. 따라서 단일 p 값을 사용하는 경우에는 불논리 연산기호들을 가능한 한 유연하게 처리하기 위해 p 값이 허용되는 범위 안에서 낮은 값을 부여하는 것이 적절한 것으로 여

$$\mathrm{Sim}(D, Q_{or(p)}) = [\frac{a_1^p d_{A1}^p + a_2^p d_{A2}^p + \text{L } a_n^p d_{An}^p}{a_1^p + a_2^p + \text{L } a_n^p}]^{\frac{1}{p}}$$

$$\mathrm{Sim}(D, Q_{\mathrm{and}(p)}) = 1 - [\frac{a_1^p (1-d_{A1})^p + a_2^p (1-d_{A2})^p + \text{L } a_n^p (1-d_{An})^p}{a_1^p + a_2^p + \text{L } a_n^p}]^{\frac{1}{p}}$$

$$\mathrm{Sim}(D, Q_{\mathrm{not}(p)}) = 1 - \mathrm{sim}(D, Q)$$

겨지고 있으며, Salton 등은 모든 연산기호에 동일한 p 값을 적용할 경우 1-2 사이가 가장 효과가 있다고 밝힌바 있다(Salton et al. 1983).

P-norm 검색을 포함한 모든 검색모형에서도 문헌을 순위화하는 가장 일반적인 방법은 용어의 가중치를 이용하는 것이다. 용어에 가중치를 부여하는 방법은 크게 두 가지로 나눌 수 있는데 하나는 문헌 및 전체문헌집단 내에서의 용어의 출현빈도에 근거하여 가중치를 산출하는 통계적 기법이며, 다른 하나는 용어의 출현빈도 외의 다른 특성을 이용하는 비통계적 기법이다. 통계적 기법으로는 단어빈도, 역문헌빈도, 그리고 적합성 가중치를 들 수 있다.

용어에 가중치를 부여한 다음 문헌과 질문과의 유사도를 측정하여 문헌을 순위화한 최초의 연구는 Salton과 그의 동료들이 SMART 시스템을 이용하여 실시한 일련의 실험연구에서 찾아볼 수 있다(Salton 1971; Salton and Buckley 1988). 그들은 이 시스템을 이용하여 정보검색의 많은 분야에 걸쳐 다양한 실험을 실시하였는데, 실험연구들 가운데 중요한 한 분야가 문헌의 순위화에 관련된 연구들이었다. 그들은 효과적인 문헌 순위화 기법을 개발하기 위한 방법으로 질문과 문헌과의 유사도 공식 및 용어 가중치 함수에 관련된 연구를 실시하였다.

용어의 가중치를 이용하는 기법중 대표적인 비통계적 기법으로는 문헌의 구조를 이용하는 방법이 있다. 문헌의 구조에 근거하여 가중치를 부여하는 방법은 수학적인 공식에 근거하기보다는 주로 경험적인 방법으로서 연구자에 의해 임의의 값이 부여되는 것으로 문헌의 주제를 잘 나타내는 표제어에 출현하는 용어에는 높은 가중치를 부여하고 다른 부분에 출현하는 용어에는 낮은 값을 부여하는 것이다.

Wade와 Willett(1989)은 실험시스템 SIBRIS에서 검색된 문헌의 순위를 부여하기 위한 한 요인으로 문헌의 구조를 이용하였다. 이 시스템에서는 탐색어와 색인어 간의 일치하는 단어수를 기본 유사도값으로 정한 다음 색인어가 문헌의 표제에 출현하면 기본 유사도값에 높은 점수를 추가하여 문헌값을 높이고, 그 외의 다른 곳에서 출현하면 그보다 낮은 점수를 추가하였다. 시소러스를 이용한 탐색어의 추가에 있어서도 표제어와 동의어에는 높은 점수를 추가하는 등 전체적으로 표제어와 관련된 용어에 높은 점수를 주어 문헌의 표제어와 일치되는 단어가 많을수록 문헌의 값이 높아지도록 하였다.

시소러스를 이용하여 색인과 탐색시 용어를 통제하는 시스템에서는 시소러스 상에서의 용어들 간의 관계를 사용하여 문헌의 순위를 부여한다. 시소러스를 이용하는 방법은 시소러스 상에서의 개념관계 중 특히 계층적 관계를 사용하며, 시소러스 상에서의 개념을 표시하는 용어는 노드로, 개념들 사이의 관계는 노드들을 연결하는 링크로 표시하여, 두 개념 노드를 연결하는 링크의 수가 곧 용어들간의 개념적 거리라고 가정하고 이 거

리를 측정하여 문헌의 순위를 부여한다. Rada와 Bicknell(1989)은 MeSH(Medical Subject Heading)에서의 계층적 용어관계를 이용하여 문헌을 순위화한 실험결과 용어간의 계층적 관계만을 사용했을 때 인간이 정한 문헌의 순위와 상당히 유사한 결과가 나타나 시소러스의 용어관계가 정확히 규정되어 진다면 시소러스는 문헌을 순위화하여 검색성능을 향상시킬 수 있는 도구로 사용될 수 있음을 밝혔다.

문헌에서 출현하는 두 용어간의 거리에 근거하여 문헌값을 산출하고 그 값에 따라 문헌의 순위를 부여하는 근접거리 기법도 있다. 이 기법에서는 한 문장 내에서 출현하는 탐색어들에 의해서 문헌값이 결정되는데, 용어쌍간의 평균거리를 산출하여 거리가 적은 것이 우선순위가 되게 하여 문헌을 순위화하는 방법, 두 용어가 인접하여 출현하는 정도를 일정한 범위 내의 가중치 값으로 환산하여 용어쌍들의 가중치의 합을 최종 문헌값으로 산출하여 순위를 결정하는 방법 등이 이용되고 있다. Keen(1994)은 용어쌍간의 거리를 이용하는 방법이 단독으로 사용했을 때보다는 단어빈도나 역문헌빈도 등과 결합했을 때 더욱 효과있음을 밝혔다.

문헌의 순위를 부여하기 위한 근거로 인용문헌을 이용하는 방법도 있다. 이 경우에는 먼저 주제어를 이용한 검색을 실시하여 문헌값을 구한 다음 인용문헌 정보를 이용하여 문헌값을 재조정하는 방법을 취한다. Savoy(1997)는 인용정보를 이용하여 문헌은 노드로 인용관계는 링크로 표현된 인용망을 구성한 다음, 색인어 파일을 이용하여 검색된 문헌 중 문헌값이 높은 노드를 출발점으로 정하고 링크에 의해 연결된 노드들을 활성화시켜 문헌값을 조정하였다.

3 검색실험 환경

P-norm 검색의 효과적인 문헌 순위화 기법을 제시하기 위한 본 연구에서는 2단계의 실험이 실시된다. 1단계 실험은 가장 효과적인 용어 가중치 기법을 찾아내기 위한 실험이며, 2단계 실험은 이용자로부터의 적합성 정보를 구하지 않은 상태에서 검색효율을 더욱 증진시키기 위한 것으로 1단계 실험 결과로 검색된 문헌들을 대상으로 문헌들의 순위를 조정하는 실험이다.

3.1 실험용 데이터베이스 및 색인 파일

본 연구에서 사용된 데이터베이스는 1993년 한국통신에서 개발한 실험용 데이터베이스 KT Test Set이다. 이 데이터베이스의 문헌 파일은 정보과학 분야의 주제를 다루는 한국어 문헌으로서 1985-1993년 사이에 정보과학회 논문지에 수록된 논문 403건, 한국정보과학회 1993년 회의논문집에 수록된 논문 534건, 1984-1993년 사이 정보관리학회지에 수록된 논문 116건 등 총 1,053건을 포함하고 있다.

실험 문헌 파일의 키워드 필드에 수록된 용어들을 사용하여 색인파일들이 작성되었으며, 색인파일은 색인어 파일, 도치색인 파일, 문헌 색인어 벡터 파일로 나뉘어진다. 도치색인 파일에는 용어 가중치를 산출하는데 필요한 정보, 즉 단어빈도, 필드 식별기호, 단어 위치번호에 관한 정보가 포함되어 있으며, 문헌 색인어 벡터 파일에는 순위 조정 과정에서 문헌간의 유사도를 측정하기 위해 필요한 문헌번호와 색인어 벡터가 수록되어 있다.

3.2 P-norm 검색시스템의 구성

P-norm 검색시스템에서 새로운 문헌 순위화 기법을 사용했을 때 검색성능을 향상시킬 수 있는가를 밝히기 위해서 한국어 문헌과 질문을 처리할 수 있는 실험용 P-norm 검색시스템이 구축되었다.

검색 과정은 크게 두 부분으로 나누어지는데 첫 번째 부분은 여러 가지의 용어 가중치를 이용하여 검색된 문헌의 문헌값을 계산한 다음 문헌값에 따라 순서적으로 출력하는 과정으로 사용된 용어 가중치 종류만큼 검색횟수가 반복된다. 두 번째 부분인 순위 조정 과정은 용어 가중치를 이용하여 문헌의 순위를 부여하는 1차 검색 과정을 마친 후 그 결과로 출력된 문헌들을 대상으로 소수의 상위문헌을 기본문헌으로 선정하고 기본문헌들과 나머지 문헌들간의 유사도를 측정한 다음 유사도값 순으로 문헌의 순위를 재정렬하는 과정이다.

실험시스템에서 사용된 p 파라미터 값은 AND 연산기호와 OR 연산기호의 파라미터 값을 구별하여 $p_{and}=2$를 $p_{or}=1$을 각각 부여하였다.

3.3 탐색문 및 적합성 판정

실험에 사용된 질문은 31개로서 실험에 참여한 이용자로부터 직접 구하였다. 이들은 정보과학 분야에 관심을 가지고 있거나 연구에 참여하고 있는 컴퓨터과학과 대학원생들이다. 이용자에 의해서 작성된 탐색신청서는 박사학위를 취득한 전산학 전공자 2명과 본 연구자에 의해 탐색문으로 변환되었다. 탐색문은 불논리 연산기호 AND와 OR을 이용하여 작성되었고 NOT 연산기호는 사용되지 않았다.

검색된 문헌들을 대상으로 이용자들은 질문과의 적합성 여부를 판정하였고, 적합성 판정 척도는 '적합', '부분적합', '부적합'의 세 가지 수준으로 하였다. 적합성 판정결과 검색된 문헌 중 136건이 적합문헌으로, 169건이 부분적합문헌으로 판명되어 한 문헌 당 평균 적합문헌은 4.39건, 부분적합문헌은 5.45건이었다. 검색효율 산출시 부분적으로 적합한 문헌도 적합문헌으로 계산하였기 때문에 적합문헌과 부분적합문헌을 모두 적합문헌으로 계산했을 때 적합문헌은 총 305건으로 한 질문 당 평균 9.84건이었다.

3.4 성능척도

용어 가중치 기법 및 순위 조정 실험의 검색효율을 평가하는 척도로 평균재현율과 평균정확률을 사용하였다. 재현율은 검색된 문헌을 30건으로 제한하여 측정하였고, 정확률은 표준재현율 0.25, 0.5, 0.75 지점에서의 정확률을 각각 측정하여 평균값을 구하였다. 재현율값이 표준적인 값이 아닐 때는 보간법을 사용하여 표준적인 재현율값으로 바꿔 정확률을 산출하였다.

검색효율 비교는 용어 가중치를 이용한 1차 순위화 실험에서는 각 가중치 기법의 평균재현율과 평균정확률의 차이를 비교하였고, 검색된 문헌의 순위를 조정하는 2차 순위 조정 실험에서는 순위 조정 이전의 정확률을 기준으로 하여 순위 조정 이후의 정확률의 증가율을 계산하였다. 또한 표준재현율 11개 지점에서의 정확률을 측정하여 성능곡선을 그려봄으로써 각 기법의 특성도 관찰하였다.

순위 조정 과정에서는 키워드 탐색 후 검색된 문헌들을 대상으로 다시 한번 문헌간의 유사도를 측정하는 과정이 추가되어 탐색시간이 추가로 소

요되기 때문에, 검색효율과 더불어 순위 조정 과정에 필요한 소요 시간을 측정하였다.

4 검색실험

4.1 용어 가중치 기법 실험

색인어 가중치로 용어의 출현위치에 따라 가중치를 부여하는 비통계적 기법을 사용했을 경우 단어의 출현빈도를 이용하는 통계적 기법에 비하여 검색성능을 향상시킬 수 있는가를 비교·분석할 목적으로 비통계적 기법 및 통계적 기법을 이용한 검색을 각각 실시한 다음 그 결과를 비교하였다. 비통계적 기법으로는 필드 가중치와 근접거리 가중치를 사용하였으며, 통계적 기법으로는 단어빈도 가중치, 역문헌빈도 가중치, 그리고 단어빈도·역문헌빈도 가중치를 사용하였다.

필드 가중치에서는 문헌의 내용과 관련된 표제와 초록 필드만 이용하였으며, 질문에서 사용된 용어가 한 필드에 출현하는 문헌보다는 여러 필드에 출현하는 문헌이 질문에 더욱 적합할 가능성이 높다고 보고 여러 필드에 출현하는 용어에 높은 가중치를 부여하는 방법을 사용하였다. 필드에 따라 차등 부여되는 가중치값은 다음과 같다.

① 표제 필드와 초록 필드에 모두 출현하는 경우 = 1.0
② 표제 필드에만 출현하는 경우 = 0.8
③ 초록 필드에만 출현하는 경우 = 0.5

용어간의 거리를 이용하여 문헌의 순위를 부여하는 기법은 그 거리를 문헌값으로 사용하는 방법이 이용되고 있으나 본 실험에서는 P-norm 검색 시스템에 적용하기 위하여 용어간의 근접거리를 용어 가중치로 변형하여 사용하였다. 용어간의 근접거리를 가중치로 이용하는 구체적인 방법은 다음과 같다.

(1) 같은 문장에서 두 용어가 나란히 출현하면 두 용어에 각각 1.0씩 부여한다.

(2) 두 용어 사이에 출현하는 단어수에 따른 가중치값은 다음과 같다.

① 1~2 단어가 출현하는 경우 = 0.9
② 3~4 단어가 출현하는 경우 = 0.8
③ 5~6 단어가 출현하는 경우 = 0.7
④ 7~8 단어가 출현하는 경우 = 0.6
⑤ 9~10 단어가 출현하는 경우 = 0.5

(3) 그 밖의 모든 경우는 가중치값 0.1을 부여한다.

(4) 두 용어의 거리가 여러 값을 가질 때는 가장 높은 값을 사용한다.

비통계적인 기법과 이미 널리 사용되고 있는 통계적인 기법을 결합하였을 경우 검색성능이 향상될 수 있는가를 밝히기 위하여 가중치 기법들을 결합한 실험도 실시하였으며, 또한 비통계적 기법들만을 결합한 경우의 검색성능을 측정하기 위해 필드 가중치와 근접거리 가중치를 결합하였다.

탐색어 가중치로는 다양한 기법을 이용하여 색인어 가중치를 부여하기 때문에 모든 탐색어에 동일하게 '1'을 부여하는 이진값을 사용하였다.

4.2 문헌간의 유사도를 이용한 순위 조정

키워드를 이용한 검색은 동일한 주제를 표현하는 데 있어 질문에서 사용된 탐색어의 형태와 문헌 색인어의 형태가 정확하게 일치하지 않으면 질문에 적합한 문헌일지라도 누락되는 문제점이

있다. 이러한 문제를 보완하려는 검색기법 중 하나로 인용문헌을 이용하는 방법이 제시되었으나 데이터베이스에 인용 사항이 포함되어 있지 않는 경우는 사용할 수 없다는 문제가 있다.

본 실험에서는 도치색인 파일을 이용한 검색 시스템에서 쉽게 구할 수 있는 색인어 정보를 이용하여, 용어 가중치를 이용한 1차 검색 결과 질문과의 유사도가 높은 것으로 판정된 상위의 문헌들과 내용상으로 가까운 문헌의 순위를 높여주도록 함으로써 검색된 문헌들의 순위를 조정하여 검색성능을 개선하도록 하였다.

유사도 측정 방법은 용어 가중치를 이용한 1차 검색결과 각 탐색문별로 1위부터 5위까지 출력된 상위문헌과 나머지 문헌간의 유사도를 측정하는 것으로서, 문헌간의 유사도는 각 문헌의 색인어들을 비교하여 측정된다. 유사도 측정공식으로는 코사인 유사계수를 사용하였다.

유사도 측정 대상은 각 탐색문별 1차 검색 후 검색된 문헌 중 처음 30건으로 제한하였으며, 순위 조정 실험에는 1차 실험결과 검색된 문헌수를 상위 30건으로 제한했을 때 재현율이 낮아서 두 실험 결과를 비교하는데 문제점이 있는 탐색문 5개를 제외한 탐색문 26개가 사용되었다.

검색된 상위의 문헌과 나머지 문헌과의 유사도 측정은 상위문헌 몇 건을 유사도 측정에 사용하느냐에 따라 유사도 값이 달라져 검색효율에 영향을 미친다. 본 실험에서는 용어 가중치를 이용한 1차 검색 후 출력된 상위문헌 1건을 기본문헌으로 사용하여 상위문헌 1건과 나머지 문헌과의 유사도를 측정하여 순위를 조정한 실험에서부터, 기본문헌을 1건씩 증가시켜 상위문헌 5건과 나머지 문헌과의 유사도를 측정하여 순위를 조정하는 실험까지 모두 5번에 걸친 순위 조정 실험을 실시하였다.

순위 조정 실험에 사용된 기본문헌의 특성이 순위 조정 실험결과의 검색성능에 미치는 영향을 분석하기 위하여 기본문헌으로 사용되는 상위문헌 5건을 분석한 결과 기본문헌 중 적합문헌으로 판명된 문헌수는 〈표 1〉과 같다. 괄호 안의 비율은 순위 조정에 사용되는 전체 기본문헌 중 적합문헌이 차지하는 비율이다.

순위 조정 실험에서 기본문헌을 2건 이상 사용하여 나머지 문헌들과의 유사도값을 산출할 때는 각 기본문헌과 나머지 문헌을 각각 비교하여 유

〈표 1〉 순위 조정에 사용된 기본문헌 중 적합문헌수

구분 / 기본문헌	단어빈도 가중치	역문헌빈도 가중치	단어빈도 · 역문헌빈도 가중치	필드 가중치	근접거리 가중치
첫번째 문헌	25(96.15%)	22(84.62%)	23(88.64%)	23(88.46%)	22(84.62%)
두번째 문헌	16(78.85%)	17(75.00%)	18(78.85%)	21(84.62%)	18(76.92%)
세번째 문헌	12(67.95%)	16(70.51%)	14(70.51%)	18(79.49%)	15(70.51%)
네번째 문헌	17(67.31%)	13(65.38%)	14(66.35%)	17(75.96%)	15(67.31%)
다섯번째 문헌	11(62.31%)	13(62.31%)	13(63.08%)	15(72.31%)	13(63.85%)

사도값을 산출한 다음 평균값을 구하였다. 순위 조정에 사용되는 기본문헌이 2건 이상 일 때 유사도값 산출 공식은 다음과 같다.

$$\mathrm{Sim}(D, Q) = \frac{1}{k} \sum_{i=k+1}^{z} \sum_{j=1}^{k} S(X_i, Y_j)$$

k = 순위 조정에 사용되는 기본문헌 수

z = 유사도 측정 대상이 되는 문헌 수

X_i, Y_j = 1차 검색후 질문과 문헌간의 유사도 값으로 정렬된 문헌 번호

검색된 문헌의 최종 순위는 먼저 순위 조정을 위해 사용된 기본문헌들에게 우선 순위가 부여되고 이어서 위의 공식을 이용하여 산출된 유사도 값의 크기 순으로 순위가 정해진다. 예를 들면 기본문헌으로 상위문헌 2건을 사용했을 때는 최종 문헌의 순위는 기본문헌 2건이 각각 첫 번째 문헌과 두 번째 문헌이 되며 세 번째 문헌부터는 문헌간의 유사도값의 크기 순으로 결정된다.

5 실험결과의 분석

5.1 용어 가중치의 이용 결과

5.1.1 통계적 기법 및 비통계적 기법 결과

통계적 가중치 및 비통계적 가중치를 이용하여 문헌을 순위화한 실험결과는 〈표 2〉와 같다. 검색효율을 측정하기 위하여 적합문헌총수를 산출할 때 본 실험에서는 각 가중치 기법의 검색결과 상위 30건으로 문헌수를 제한하였기 때문에, 평균재현율은 이와 같이 제한된 범위 내에서 측정되었다.

통계적 기법 중 단어빈도 가중치를 사용하여 문헌을 순위화했을 때의 재현율은 0.8554, 역문헌빈도 가중치를 사용하였을 때의 재현율은 0.8571, 그리고 단어빈도 · 역문헌빈도 가중치의 재현율은 0.8571로 통계적 기법들 사이에는 차이가 거의 없었다. 그러나 비통계적 기법 중 필드 가중치를 사용했을 때의 재현율은 0.9018이고, 근접거리 가중치의 재현율은 0.8636으로 두 기법간에는 차이를 보이고 있다. 따라서 통계적 기법들 간에는 검색되는 적합문헌수에 있어서 차이가 없었으나, 비통계적 기법인 필드 가중치를 사용했을 때는 통계적 기법을 사용했을 때보다 재현율이 4.47~4.64% 높아져 검색되는 적합문헌수가 증가되는 것으로 나타났다. 비통계적 기법 중 근접거리 가중치를 사용했을 때는 통계적 기법들을 사용했을 때보다는 재현율이 높아졌으나 그 정도가 1% 미만이어서 사실상의 차이는 없는 것으로 나타났다.

〈표 2〉 용어 가중치를 이용한 실험결과의 검색효율 비교

가중치 구분 / 척도	〈통계적 기법〉			〈비통계적 기법〉	
	단어빈도	역문헌빈도	단어빈도 · 역문헌빈도	필드	근접거리
평균재현율	0.8554	0.8571	0.8571	0.9018	0.8636
평균정확률	0.5686	0.5618	0.5771	0.6386	0.5723

통계적 기법을 사용했을 때의 정확률은 단어빈도·역문헌빈도 가중치의 경우가 0.5771로 단어빈도 가중치의 정확률 0.5686과 역문헌빈도 가중치의 정확률 0.5618보다 약간 높게 나타났다. 비통계적 기법 중 필드 가중치를 사용하였을 때 정확률은 0.6386으로 현재 가장 보편적으로 사용되고 있는 단어빈도·역문헌빈도 가중치 경우보다 6.15% 높았다. 따라서 필드 가중치를 사용했을 때는 재현율과 마찬가지로 정확률에 있어서도 통계적 기법과는 현저한 차이를 보였다. 근접거리 가중치를 사용했을 때의 정확률은 0.5723으로 단어빈도·역문헌빈도 가중치의 경우보다는 근소한 차이로 낮아졌고, 단어빈도 가중치 및 역문헌빈도 가중치의 경우보다는 근소한 차이로 높아져, 근접거리 가중치는 검색효율에 있어서 통계적 기법과 차이가 거의 없는 것으로 밝혀졌다.

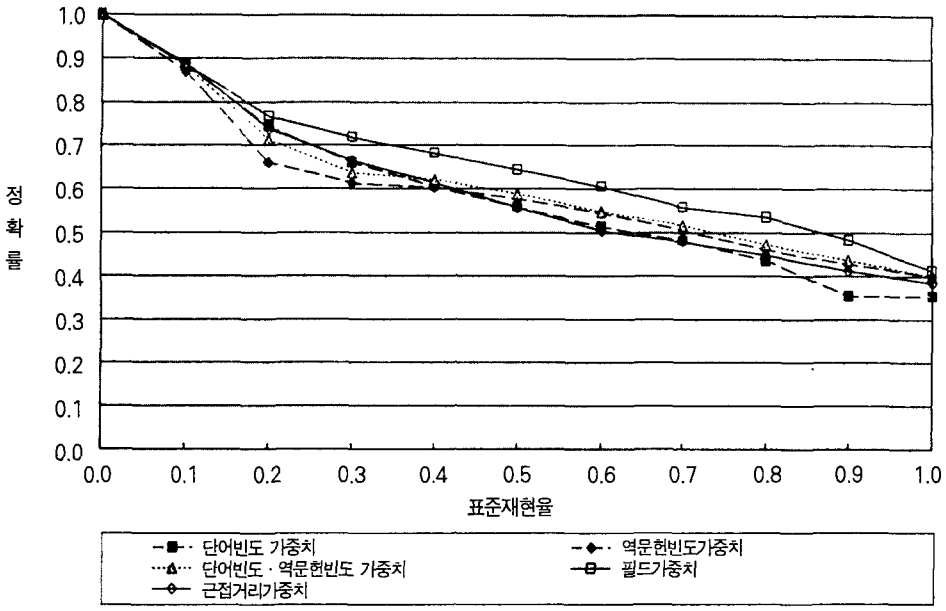
적합문헌들의 평균순위에 있어서 통계적 기법과 비통계적 기법 사이에 차이가 있는가를 관찰하기 위해서 검색된 적합문헌들의 평균순위를 측정하였다. 문헌의 적합성을 판정할 때 적합문헌을 '적합'과 '부분적합'으로 구분하였기 때문에, 적합문헌의 순위와 부분적합문헌의 순위를 각각 측정하여 각 기법간의 차이뿐만 아니라 한 기법 내에서의 적합문헌의 순위와 부분적합문헌의 순위도 비교하였다. 각 기법별 적합문헌의 순위와 부분적합문헌의 평균순위는 〈표 3〉과 같다.

적합문헌의 평균순위가 가장 앞선 기법은 필드 가중치 기법이며, 이어서 단어빈도·역문헌빈도 가중치, 근접거리 가중치, 역문헌빈도 가중치, 그리고 단어빈도 가중치 기법 순이었다. 필드 가중치의 경우 적합문헌과 부분적합문헌을 합한 적합문헌의 평균순위는 11.3으로 단어빈도·역문헌빈도 가중치의 평균순위 14.7 보다 3.4순위가 앞섰다. 따라서 필드 가중치를 사용할 경우 단어빈도·역문헌빈도 가중치를 사용할 때보다 적합문헌들은 약 3건 정도 앞서 출현하는 것으로 밝혀졌다. 근접거리 가중치의 경우는 단어빈도·역문헌빈도 가중치를 사용했을 때보다 근소한 차이로 뒤쳐져 큰 차이가 없는 것으로 나타났다. 통계적 기법들 중에서는 단어빈도·역문헌빈도 가중치를 사용했을 때의 순위가 단어빈도 가중치와 역문헌빈도 가중치를 각각 사용한 경우보다 약간 앞서는 것으로 나타났다.

적합문헌의 순위와 부분적합문헌의 순위의 차이를 살펴보면, 적합문헌의 순위는 기법에 따라 9.5-12.5, 부분적합문헌의 순위는 13.1-18.2로 적합문헌은 부분적합문헌에 비하여 기법에 따라 3.6-5.7순위가 앞서 적합문헌집단과 부분적합문

〈표 3〉 '적합' 문헌과 '부분적합' 문헌의 평균순위 비교

구분 / 가중치 기법	적합문헌 순위	부분적합문헌 순위
단어빈도	12.4	18.2
역문헌빈도	12.5	17.6
단어빈도·역문헌빈도	12.2	17.1
필드	9.5	13.1
근접거리	12.3	17.3



〈그림 1〉 용어 가중치를 이용한 실험 결과의 성능곡선

헌집단은 순위에 있어서 차이가 있는 것으로 나타났다. 또한 각 기법들은 적합문헌에서 보다는 부분적합문헌의 순위에서 더 큰 차이를 보여, 우수한 기법과 그렇지 못한 기법은 부분적합문헌의 순위에 더 큰 영향을 미치는 것으로 나타났다.

〈그림 1〉은 각 기법의 특성을 비교해보기 위해 표준재현율 지점에서의 정확률을 측정하여 성능곡선으로 나타낸 것이다. 성능곡선을 살펴보면 필드 가중치의 성능곡선이 가장 상위에 위치하여 검색성능이 가장 우수한 것을 볼 수 있다. 필드 가중치 곡선은 재현율 0.1지점까지는 다른 곡선들과 구별되지 않다가 그 이후부터 점차 차이를 보이기 시작하여 0.3지점부터는 큰 차이를 보여주고 있다. 근접거리 가중치 곡선과 단어빈도 가중치 곡선은 상위지점에서는 필드 가중치 다음으로 성능이 우수하였으나 중간지점부터는 급격하게 저하되는 것으로 나타났다. 단어빈도 · 역문헌빈도 가중치 곡선은 상위부분에서는 상대적으로 성능이 저조하였으나 중간지점부터는 필드 가중치 곡선 바로 아래에 위치하여 단어빈도 · 역문헌빈도 가중치를 사용할 경우에는 중간부분에서 다른 통계적 가중치 기법보다 성능이 우수한 것을 볼 수 있다. 역문헌빈도 가중치 곡선은 상위지점에서는 가장 아래에 위치하여 다른 기법들보다 성능이 저조하였으나 중간지점부터는 성능 감소의 폭이 적어 단어빈도 · 역문헌빈도 가중치 곡선에 근접하였다.

5.1.2 통계적 기법과 비통계적 기법의 결합 결과

통계적 기법과 비통계적 기법의 결합형태는 필드 가중치와 통계적 기법들을 결합한 세 가지 형태와 근접거리 가중치와 통계적 기법들을 결합한 세 가지 형태이며, 비통계적 기법들만의 결합인

〈표 4〉 필드 가중치와 다른 가중치 기법을 결합한 실험결과

가중치 / 구분 / 척도	필드 + 단어빈도	필드 + 역문헌빈도	필드 + 단어빈도 · 역문헌빈도	필드 + 근접거리
평균재현율	0.8998	0.9247	0.9306	0.8977
평균정확률	0.6331	0.6520	0.6586	0.6257

필드 가중치와 근접거리 가중치의 결합형태도 실험되었다. 필드 가중치와 통계적 가중치를 결합한 실험결과 및 필드 가중치와 근접거리 가중치를 결합한 실험결과는 〈표 4〉와 같다.

필드 가중치와 통계적 기법을 결합했을 때의 재현율은 필드 가중치와 단어빈도 · 역문헌빈도 가중치를 결합하였을 경우 가장 높았으며, 그 다음은 필드 가중치와 역문헌빈도 가중치를 결합한 경우였고, 필드 가중치와 단어빈도 가중치를 결합한 경우가 세 가지 형태 중 가장 낮았다. 필드 가중치+단어빈도 · 역문헌빈도 가중치의 재현율 0.9306은 필드 가중치만을 사용했을 때보다 2.88% 높아졌고, 단어빈도 · 역문헌빈도 가중치만을 사용했을 때보다 7.35% 높아진 것이다. 필드 가중치와 근접거리 가중치를 결합한 경우에는 필드 가중치만을 사용했을 때와 차이가 거의 없었으며, 근접거리 가중치만을 사용했을 때의 재현율보다 3.41% 높아졌다.

필드 가중치와 통계적 기법을 결합했을 때 정확률이 가장 높은 기법은 재현율에서와 마찬가지로 필드 가중치와 단어빈도 · 역문헌빈도 가중치를 결합한 것이었다. 필드+단어빈도 · 역문헌빈도 가중치 기법의 정확률 0.6586은 필드 가중치만을 사용했을 때보다 2.0% 높아진 것이며, 단어빈도 · 역문헌빈도 가중치만을 사용했을 때보다는 8.15% 높아진 것이다.

필드+근접거리 가중치 기법의 정확률은 필드 가중치와 통계적 기법을 결합한 경우보다 낮았으며, 필드 가중치만을 사용했을 때와는 차이가 거의 없었다. 그러나 통계적 기법들만의 결합인 단어빈도 · 역문헌빈도 가중치 기법과 비교해보면 4.86%가 높아진 것으로 같은 특성을 지니는 두 기법을 결합할 경우에는 통계적 기법들의 결합보다는 비통계적 기법들의 결합이 더욱 효과가 있음을 보여주고 있다.

필드 가중치와 다른 가중치 기법을 결합하였을 경우에는 기법에 따라 약간씩 차이가 있기는 하지만 전체적으로 볼 때 어떤 기법의 경우에도 재현율과 정확률이 크게 떨어지지 않고 일정한 수준을 유지하고 있어서 필드 가중치는 검색효율을 증가시킬 수 있는 가장 효과적인 방법으로 밝혀졌다.

필드 가중치와 통계적 기법들을 결합했을 때, 그리고 필드 가중치와 근접거리 가중치를 결합했을 때 적합문헌들의 평균순위가 어떻게 변화하는가를 분석하기 위하여 적합문헌과 부분적합문헌의 순위를 측정한 결과는 〈표 5〉와 같다.

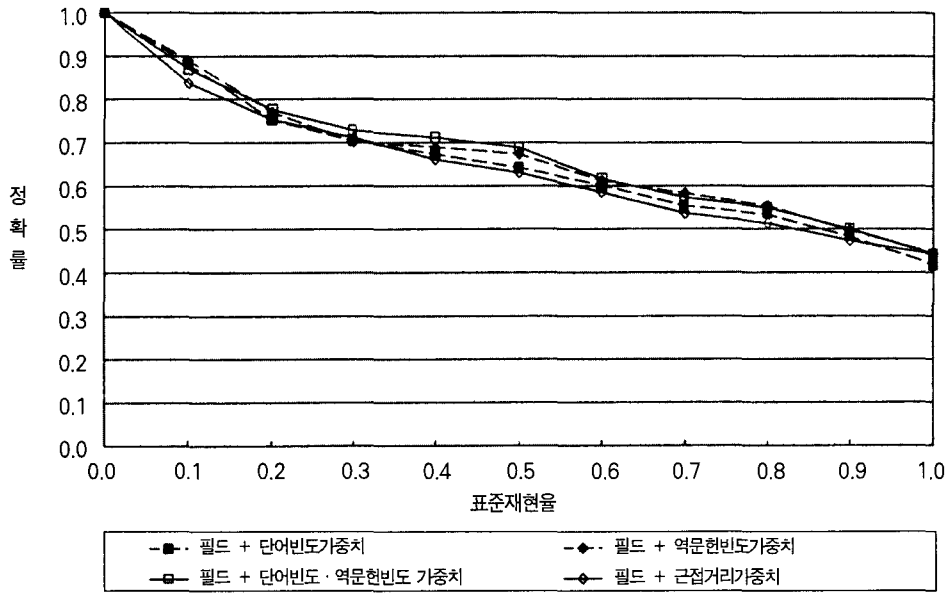
필드 가중치와 다른 기법들을 결합했을 때의 적합문헌의 평균순위는 필드+단어빈도 · 역문헌빈도 가중치 기법을 제외하면 필드 가중치만을 사용했을 때의 적합문헌의 순위와 거의 동일하였다. 필드+단어빈도 · 역문헌빈도 가중치 기법의

〈표 5〉 필드 가중치와 다른 기법을 결합했을 때의 적합문헌의 평균순위

가중치 구분	필드 + 단어빈도		필드 + 역문헌빈도		필드 + 단어빈도 · 역문헌빈도		필드 + 근접거리	
	적합	부분적합	적합	부분적합	적합	부분적합	적합	부분적합
순위	9.7	13.1	8.9	12.9	8.7	12.8	9.7	13.2

적합문헌 및 부분적합문헌의 순위는 필드 가중치 기법보다 각각 0.8, 0.3 순위 빨라져 필드 가중치와 단어빈도 · 역문헌빈도 가중치를 함께 사용했을 때는 필드 가중치만을 사용했을 때보다 적합문헌들이 약 1건 정도 앞서 출현하는 것으로 나타났다.

〈그림 2〉는 필드 가중치와 다른 가중치 기법들을 결합했을 때의 검색성능의 특성을 관찰하기 위하여 표준재현율에서의 정확률을 측정하여 성능곡선으로 나타낸 것이다. 성능곡선을 보면 필드+단어빈도 · 역문헌빈도 가중치 곡선과 필드+역문헌빈도 가중치 곡선은 중상위 지점에서 필드+단어빈도 · 역문헌빈도 가중치 곡선이 약간의 차이로 상위에 있는 것을 제외하고는 전체적으로 거의 일치하여 두 기법간의 검색성능이 비슷한 것을 볼 수 있다. 필드+단어빈도 가중치 곡선은 위의 두 곡선보다 하위에 위치하여 검색성능이 두 기법에 비해 떨어지는 것으로 나타났다. 필드



〈그림 2〉 필드 가중치 기법과 다른 가중치를 결합한 결과의 성능곡선

〈표 6〉 근접거리 가중치와 통계적 기법을 결합한 실험결과

척도 \ 가중치 구분	근접거리 + 단어빈도	근접거리 + 역문헌빈도	근접거리 + 단어빈도 · 역문헌빈도
평균재현율	0.8511	0.8590	0.8590
평균정확률	0.5504	0.5764	0.5809

〈표 7〉 근접거리 가중치와 통계적 기법을 결합했을 때의 적합문헌의 평균순위

가중치 구분	근접거리 + 단어빈도		근접거리 + 역문헌빈도		근접거리 + 단어빈도 · 역문헌빈도	
	적합	부분적합	적합	부분적합	적합	부분적합
순위	12.5	18.6	12.3	17.5	12.3	17.3

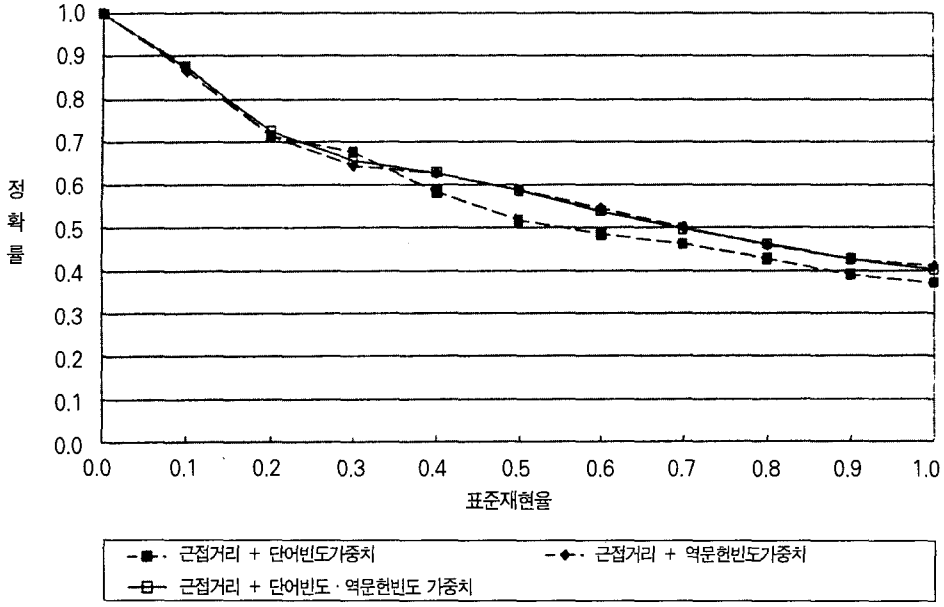
+근접거리 가중치 곡선은 전반적으로 가장 아래에 위치하여 필드 가중치와 통계적 기법의 결합보다는 검색성능이 뒤떨어지는 것을 볼 수 있다.

근접거리 가중치 기법은 질문에서 사용된 용어들이 문헌에서 얼마나 가까이 출현하느냐에 따라 가중치를 부여하는 비통계적 기법이지만, 단일기법으로 사용했을 때는 통계적 기법들에 비하여 검색효율이 크게 향상되지 않는 것으로 밝혀졌다. 그러나 근접거리 가중치와 통계적 가중치를 함께 이용할 경우에는 검색효율이 향상될 수 있는가를 살펴보기 위해서 두 종류의 가중치 기법들을 결합하였다. 결합 형태는 근접거리+단어빈도 가중치, 근접거리+역문헌빈도 가중치, 그리고 근접 거리+단어빈도 · 역문헌빈도 가중치 기법이다.

근접거리 가중치와 통계적 가중치를 결합했을 때의 실험결과는 〈표 6〉과 같다. 근접거리 가중치에 통계적 가중치를 결합했을 때의 재현율과 정확률은 한 기법만을 사용했을 때보다 기법에 따라 약간씩 다르기는 하지만 큰 차이가 없는 것으로 나타나, 근접거리 가중치는 단일 기법으로 사용했을 때와 마찬가지로 통계적 기법과 결합해서 사용할 경우에도 큰 효과가 없는 것으로 밝혀졌다.

근접거리 가중치와 통계적 가중치를 함께 사용했을 때의 적합문헌과 부분적합문헌의 평균순위는 〈표 7〉과 같다. 근접거리 가중치와 통계적 가중치를 함께 사용했을 때는 근접거리+단어빈도 · 역문헌빈도 가중치 기법의 적합문헌의 순위가 가장 빨랐으며 다음은 근접거거리+역문헌빈도 가중치 그리고 근접거리+단어빈도 가중치 순이었다. 그러나 전체적으로 볼 때 근접거리 가중치와 통계적 가중치들을 결합하였을 때의 적합문헌의 순위도 각각 사용했을 때와 거의 차이가 없는 것으로 나타났다.

〈그림 3〉은 근접거리 가중치와 통계적 가중치를 결합했을 때의 정확률을 표준재현율 지점에서 측정하여 성능곡선으로 나타낸 것이다. 근접거리



〈그림 3〉 근접거리 가중치와 통계적 기법을 결합한 결과의 성능곡선

가중치와 통계적 가중치를 결합한 곡선들의 특성을 살펴보면 근접거리+단어빈도 · 역문헌빈도 가중치 곡선과 근접거리+역문헌빈도 가중치 곡선은 전구간에 걸쳐 거의 일치하고 있어 두 기법간의 검색성능이 비슷한 것을 볼 수 있다. 그러나 근접거리+단어빈도 가중치 곡선은 상위지점에서는 다른 곡선들과 거의 구별이 되지 않다가 중간지점 이후부터는 다른 두 곡선들보다 성능이 떨어지는 것으로 나타났다.

5.2 순위 조정 결과

5.2.1 순위 조정에 사용된 기본문헌수에 따른 결과

순위 조정 실험에서는 상위문헌 몇 건을 기본문헌으로 사용하느냐에 따라 검색효율이 다르게 나타나기 때문에, 상위문헌 1건을 사용하는 경우에서부터 상위문헌 5건을 사용하는 경우까지 기본문헌수를 1건씩 증가시키면서 5번에 걸친 실험을 실시하여 실험결과를 분석하였다. 순위 조정 실험의 효과를 측정하기 위해 먼저 탐색문 26개의 순위 조정 이전의 평균정확률을 산출한 다음 순위 조정 실험 결과의 정확률과 비교하였다. 순위 조정 실험 이전의 정확률 및 순위 조정에 사용된 기본문헌수에 따른 정확률은 〈표 8〉과 같다.

상위문헌 1건을 사용하여 검색된 문헌의 순위를 조정했을 때의 평균정확률을 보면 필드 가중치를 사용했을 때가 0.7152로 가장 높았으며, 다음은 근접거리 가중치의 경우로 0.6971이었고, 단어빈도 가중치를 사용했을 때는 0.6862였다. 단어빈도 · 역문헌빈도 가중치의 경우와 역문헌빈도 가중치의 경우는 각각 0.6728과 0.6727로 비슷한 수준이었다.

순위 조정 이전의 정확률과 비교하면 필드 가

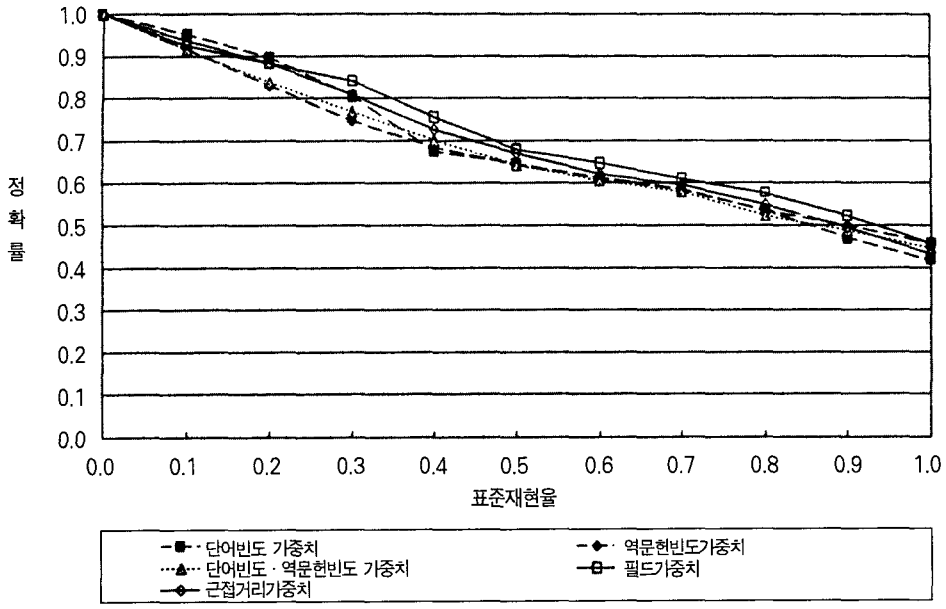
〈표 8〉 순위 조정에 사용된 기본문헌 중 적합문헌수

기본문헌수 \ 가중치 구분	단어빈도	역문헌빈도	단어빈도 · 역문헌빈도	필드	근접거리
순위조정 이전	0.6244	0.6147	0.6304	0.7006	0.6305
상위문헌 1건 (증가율)	0.6862 (+9.90%)	0.6727 (+9.44%)	0.6728 (+6.73%)	0.7152 (+2.08%)	0.6971 (+10.56%)
상위문헌 2건	0.6792 (+8.78%)	0.6781 (+10.31%)	0.7009 (+11.18%)	0.7658 (+9.31%)	0.6989 (+10.85%)
상위문헌 3건	0.6707 (+7.42%)	0.6693 (+8.88%)	0.6767 (+7.34%)	0.7718 (+10.16%)	0.6851 (+8.66%)
상위문헌 4건	0.6861 (+9.88%)	0.6652 (+8.22%)	0.6749 (+7.06%)	0.7628 (+8.88%)	0.6842 (+8.52%)
상위문헌 5건	0.6789 (+8.78%)	0.6618 (+7.66%)	0.6779 (+7.53%)	0.7595 (+8.41%)	0.6839 (+8.47%)

중치 기법은 2.08% 증가하여 다른 기법들에 비해 향상 정도가 가장 적었으며, 근접거리 가중치 기법은 10.56% 증가하여 가장 큰 폭으로 향상되었다. 단어빈도 가중치 기법은 9.90%, 그리고 역문헌빈도 가중치 기법은 9.44% 증가하였으나, 단어빈도 · 역문헌빈도 가중치 기법은 6.73% 증가로 두 가중치를 각각 사용했을 때보다 증가의 폭이 적은 것으로 나타났다. 상위문헌 1건을 기본문헌으로 사용하여 순위를 조정했을 때는 전체적으로 정확률은 향상되었지만 순위 조정 이전에서와 같이 가중치 기법간의 차이는 크지 않은 것으로 나타났다. 특히 필드 가중치 기법의 경우에는 정확률 향상의 폭이 적어 순위 조정 이전과는 달리 다른 기법들과의 차이가 적었다.

상위문헌 1건을 기본문헌으로 사용하여 순위를 조정했을 때의 성능곡선은 〈그림 4〉와 같다. 순위 조정 이전의 곡선들과 비교해 보면(그림 1 참조) 순위 조정 이전의 성능곡선들이 상위의 재현율 지점에서 급격한 경사를 이루며 하강하는 것과는 달리 순위 조정 이후의 곡선들은 전체적으로 완만한 경사를 이루며 서서히 하강하는 것으로 나타나 상위지점에서의 검색성능이 더욱 향상되는 것으로 밝혀졌다.

상위문헌 2건을 사용했을 때의 정확률은 필드 가중치 기법을 사용했을 때 0.7658로 가장 높았으며, 그 다음은 단어빈도 · 역문헌빈도 가중치 기법으로 0.7009이었고, 이어서 근접거리 가중치 기법은 0.6987, 단어빈도 가중치 기법은 0.6792, 그리고 역문헌빈도 가중치 기법은 0.6781로 가장 낮았다. 상위문헌 1건을 사용했을 때와 비교하면 단어빈도 가중치 기법을 사용했을 때 정확률이 약간 낮아진 것을 제외하고는 모든

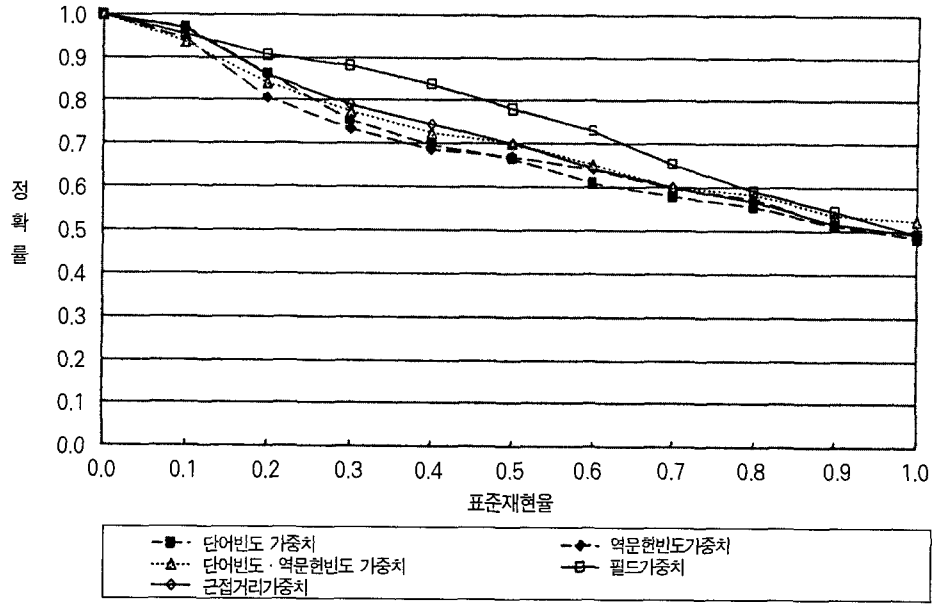


〈그림 4〉 상위문헌 1건을 기본문헌으로 사용한 순위 조정 결과의 성능곡선

기법에서 정확률이 높아져 상위문헌 2건을 사용하였을 때가 1건을 사용하였을 때보다 검색효율이 더욱 향상되는 것으로 밝혀졌다.

정확률 증가율을 기법별로 살펴보면 먼저 필드 가중치 기법의 정확률은 순위 조정 이전보다 9.31% 증가함으로써 1건을 사용했을 때보다 큰 폭으로 향상되어 다른 기법들과의 차이가 다시 현저해졌다. 근접거리 가중치 기법은 순위 조정 이전보다 10.85% 증가하여 높은 증가율을 보였고, 단어빈도 · 역문헌빈도 가중치 기법도 11.18% 증가하여 모든 순위 조정 실험 중 가장 높은 증가율을 보였다. 단어빈도 가중치 기법과 역문헌빈도 가중치 기법의 경우는 다른 기법들에 비해 상대적으로 정확률이 낮았고, 특히 단어빈도 가중치 기법은 순위 조정 이전에 비해 8.78% 증가하여 모든 기법 중 가장 적은 증가율을 보였을 뿐만 아니라 1건을 사용했을 때보다 정확률이 약간 낮아졌다. 그러나 역문헌빈도 가중치 기법은 순위 조정 이전보다는 10.31% 증가하여 단어빈도 가중치와의 차이가 상대적으로 적어졌다.

상위문헌 2건을 사용하여 문헌의 순위를 조정하였을 때의 성능곡선은 〈그림 5〉과 같다. 상위문헌 2건을 사용했을 때의 성능곡선들의 특징은 필드 가중치 기법의 곡선과 나머지 기법의 곡선들과의 간격의 폭이 커져 성능의 차이가 확연히 드러난다는 점이다. 특히 이러한 차이는 중간부분에서 가장 두드러진다. 근접거리 가중치 기법의 곡선은 상위지점에서는 성능이 우수하였으나 중간지점부터는 성능이 급격하게 감소하였고, 단어빈도 · 역문헌빈도 가중치 기법의 곡선은 중간지점부터는 필드 가중치를 제외한 다른 기법들 보다 상위에 위치하여 상위지점보다는 하위지점에서의 검색성능이 상대적으로 우수한 것으로 나타났다. 단어빈도 가중치 기법의 성능곡선은 상위



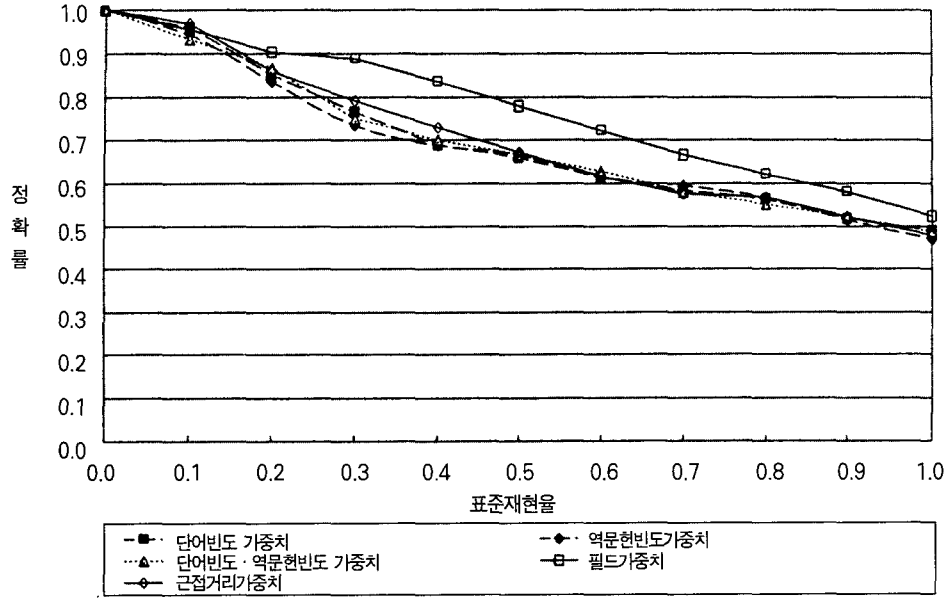
〈그림 5〉 상위문헌 2건을 기본문헌으로 사용한 순위 조정 결과의 성능곡선

지점에서는 상대적으로 높았으나 하위지점에서는 가장 아래에 위치하여 검색성능이 급격하게 감소하였고, 역문헌빈도 가중치 기법의 곡선은 대부분의 지점에서 가장 아래에 위치하여 다른 기법들에 비해 성능이 뒤떨어지는 것으로 나타났다.

상위문헌 3건을 기본문헌으로 사용했을 때의 정확률을 보면 필드 가중치 기법의 정확률이 0.7718로 가장 높았고, 이어서 근접거리 가중치 기법의 정확률 0.6851, 단어빈도 · 역문헌빈도 가중치 기법의 정확률 0.6767, 단어빈도 가중치 기법의 정확률 0.6707 순이었으며, 역문헌빈도 가중치 기법의 정확률은 0.6693으로 가장 낮았다.

정확률을 순위 조정 이전과 비교하여 증가율을 측정해보면 필드 가중치 기법의 정확률은 10.16% 증가되어 2건을 사용했을 때의 정확률 증가보다 많았다. 그러나 근접거리 가중치 기법은 8.66%, 단어빈도 · 역문헌빈도 가중치 기법은 7.34%, 단어빈도 가중치 기법은 7.42%, 역문헌빈도 가중치 기법은 8.88% 증가에 그쳐 2건을 사용했을 때의 증가율보다 적었다. 따라서 상위문헌 3건을 사용하여 검색된 문헌들의 순위를 조정하였을 때는 필드 가중치 기법을 제외한 나머지 기법들은 2건을 사용했을 때보다 오히려 검색효율이 감소되는 것으로 밝혀졌다.

〈그림 6〉은 상위문헌 3건 사용하여 순위를 조정했을 때의 성능곡선이다. 3건을 이용했을 때는 필드 가중치 기법의 곡선과 다른 기법의 곡선들과의 간격의 차이가 상위지점에서뿐만 아니라 마지막 지점까지 뚜렷이 나타나 필드 가중치 기법의 성능과 나머지 기법간의 성능 차이가 더욱 커지고 있음을 보여주고 있다. 근접거리 가중치 기법의 곡선은 상위지점에서 통계적 가중치 기법의 곡선들과 약간의 차이를 보이고 있으나 중간지점부터는 전혀 차이를 보이지 않고, 통계적 가중치



〈그림 6〉 상위문헌 3건을 기본문헌으로 사용한 순위 조정 결과의 성능곡선

기법의 곡선들간에는 상위부분에서 역문헌빈도 가중치 곡선이 약간의 차이로 뒤떨어지는 것을 제외하고는 전구간에 걸쳐 거의 구별이 되지 않고 있어 통계적 가중치 기법간의 성능 차이가 거의 없는 것으로 나타났다.

기본문헌수를 한 건 더 추가하여 4건을 사용했을 경우 3건을 사용했을 때와 같이 정확률 감소 현상이 계속되는지를 살펴보기 위하여 상위문헌 4건을 기본문헌으로 사용한 결과, 단어빈도 가중치 기법을 제외하고는 3건을 사용했을 때보다 정확률이 약간씩 낮아지는 것으로 나타났다. 정확률의 순으로 보면 필드 가중치 기법의 정확률이 0.7628로 가장 높았고, 다음은 단어빈도 가중치 기법과 근접거리 기법으로 각각 0.6861과 0.6842로 서로 비슷한 수준이었으며, 단어빈도·역문헌빈도 가중치 기법은 0.6749, 그리고 역문헌빈도 가중치 기법의 정확률은 가장 낮은 0.6652였다. 마지막으로 상위문헌 5건을 사용하여 검색된 문헌의 순위를 조정했을 때의 정확률은 〈표 8〉에 나타난 바와 같이 4건을 사용했을 때보다 대체적으로 약간씩 낮아져 정확률 감소현상은 계속되는 것으로 나타났다. 단어빈도 가중치 기법의 경우에도 5건을 사용했을 때는 다시 정확률이 낮아지기 시작하여 4건을 사용했을 때의 증가는 일시적인 현상인 것으로 볼 수 있다.

이상과 같이 용어 가중치를 이용한 1차 검색결과로 출력된 상위 문헌과 나머지 문헌간의 유사도를 측정하여 검색된 문헌의 순위를 조정하는 실험에서는 필드 가중치 기법을 제외한 나머지 기법들은 상위문헌 2건을 기본문헌으로 사용하여 나머지 문헌들과의 유사도를 측정했을 때 정확률이 가장 큰 폭으로 증가하였고 그 이후부터는 서서히 감소였으므로, 순위 조정 과정에서는 상위문헌 2건을 사용할 때 가장 효과적인 것으로 밝

혀졌다. 다만 필드 가중치 기법의 경우는 상위문헌 3건을 사용했을 때 정확률이 가장 큰 폭으로 증가했다가 그 이후부터 감소하기 시작하여 상위문헌 3건을 사용했을 때 가장 검색효율이 높은 것으로 나타났다.

5.2.2 가중치 기법에 따른 성능변화의 특성

순위 조정 실험에서는 순위 조정에 사용된 기본문헌수가 동일하여도 정확률 증가율은 용어 가중치 기법에 따라 달라지는 특성을 보여주었는데 이러한 현상은 순위 조정에 사용된 기본문헌수 및 기본문헌 중 적합문헌이 차지하는 비율과 관련이 있는 것으로 분석된다.

순위 조정 실험결과 대체적으로 상위문헌 2건을 사용했을 때 정확률이 가장 높아졌다가 그 이후에는 점점 낮아지는 현상을 보였는데, 근접거리 가중치 기법, 역문헌빈도 가중치 기법, 그리고 단어빈도 · 역문헌빈도 가중치 기법이 이와같은 일반적인 양태를 따르는 것으로 나타났다.

그러나 단어빈도 가중치 기법은 상위문헌 1건을 사용했을 때 정확률이 가장 높았고 이어서 점차로 감소되다가 4건을 사용했을 때 일시적으로 상승하는 불규칙적인 현상을 보이고 있는데 이러한 현상은 순위조정에 사용된 기본문헌의 특성과 관련이 있는 것으로 분석된다. 〈표 1〉에 나타나 바와 같이 단어빈도 가중치 기법을 사용한 1차 검색결과 상위 첫 번째 문헌이 적합문헌인 경우는 탐색문 26개중 25개로 다른 기법들에 비하여 상대적으로 높았으나, 두 번째 문헌과 세 번째 문헌이 적합문헌인 경우는 각각 16개, 12개로 급격하게 감소하였다. 따라서 상위문헌 1건만을 기본문헌으로 사용하여 순위 조정에 사용했을 때는 정확률 증가의 폭이 컸으나, 두 번째 문헌과 세 번째 문헌도 기본문헌으로 사용했을 때는 기본문헌 중 적합문헌수가 상대적으로 적어 정확률이 낮아지는 것을 볼 수 있다. 그러나 네 번째 문헌이 적합문헌인 경우는 다른 기법에 비해 상대적으로 많아 상위문헌 4건을 사용했을 때는 정확률이 일시적으로 상승하였다.

필드 가중치 기법에서 상위문헌 1건을 기본문헌으로 사용했을 때는 다른 기법들과 차이가 적었으나 상위문헌 3건을 사용했을 때는 정확률이 가장 높아져 다른 기법들과의 차이가 현저해지는 현상도 순위 조정에 사용된 기본문헌의 특성으로 설명될 수 있다. 필드 가중치 기법을 사용하여 1차 검색을 실시한 결과 상위 첫 번째 문헌이 적합문헌인 경우는 다른 기법들과 비슷하였으나, 두 번째 문헌부터는 다른 기법들에 비해 상대적으로 적합문헌수가 많아 기본문헌 중 적합문헌이 차지하는 비율이 커지기 시작해 세 번째 문헌에서는 그 차이가 가장 크기 때문인 것으로 해석된다.

가중치 기법별로 순위 조정 실험의 정확률 변화를 살펴본 결과 순위 조정 실험의 성능은 순위조정에 사용되는 기본문헌 중 적합문헌이 차지하는 비율에 따라 크게 영향을 받는 것으로 나타났다. 필드 가중치를 제외한 기법들은 대체적으로 검색효율이 가장 큰 폭으로 증가할 때가 상위문헌 2건을 기본문헌으로 사용하였을 때였는데 이때 기본문헌 중 적합문헌이 차지하는 비율은 75%~78.85%였고, 상위문헌 3건을 사용했을 때는 증가율이 감소하기 시작했는데 이때의 비율은 67.95%~70.52%였다. 그러나 필드 가중치 기법의 경우는 상위문헌 3건을 사용했을 때도 정확률이 높아져 가장 큰 증가율을 보였는데 이때 기본문헌 중 적합문헌이 차지하는 비율은 79.49%였다. 따라서 순위 조정 실험에서 가장 검색효율을 증진시킬 수 있는 기본문헌수는 순위조정에 사용되는 기본문헌 중 적합문헌이 차지하

〈표 9〉 순위 조정 과정에 필요한 추가 소요시간(초)

가중치 구분 / 기본문헌수	단어빈도	역문헌빈도	단어빈도·역문헌빈도	필드	근접거리	평균
1개 문헌	0.866	0.894	0.882	0.895	0.881	0.884
2개 문헌	0.997	0.972	0.986	0.966	0.998	0.984
3개 문헌	1.078	1.059	1.085	1.082	1.071	1.075
4개 문헌	1.162	1.157	1.153	1.154	1.161	1.157
5개 문헌	1.240	1.253	1.251	1.234	1.236	1.243

는 비율이 약 75% 정도가 될 때인 것으로 볼 수 있다.

각 기법의 성능을 비교해보면 실험된 다섯 가지 가중치 기법 가운데 가장 효율이 우수한 가중치 기법은 필드 가중치 기법이었다. 필드 가중치 기법은 순위 조정에 사용된 문헌수에 따라 정도의 차이는 있지만 다섯 번의 순위 조정 결과 모두 정확률이 가장 높게 나타났다. 근접거리 가중치 기법은 순위 조정 실험에서 필드 가중치 기법 다음으로 검색효율이 우수하였다. 다섯 번의 실험 중에서 세 번의 경우는 정확률이 전체 기법 중 두 번째로 우수하였고, 두 번의 경우는 세 번째로 우수하여 통계적 기법인 단어빈도, 역문헌빈도, 단어빈도·역문헌빈도 가중치 기법보다 우수한 것으로 나타났다. 통계적 기법들 중에서 단어빈도 가중치 기법과 단어빈도·역문헌빈도 가중치 기법은 순위 조정 실험에서는 정확률들이 차지하는 순위들이 서로 비슷하여 두 기법간의 검색효율은 비슷한 수준이었다. 역문헌빈도 가중치 기법은 다섯 번의 실험 모두 정확률이 가장 낮아 다른 기법에 비해 검색효율이 가장 떨어지는 것으로 밝혀졌다.

5.2.3 소요시간

순위 조정 과정에 추가로 소요된 시간은 〈표 9〉와 같다. 추가 소요 시간은 가중치 기법에 따라 약간씩의 차이는 있으나 상위문헌 1건을 기본문헌으로 사용했을 때의 평균 추가 소요시간은 0.884초였고, 2건을 사용했을 때는 0.984초였다. 그러나 3건을 사용했을 때는 1.075초, 4건을 사용했을 때는 1.175초, 5건을 사용했을 때는 1.243초가 추가로 소요되어 기본문헌수를 1개씩 증가시킬수록 추가로 소요되는 시간은 증가하였으나 증가의 폭은 점점 더 줄어들었다.

검색 소요 시간은 검색효율과 함께 검색성능 평가의 중요한 한 요소로 검색효율이 높다 하더라도 검색시간이 많이 소요되면 시스템의 효율성이 떨어지는 것으로 볼 수 있다. 그러나 본 실험에서 사용된 순위 조정 기법은 상위문헌 2건을 사용했을 경우 검색효율이 증가되면서도 추가로 소요되는 검색시간은 1초 미만이어서 검색시간이 시스템의 성능 저하에는 크게 영향을 미치지 않

는 것으로 밝혀졌다.

6 결 론

한국어 P-norm 검색시스템에서의 용어 가중치를 이용한 문헌 순위화 기법 및 검색된 문헌의 순위 조정 실험결과는 다음과 같다.

첫째, 비통계적 기법 중 문헌의 구조를 이용하는 필드 가중치 기법은 실험된 다섯 가지 가중치 기법 중 가장 우수하였다. 필드 가중치 기법은 통계적 기법 중 검색효율이 가장 높은 단어빈도·역문헌빈도 가중치 기법보다 재현율은 4.47%, 정확률은 6.15% 높았다. 또한 필드 가중치 기법에서의 적합문헌의 평균순위는 단어빈도·역문헌빈도 가중치 기법의 평균순위보다 3.35 순위가 빨라져 적합문헌들이 약 3건 앞서 출현하는 것으로 밝혀졌다. 따라서 필드 가중치 기법은 적용하기가 용이하면서도 검색성능은 통계적 기법들보다 우수한 것으로 밝혀졌다.

둘째, 비통계적 가중치 기법 중 단어의 위치를 이용한 근접거리 가중치 기법은 검색효율에 있어서 통계적 기법들과 거의 차이가 없는 것으로 밝혀졌다. 재현율은 단어빈도·역문헌빈도 가중치 기법보다 근소한 차이로 높았으나, 정확률은 단어빈도·역문헌빈도 가중치 기법보다 근소한 차이로 낮아졌다. 그리고 적합문헌의 평균순위에 있어서도 거의 차이가 없어 근접거리 가중치 기법은 통계적 기법들과 큰 차이가 없는 것으로 밝혀졌다.

셋째, 비통계적 기법과 통계적 기법을 결합하였을 때는 필드 가중치와 단어빈도·역문헌빈도 가중치를 결합하였을 때 가장 성능이 우수하였다. 두 가지 기법을 결합하였을 때는 필드 가중치만을 사용했을 때보다 재현율은 2.28%, 정확률은 2.0% 높았고, 단어빈도·역문헌빈도 가중치 기법만을 사용했을 때보다 재현율은 7.35%, 정확률은 8.15% 높아졌다. 근접거리 가중치와 단어빈도·역문헌빈도 가중치를 결합하였을 때는 근접거리 가중치만을 사용했을 때에 비해 재현율과 정확률 모두 1% 미만의 차이를 보였고, 단어빈도·역문헌빈도 가중치를 사용했을 때에 비해서도 마찬가지로 거의 차이가 없었다.

넷째, 2단계 실험으로 검색된 상위 문헌과 나머지 문헌간의 유사도를 측정하여 문헌간의 유사도값 순으로 순위를 재정렬하는 순위 조정 실험은 검색효율을 증진시키는 것으로 밝혀졌다. 순위 조정 실험에서는 유사도 측정에 사용되는 기본문헌수에 따라 검색성능이 달라졌는데 전체적으로 볼 때 검색된 문헌 중 상위문헌 2건을 이용했을 때 검색효율이 가장 큰 폭으로 증가하였으며, 상위문헌 3건 이상을 사용했을 때부터는 검색효율의 증가폭이 서서히 감소하는 현상을 보였다. 상위문헌 2건을 기본문헌으로 사용하여 순위를 조정했을 때는 순위 조정 이전에 비해 정확률이 기법에 따라 8.78~11.18% 증가되었다. 그러나 필드 가중치 기법에서는 상위문헌 3건을 기본문헌으로 사용했을 때 가장 검색효율이 높아져 순위 조정 이전에 비해 10.61% 증가하였다.

위와 같은 용어 가중치를 이용한 문헌 순위화 기법 및 문헌간의 유사도를 이용한 순위 조정 실험을 통하여 본 연구에서 이룬 성과는 다음과 같다.

첫째, 불논리 탐색문과 도치색인 파일을 사용하는 현재의 검색시스템을 크게 수정하거나 새로운 시스템 구성 요소를 추가하는 과정 없이 용이하게 적용할 수 있도록 시스템을 구축하면서 동시에 검색효율을 향상시킬 수 있는 문헌 순위화

기법을 제시하였다.

둘째, 최초의 탐색질문을 작성하는 것 외에는 이용자의 부담이 전혀 요구되지 않는 상태에서 시스템이 자동적으로 색인어 정보를 이용하여 1차로 검색된 문헌의 순위를 재정렬함으로써 검색 효율을 향상시킬 수 있다.

본 연구의 제한점 및 향후 연구를 위한 사항은 다음과 같다.

첫째, 본 실험의 범위가 한정되어 있다는 것이다. 본 실험은 여러 데이터베이스를 대상으로 실시되지 않고 한 개의 데이터베이스를 사용하였다. 그리고 사용된 실험 문헌 집단의 크기가 작고, 실험에 사용된 탐색문도 31개였다. 따라서 본 연구에서 발견한 결과를 일반화시키기 위해서는 특성이 다른 여러 개의 데이터베이스를 상대로 실험이 확대되어야 하며, 실험규모도 크게 확대할 필요가 있다.

둘째, 순위 조정 실험에서 문헌간의 유사도값을 측정할 때 동의어나 관련어 관계를 처리하지 않았다는 점이다. 탐색어의 동의어나 관련어는 탐색문에서 처리하였으나, 색인어 비교 시에는 완전일치 방법을 사용하였기 때문에 비교되는 문헌이 동의어를 색인어로 가진 경우에는 같은 주제를 다룬 문헌들일지라고 유사도값이 적어지는 결과를 낳게 되었다. 따라서 동의어나 관련어를 색인어로 포함하고 있는 경우 유사도값을 높여줄 수 있는 처리 과정을 추가할 필요가 있다.

참 고 문 헌

김성혁 외. 1994. 자동색인기 성능시험을 위한 Test Set 개발. 『정보관리학회지』, 11(1): 81-101.

이준호. 1993. 시소러스의 연관성 정보를 이용한 문서의 순위 결정 방법. 『정보관리학회지』, 10(2): 3-22.

이효숙. 1993. 『적합성 가중치 검색 및 P-norm 검색에 관한 연구』. 박사학위논문, 이화여 자대학교 대학원, 도서관학과.

정영미. 1993. 『정보검색론』. 개정판. 서울: 구미무역.

Bookstein, A. 1978. "On the perils of merging Boolean and weighted retrieval systems." *Journal of the American Society for Information Science*, 29(3): 156-157.

Bookstein, A. 1981. "A Comparison of two systems of weighted Boolean retrieval." *Journal of the American Society for Information Science*, 32(4): 275-279.

Croft, W. B. and D. J. Harper. 1979. "Using probabilistic models of document retrieval without relevance information." *Journal of Documentation*, 35(4): 285-295.

Fox, E. A. 1983. *Extending the Boolean and vector space models of information retrieval with P-norm queries and*

multiple concept types. Ph.D. Dissertation, Cornell University.

Frakes, W. B. and R. Beaza-Yates ed. 1992. *Information retrieval: data structures & algorithms*. Englewood Cliffs, N. J.: Prentice Hall.

Griffiths, A., H. C. Luckhurst, and P. Willett. 1986. "Using interdocument similarity information in document retrieval systems." *Journal of the American Society for Information Science*, 37(1): 3-11.

Keen, E. M. 1991. "The Use of term position devices in ranked output experiment." *Journal of Documentation*, 47(1): 1-22.

Keen, E. M. 1994. "Designing and testing an interactive ranked retrieval system for professional searchers." *Journal of Information Science*, 20(6): 389-398.

Noreault, T., M. McGill, and M. Koll. 1977. "Automatic ranked output from Boolean searches in SIRE." *Journal of the American Society for Information Science*, 28(6): 333-341.

Oddy, R. N. et al. 1981. *Information retrieval research*. London: Butterworths.

Rada, R. and E. Bicknell. 1989. "Ranking documents with a thesaurus." *Journal of the American Society for Information Science*, 40(5): 304-310.

Radecki. T. 1988. "Probabilistic methods for ranking output documents in conventional Boolean retrieval systems." *Information Processing & Management*, 24(3): 281-301.

Robertson, S. E. 1977. "Theories and models in information retrieval." *Journal of Documentation*, 33(2): 126-148.

Sallton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley.

Sallton, G. ed. 1971. *The SMART retrieval system*. Englewood, N. J.: Prentice-Hall.

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

Salton, G. and E. Voorhees. 1985. "Automatic assignment of soft Boolean operators". *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery*, N.Y.: 54-69.

Salton, G., E. A. Fox, and H. Wu. 1983. "Extended Boolean information retrieval." *Communications of the ACM*, 26(11): 1022-1036.

Savoy, J. 1997. "Ranking schemes in hybrid Boolean systems: a new approach." *Journal of the American Society for Information Science*, 48(3): 235-253.

Smith, M. E. 1990. *Aspects of the P-norm model of information retrieval: syntactic query generation, efficiency,*

and theoretical properties. Ph.D. Dissertation, Cornell University.

Sparck Jones, K. 1972. "A Statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation*, 28(1): 11-20.

Wade, S. J. and P. Wilett. 1989. "SIBRIS: the sandwich interactive browsing and ranking information system." *Journal of Information Science*, 15: 249-260.