용어 분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정

(Query Term Expansion and Reweighting using Term-Distribution Similarity)

김 주 연 † 김 병 만 †† 박 혁 로 †††

(Ju Youn Kim) (Byeong Man Kim) (Hyuk Ro Park)

요 약 본 논문에서는 사용자의 적합 피드백을 기반으로 피드백 문서들에서 발생하는 용어들과 초기 질의와의 관련 정도를 이용하여 용어의 가중치를 산정하는 방법에 대하여 제안한다. 피드백 문서들에서 발생하는 용어들 중에서 불용어를 제외한 모든 용어들을 질의로 확장될 수 있는 후보 용어들로 선택하고, 피드백 문서들에서 발생 빈도 유사성을 이용하여 초기 질의에 대한 후보 용어의 관련 정도를 산정하며, 피드백 문서들에서의 가중치와 관련 정도를 결합하여 후보 용어들의 가중치를 산정 하였다. 본 논문에서는 성능을 평가하기 위하여 KT-set 1.0과 KT-set 2.0을 사용하였으며, 성능의 상대적인 평가를 위하여 질의어를 확장하지 않은 방법, Dec-Hi방법들을 정확률-재현율을 사용하여 평가 하였다.

Abstract We propose, in this paper, a new query expansion technique with term reweighting. All terms in the documents feedbacked from a user, excluding stopwords, are selected as candidate terms for query expansion and reweighted using the relevance degree which is calculated from the term-distribution similarity between a candidate term and each term in initial query. The term-distribution similarity of two terms is a measure on how similar their occurrence distributions in relevant documents are. The terms to be actually expanded are selected using the relevance degree and combined with initial query to construct an expanded query. We use KT-set 1.0 and KT-set 2.0 to evaluate performance and compare our method with two methods, one with no relevance feedback and the other with Dec-Hi method which is similar to our method, based on recall and precision.

1. 서 론

정보 검색 시스템은 사용자의 질의 요구에 가장 적합한 문서들을 제공하기 위하여, 대용량의 데이타로부터 주어진 시간내에 원하는 정보를 발견할 수 있도록 도와주는 시스템이다[1]. 그러므로, 정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법(Ranking)을 적용하는 것이며, 문서 순위 결정 방법은 문서와 질의 사이의 관련 정도를 나타내는 유사도(Similarity)를 계산하고, 계산된 유사도에 따라 문서에 순위가 부여된다. 이때, 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다.

정보 검색 시스템에서는 질의로 사용한 용어가 문서에서 발생할 경우에만 검색이 가능하며, 질의어의 의미가 유사한 동의어를 이용하여 문서를 작성할 경우 검색을 할 수 없는 용어 불일치 문제라는 가장 기본적인 문제를 가지고 있다. 간단한 예로, 사용자가 질의를 기술할 때 문서에서 작가가 어떤 개념을 기술하기 위하여 사용한 용어와 의미가 유사한 용어를 사용하여 질의를 한다고 가정하면, 정보 검색 시스템은 질의어로 주어진

· 본 연구는 한국과학재단 박사후 해외연수 프로그램(1997년도)의 지원을 받아 수행되었음

† 비 회 원 : 금오공과대학교 전자계산기공학과
jykim@cespc1.kumoh.ac.kr
†† 종신회원 : 금오공과대학교 컴퓨터공학부 교수
bmkim@cespc1.kumoh.ac.kr
††† 비 회 원 : 전남대학교 컴퓨터정보학부 교수
hyukro@chonnam.ac.kr

논문접수 : 1999년 8월 25일
심사완료 : 1999년 12월 24일

용어가 문서에 발생하지 않음으로 질의에 적합 문서임에도 검색이 되지 않는 문제가 발생한다. 이와 같은 용어 불일치 문제를 해결하기 위한 가장 간단한 방법은 질의어를 다량으로 입력함으로써 질의어와 적합 문서에서의 용어가 일치할 기회를 높이는 것이다. 그러나 많은 정보 검색 시스템들에서의 질의어는 매우 짧은 경우가 대부분으로 통계에 의하면 World-Wide-Web을 통한 정보 검색 시스템에서의 평균 질의어 길이는 2Word라는 사실이 밝혀졌다[2].비록 이러한 통계가 정보 검색 시스템 중에서 극단적인 예일지는 모르지만, 대부분의 정보 검색 시스템에서는 질의어가 길지 않고, 용어 불일치 문제를 해결해야 할 필요가 있음을 나타낸다.

용어 불일치 문제를 해결하고 검색 성능을 향상시키기 위하여 제안된 방법이 질의어 수정이다. 질의어는 문서내 용어들이나 혹은 질의에서 발생하는 용어들과 의미가 유사한 용어를 사용하여 질의어를 확장하고 관련 문서에서 용어들이 일치할 기회를 증가시킨다. 따라서 이러한 용어가 일치할 기회를 증가시키기 위한 용도로 사용할 수 있는 것이 시소러스이다. 그러나 실험 결과에 의하면, 비록 확장될 용어들이 주제의 전문가에 의해 선택되었다 하더라도 실제 검색에서는 검색 효율을 크게 향상시키지는 못하였고, 검색되는 전체 문서에서 자동으로 구문을 분석하여 질의를 확장하는 방법이 수동으로 구축된 시소러스를 이용하는 방법보다 더욱 효과적이라는 사실이 밝혀졌다[3].

질의어 수정의 초기 연구는 문서들에서 동시에 발생하는 용어들을 분류하여 질의어로 확장하는 방법을 사용한 Sparck Jones에 의해 수행되어졌으며[4], 사용자에 의한 적합 피드백을 기반으로 적합 문서내에서 확장될 용어를 선택하는 방법은 1971년 Rocchio가 용어 가중치 재산정과 질의어 확장을 조합한 질의 수정에 관한 실험 결과를 발표하면서부터이다[5]. 따라서 현재까지의 적합 피드백을 기반으로 질의어를 수정하는 방법을 종합해 보면 1)질의 용어 가중치 재산정 방법, 2)질의 용어 확장 방법, 3)질의 용어 가중치 재산정 및 질의 용어 확장 방법으로 구분 할 수 있다.

본 논문에서는 순위 부여 모델들중에서 벡터 스페이스 모델을 기반으로 사용자의 적합 피드백을 이용한 용어 가중치 산정 및 질의어 확장을 위하여 초기 질의에 대한 피드백 문서들에서 용어들의 발생 빈도 분포를 이용하여 확장될 용어와 원 질의어간의 관련 정도를 산정하는 방법과, 산정된 관련 정도를 이용하여 확장 용어의 가중치를 산정하는 방법에 대하여 제안한다. 사용자에 의해 판단된 초기 질의에 대한 피드백 문서들을 근거로 피드백 문서들내에서 발생하는 용어들을 질의어로 확장될 수 있는 후보 용어들로 선택하고, 피드백 문서들에서 후보 용어와 원 질의어들의 피드백 문서내 발생 빈도(TF) 분포를 이용하여 원 질의어에 대한 후보 용어의 관련 정도를 산정하며, 피드백 문서내에서 후보 용어의 가중치와 관련 정도를 결합하여 확장될 후보 용어의 가중치를 산정하였다. 또한 후보 용어의 관련 정도를 기준으로 관련 정도가 높은 용어들 순으로 확장될 용어의 수를 제한하여 성능에 미치는 영향을 실험하였다. 그리고 상대적인 성능 평가를 위하여 질의를 확장하지 않은 방법(BaseLine), 적합 피드백에서 많이 사용되는 Ide Dec-Hi방법등 2가지 방법을 정확률-재현율을 이용하여 성능을 비교 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 적합 피드백에 대한 기존 연구들을 분석하고 적합 피드백 분야에서 많이 사용되고 있는 Ide Dec-Hi 방법의 문제점들을 지적하며, 3장에서는 본 논문에서 제안하는 적합 문서내에서의 발생 빈도 분포를 이용한 원 질의어와의 관련 정도 산정과, 관련 정도를 결합하여 질의어로 확장될 후보 용어의 가중치를 산정하는 방법에 대하여 설명한다. 4장에서는 재현율과 정확율을 사용하여 제안하는 방법과 기존 방법의 성능을 다양한 방법을 이용하여 비교 평가하였다. 마지막으로 5장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구 및 문제점

2.1 선행 연구

적합 피드백을 이용하여 검색 효율을 향상하고자 하는 최근의 많은 노력들 중에서 질의어 수정 방법은 크게 3가지 분류로 구분할 수 있다.

1) 질의 용어 가중치 재산정 방법: 사용자의 질의와 관련이 있는 적합 문서와 관련이 없는 비적합 문서들에서 발생하는 질의 용어들의 중요도 분포를 계산하여 질의 용어의 가중치만을 재산정하고 질의 용어는 확장하지 않는 방법이다[6]. 이러한 방법은 1976년 Robertson과 Sparck Jones에 의해 처음 사용되었으며 적합, 비적합 문서에서의 질의 용어 분포를 바탕으로 용어의 가중치를 산정하였다[7]. 그러나 이러한 방법은 검색 시스템의 가장 기본적인 문제인 용어 불일치 문제를 해결할 수 없다.
2) 질의 용어 확장 방법: 수동으로 구성된 시소러스를 이용하거나 자동으로 생성된 용어를 질의어에 포함하여 검색 시 질의어를 확장하는 방법[8]으로

서 이러한 연구들의 대부분은 용어와 용어간의 관련성을 이용하거나 클러스트링 방법을 이용하여 진행되어지고 있다.

3) 질의 용어 가중치 재산정 및 질의 용어 확장 방법: 질의 용어 가중치 재산정 방법과 질의 용어 확장 방법을 결합한 방법으로 적합 피드백 기반의 검색에서 현재 가장 많이 연구되고 있는 분야이며, 실험에서도 3가지 방법중 평균 정확도 면에서 가장 우수한 성능을 나타냈다[9].

Jinxi와 Croft는 Local분석과 Global분석을 조합한 새로운 방법인 Local Context분석 방법을 사용하여 질의어 확장을 시도하였다[10]. Global분석 방법은 전체 문서를 일정한 구절별로 분석하여 전체 문서중 출현 빈도수가 대단히 높은 용어와 출현 빈도수가 대단히 낮은 용어를 제외한 이외의 용어를 집합으로 구성하여 질의어로 확장하는 것을 의미한다. 또한 Local 분석은 사용자의 적합 피드백을 근거로 최상위의 문서들에 대하여 용어를 추출하는 방법이다.

Local Context 분석 방법은 Global분석 방법과 Local분석 방법을 조합한 형태의 분석 방법으로 사용자의 적합 피드백을 근거로 최상위의 문서들에 대하여 Global분석에서와 같이 일정한 구절별로 문서를 분석하여 확장될 용어를 선택하고, 역문헌 빈도수(IDF), 적합 문서내 용어의 발생빈도수(TF)를 결합한 수식을 사용하여 원질의어-용어의 관련정도를 산정하고, 관련 정도순위로 기준으로 용어들을 확장하였다. 이 실험에서는 Global 분석 방법과 Local분석 방법, Local Context분석 방법을 비교하였으며, 실험 결과는 Global 분석이 가장 낮으며, Local분석은 부분적으로 성능 향상을 이루었으며, Local context 분석은 4개의 test set에서 최고 53.8%, 최저 23.5%의 성능 향상을 보였다.

Salton과 Buckley는 6개의 실험 문서들을 통해서 Ide Dec-Hi, Ide Regular, Rocchio방법들을 실험하였다[9]. 이와 같은 세 가지 방법의 기본적인 연산 절차는 문서 벡터와 원래의 질의 벡터를 병합하는 것이다. 이것은 적합 문서들에 해당 질의어의 발생으로부터 가중치를 부가하고, 비적합 문서에 대하여 가중치를 줄여줌으로서 질의어에 자동적으로 가중치가 다시 부여되도록 한다. 질의어는 원래의 질의에 없었던 용어에 대하여 적합 문서에서 발생한 것인지, 아니면 비 적합 문서에서 발생한 것인가의 판단에 따라 양의 가중치와 음의 가중치가 부여된다. 또한 음의 가중치를 가지는 용어는 질의어로 확장되지 않는다.

Ide dec-hi방법은 사용자에게 보여진 집합 내에서 검색되어진 비적합 문서들 전체에 대한 평가 대신에 적합 평가에 대한 최상위의 비적합 문서를 사용하며, Rocchio 방법은 적합과 비적합 문서의 적합 정도의 조정을 허락하였다. Salton과 Buckey의 실험 결과는 여섯 개의 실험문서 집합에서 거의 차이가 없었지만, Ide dec-hi 방법을 사용했을 때 가장 좋은 결과를 얻었었으며, Rocchio방법에서는 적합, 비적합 문서에 0.75, 0.25의 가중치를 부여하여 최상의 결과를 생성하였다. 또한 이 실험에서는 5개의 실험 문서에서 평균 60%-90%의 정확도가 증가하였다.

2.2 Ide Dec-Hi방법의 문제점

Salton과 Buckey의 실험 결과에서 가장 우수한 성능을 나타내는 Ide Dec-Hi 방법에서는 확장될 용어들의 가중치를 부여할 때 초기 질의어와의 관련성, 질의어로서의 중요성을 반영하지 못하는 문제점을 가지고 있다. 즉, Dec-Hi방법에서는 적합, 비적합 문서내의 발생 빈도(TF)와 전체 문헌에서의 역문헌 빈도수(IDF)만을 이용하여 용어의 가중치를 산정하게 되므로 적합 문서들에서 TF가 높은 용어는 높은 가중치를 가지게 되고, 이러한 결과는 질의어로서 중요하지 않은 용어들도 단지 적합 문서내에서만 자주 발생하게 되면 높은 가중치를 부여받게 된다는 것을 의미한다.

Ide Dec-Hi 방법에서의 문제점은 질의어 수정을 위하여 확장될 용어들을 자동으로 선택할 경우 더욱 분명하게 나타난다. 자동으로 확장될 용어를 선택하는 경우에는 질의어로서 중요하지 않은 용어들도 많이 선택될 수 있으며, 이러한 용어들은 문서에서의 발생 빈도가 높다는 특성이 있다. 이때, Ide Dec-Hi 방법을 이용하여 가중치를 산정할 경우 질의어로서 중요하지 않은 일반적인 용어들에 높은 가중치를 부여하는 경우가 발생하며, 이러한 결과는 검색 성능을 현저히 저하 시키는 요인으로 작용할 수 있다. 비록 이러한 문제점을 보완하기 위하여 가중치 산정시 문서내의 용어 빈도수를 정규화하고 역문서 빈도수(IDF)를 이용하여 가중치를 부여하고 있지만 문제점을 해결하지는 못하고 있다.

예를 들어 "데이타베이스"라는 용어가 1.0의 역문서 빈도(IDF)값를 가지고 A 적합 문서에 1회, B 적합 문서에 1회 발생했고, "논문"이라는 용어가 0.3의 역문서 빈도값을 가지고 A적합 문서에 4회, B 적합 문서에 4회 발생할 경우 Dec-Hi방법을 이용하여 2개의 용어에 가중치를 부여할 경우 "데이타베이스"라는 용어는 빈도수를 정규화하여 가중치를 부여할 경우 0.5, 정규화하지 않은 경우 2.0의 가중치가 부여되는데 반하여 "논문"이라는 용어는 정규화 할 경우 0.6, 정규화 하지 않을 경

우 2.4의 가중치가 부여된다. 비록 이것은 극단적인 예일지는 모르지만 적합 문서내 용어의 발생만을 이용하여 가중치를 산정하는 것은 위의 예에서와 같이 중요하지 않은 용어에 높은 가중치를 부여하게 될 우려가 있다. 그러므로, 확장 용어의 가중치는 적합 문서내 발생(TF)뿐만 아니라 초기 질의어와의 관련성을 이용한 용어의 중요성을 고려하여 산정될 필요가 있음을 알 수 있다.

본 논문에서는 Salton과 Buckey의 실험 결과에서 가장 우수한 성능을 나타내는 Ide Dec-Hi 방법을 참조하여 초기 질의어와의 관련 정도를 고려하여 확장 용어의 가중치를 산정하였고, 관련 정도를 기준으로 확장 용어의 수를 제한하여 실험하였다. 또한 본 논문에서 제안하는 방법은 관련 정도를 이용하여 확장될 용어의 가중치를 산정할 때 Ide Dec-Hi방법과 유사한 방법을 이용하여 가중치를 산정하게 됨으로 제안하는 방법의 성능을 상대 평가하기 위하여 Ide Dec-Hi방법을 함께 실험하였다.

3. 용어의 관련 정도를 이용한 가중치 산정 및 질의어 확장

적합 피드백은 피드백 문서내에서 발생하는 용어들중 질의어로 확장될 용어(이하 후보 용어)를 선택하는 단계와 후보 용어에 가중치를 부여하는 단계로 구분할 수 있으며, 이 장에서는 본 논문에서 제안하는 원 질의어와 후보 용어간 관련 정도를 산정하기 위한 가정을 정의하며, 관련 정도를 이용하여 후보 용어의 가중치를 산정하는 방법에 대하여 기술한다.

3.1 후보 용어의 관련 정도 산정을 위한 기본 가정

그림 1에서는 후보 용어와 질의 용어들간의 관련 정도를 측정하는데 필요한 기본적인 가정을 나타내고 있다. 각 피드백 문서들에서 질의 용어의 발생 분포와 후보 용어들의 발생 분포를 벡터로 나타낼 때 그림에서와 같이 질의 용어와 후보 용어들간의 각 피드백 문서내에서의 발생 분포가 가장 유사한 후보 용어가 질의 용어와 가장 관련 정도가 높다는 가정이다. 그림에서는 후보 용어 A가 질의 용어와 가장 관련이 높으며 B,C,D순으로 관련 정도가 산정된다. 이러한 가정은 적합 문서가 피드백된 상황에서 특정 피드백 문서에서만 발생 빈도수(TF)가 높거나, 혹은 발생 빈도수가 대단히 낮은 용어들은 대부분 질의와 관련이 없는 개념을 설명하기 위한 용어이거나 질의어로서 중요하지 않은 일반적인 용어일 가능성이 높다는 것이며, 질의와 유사한 발생 분포를 가지는 용어는 질의어의 동의어는 아닐수 있지만, 확장 용어로서는 중요한 용어일 수 있다는 가정이다. 그러므로 피드백 문서들내에서 질의어와 유사한 발생 분포를 가지는 후보 용어에 대하여 관련 정도를 높게 산정하고, 발생 분포에 차이가 많은 용어들은 관련 정도를 낮게 산정함으로서 후보 용어의 가중치를 산정할 때 중요도를 반영할 수 있도록 하였다.

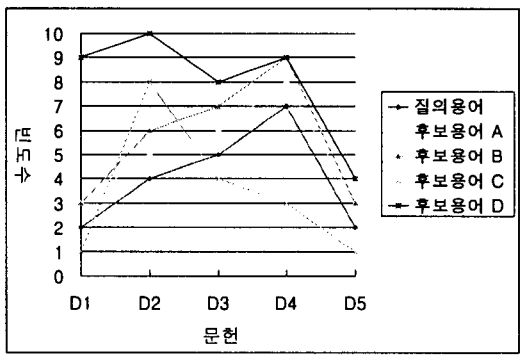

그림 1 용어 발생 분포도

3.2 후보 용어의 관련 정도 산정 및 질의어 확장을 위한 알고리즘

[단계 1] 후보 용어 집합을 생성

초기 질의어를 이용한 검색 문서중 상위 10위 이내의 문서들중에서 피드백된 문서들에 발생하는 모든 용어들을 확장될 수 있는 후보 용어들로 생성한다. 후보 용어의 선택은 저장 공간과 실행 속도를 고려하여 용어의 수를 제한할 수 있으나, 본 논문에서는 재현율을 고려하여 적합 문서들에서 발생하는 모든 용어들은 후보 용어들로 선택하였다.

[단계 2] 후보 용어-원질의어간의 관련 정도를 산정

[단계1]에서 생성된 후보 용어들과 원 질의 용어들간의 관련 정도를 산정하기 위하여 본 논문에서는 각각의 후보 용어들과 원 질의어들을 다음과 같은 벡터로 나타낸다.

$$t_i = (tfd_{i1}, tfd_{i2}, \ldots, tfd_{in})$$

$$q_j = (qfd_{j1}, qfd_{j2}, \ldots, qfd_{jn})$$

tfd_i : 피드백 문서내 후보 용어 i의 발생 빈도수

qfd_j : 피드백 문서내 원 질의 용어 j의 발생 빈도수

n : 피드백 문서의 수

각 후보 용어와 원 질의어가 벡터로 표시되면 (식 1)을 이용하여 각 피드백 문서들에서 후보 용어들과 전체 원 질의어의 관련 정도를 산정한다.

$$Rd_{ik}(Q, t_i) = 1 - \log_{10}(\sqrt{\sum_{j=1}^{m}(|qfd_{jk} - tfd_{ik}|)}) \quad (1)$$

$Rd_{ik}(Q, t_i)$: 피드백 문서 k에서 후보 용어 t_i와 원 질의어들간의 관련 정도

qfd_{jk} : 피드백 문서 k에서 원 질의 용어 j의 빈도수

tfd_{ik} : 피드백 문서 k에서 후보 용어 i의 빈도수

m : 원 질의 용어의 수

(식 1)은 각각의 피드백 문서내에서 후보 용어와 원 질의어들의 발생 빈도수 차이만을 이용하여 관련 정도를 산정하고 있다. 이와 같은 식은 3.1절의 기본 가정을 수식화한 것으로 피드백된 문서내에서 발생 빈도가 원 질의어와 유사할 경우 높은 관련 정도를 가지며, 빈도수 차이가 많을 경우 낮은 관련 정도를 가지도록 하였다. 예를 들면, 피드백 문서 d1에서 원 질의어 q1, q2, q3가 각 3, 4, 1회 발생하고, 후보 용어 t1이 2회 발생할 경우, d1에서의 원 질의어와 후보 용어의 관련 정도는 아래와 같이 산정되어 0.7이 된다.

$$Rd_{11}(Q, t_1) = 1 - \log_{10}(\sqrt{1+2+1}) = 1 - 0.30 = 0.7$$

이와 같은 관련 정도 산정에는 벡터 스페이스 모델에서 질의어와 문서의 유사도를 산정하는데 많이 사용되고 있는 코사인 측정법(Cosine measure)을 고려할 수 있으나, 코사인 측정법은 용어 발생 빈도수(TF)의 차이에 의한 유사도 차가 크지 않은 문제점이 있으므로 (식 1)과 같이 새로운 수식을 사용하였다.

[단계 3] 후보 용어의 가중치 산정

[단계 2]에서 산정된 후보 용어의 관련 정도를 이용하여 확장될 후보 용어의 가중치를 산정한다.

$$wt_i = \sum_{k=1}^{n}(wt_{ik} * Rd_{ik}) \quad (2)$$

$$wt_{ik} = freq_{ik} \times IDF_i$$

wt_i : 후보 용어 i의 피드백된 전체 문서들에서의 가중치

wt_{ik} : 피드백 문서 k에서 후보 용어 t_i의 가중치

Rd_{ik} : 피드백 문서 k에서 후보 용어 t_i와 원 질의어들간의 관련 정도

$freq_{ik}$: 피드백 문서 k에서 후보 용어 t_i의 빈도수

IDF_i : 후보 용어 t_i의 역문서 빈도수

n : 피드백된 문서의 수

(식 2)는 Ide Dec-Hi 방법의 변형으로서 각 피드백 문서내에서 후보 용어의 가중치와 원 질의어들과의 관련 정도를 결합하여 각 피드백 문서내에서의 가중치를 산정하며, 이를 전부 합산하여 전체 피드백 문서에서의 가중치를 최종적으로 산정해내고 있다. 예를 들어, 후보 용어 t1이 1.0의 IDF값을 가지고 피드백 문서 d1, d2, d3에 2, 1, 3회 발생하고, (식 1)을 이용하여 산정된 각 피드백 문서에서 관련 정도가 0.2, 0.7, 0.5라고 가정하면, 후보 용어 t1의 가중치는 아래와 같이 산정된다.

$$wt_1 = ((2.0*0.2) + (1.0*0.7) + (3.0*0.5)) = 2.6$$

이것은 직관적이고 간단한 방법으로 부가적인 연산이 필요 없이 전체 피드백 문서에서의 후보 용어 가중치를 산정할 수 있다.

[단계 4] 질의어 확장

질의어 확장은 피드백된 문서들에서 발생하는 모든 후보 용어들을 질의어로 확장한다. 원 질의를 구성하는 용어들의 집합을 Q, 그들의 가중치 벡터를 WQ라 하고, [단계 3]까지의 과정에서 생성된 후보 용어들의 집합을 T, 그들의 가중치 벡터를 WT로 표시하면, 즉,

$$Q = \{q_1, q_2, \ldots, q_n\}$$

$$WQ = (wq_1, wq_2, \ldots, wq_n)$$

wq_i : 원 질의 용어 i의 원 질의에서의 가중치

n : 원 질의 용어의 수

$$T = \{t_1, t_2, \ldots, t_m\}$$

$$WT = (wt_1, wt_2, \ldots, wt_m)$$

wt_j : 후보 용어 j의 피드백된 전체 문서들에서의 가중치

m : 후보 용어의 수

라고 하면, 최종적으로 확장된 질의어를 구성하는 용어들의 집합 P와 그들의 가중치 벡터 WP는 아래와 같이 나타낼 수 있다.

$$P = Q \cup T = \{p_1, p_2, \ldots, p_k\}$$

$$WP = (wp_1, wp_2, \ldots, wp_k)$$

이때 wp_i 는 다음과 같은 3가지 규칙을 이용하여 산정된다.

$$\text{if } (p_i \in Q) \text{ and } (p_i \in T) \text{ then } wp_i = wq_i + wt_i$$

$$\text{if } (p_i \notin Q) \text{ and } (p_i \in T) \text{ then } wp_i = wt_i$$

$$\text{if } (p_i \in Q) \text{and} (p_i \notin T) \text{ then } wp_i = wq_i$$

위의 규칙들을 예로 들어 설명하면, T, WT, Q, WQ가 아래와 같이 구성되어 있을 경우 위의 3가지 규칙을 적용하여, 확장된 후 질의어 집합 P와 가중치 벡터 WP는 다음과 같이 구성되어 진다.

$T=\{t_1, t_2, t_3\}$, $WT=(1.0, 2.0, 3.0)$

$Q=\{t_1, t_4\}$, $WQ=(3.0, 1.0)$

$P=\{t_1, t_2, t_3, t_4\}$, $WP=(4.0, 2.0, 3.0, 1.0)$

이것은 후보 용어가 원 질의어에 포함되어 있을 경우 원 질의 용어 가중치와 피드백 문서에서의 가중치를 합산하고, 만약, 원 질의어에 포함되지 않을 경우에는 피드백 문서에서의 가중치만을 가지고 확장된다. 또한 원 질의 용어가 후보 용어가 아닐 경우에는 초기 가중치 값을 가진다.

4. 성능 평가

4.1 성능 평가 환경 및 자료

본 논문에서 제안한 방법은 Sun SPARC Classic상에서 검색 엔진을 구성하였으며, 문서의 색인 및 질의 용어 추출을 위하여 한성대학교 정보전산학부 강승식 교수팀의 HAM 4.0 형태소 분석기를 사용하였다. 또한 색인 정보를 저장하기 위한 Database는 Informix를 사용하여 구현하였다.

성능 평가에 사용된 실험 자료는 한국어 테스트콜렉션인 KT-set 1.0[11]과 KT-set 2.0[12]을 사용하였다. KT-Set 1.0에는 정보과학회논문지, 1993 한국정보과학회 학술발표대회논문집, 정보관리학회지에 수록된 논문들로 구성된 1,053개의 문서들과 30개의 질의가 포함되어 있다. 입력된 모든 문서는 국문 및 영문 저자, 서명, 서지 사항, 초록, 분류 번호, 색인어 등 18개의 항목을 지니고 있으며, 각 질의에 대한 적합 문서들이 제시되어 있고, 질의어 하나의 평균 적합 문서의 수는 14개이다. 또한 KT-Set 2.0에서는 전자와 전산분야에 관련된 내용으로 논문초록, 전자신문, 잡지기사등 4,414건으로 구성되어 있으며, 테스트 문서와 함께 50개의 자연 언어 질의문으로 구성되어 있고, 질의문 하나의 평균 적합 문서의 수는 29개이다. 그리고 이와 같은 테스트콜렉션에서 <Title>부분과 <Abstract>부분만을 추출하여 색인 정보로 사용하였다.

4.2 실험 방법 및 평가 방법

본 논문에서는 실험을 수행하기 위하여 원 질의어를 이용한 검색 결과에서 상위 10위 내에 검색된 문서들중 질의에 대한 적합 문서들을 피드백된 문서로 사용하였다. 사용자는 초기 검색 문서중 전체 문서를 검색하지 않고 상위의 몇 개의 문서만을 검색하여 적합 피드백하게 되는 경우가 대부분이므로 상위 10위 내의 문서만을 대상으로 하였다. 또한 확장되는 후보 용어의 수는 전체 후보 용어의 수를 백분율로 나누어 10-100%까지 변화하여 수행하였다. 이때, 확장되는 용어의 선택을 위하여 Dec-Hi 방법에서는 역문서 빈도수(IDF)를 기준으로 변화하였고, 본 논문에서 제안하는 방법에서는 (식 1)을 이용하여 산정된 후보 용어의 관련정도를 기준으로 확장되는 용어의 수를 변화하였다. 그리고 검색 효율을 평가하기 위하여 보간 기법을 이용한 고정된 재현율에 대한 정확율을 계산하고, 재현율 0.0 - 1.0까지 11개의 재현율에서의 평균 정확율을 사용하였다.

재현율(Recall)은 전체 적합 문서들중 검색된 적합 문서의 비율을 의미하고, 정확율(Precision)은 검색 문서들에서 사용자가 원하는 적합 문서의 검색 비율을 의미한다. 또한 본 논문에서는 적합 피드백 환경에서 원 질의어와의 성능을 정확하게 비교하기 위하여 Standard 평가 방법 대신에 Residual Collection 평가 방법을 사용하였다. Standard 평가 방법은 각 질의문의 11개 재현율에서 평균 정확도를 이용하여 검색 효율을 평가하는 방법이다. 이 방법은 피드백에 이용된 문서내의 용어를 질의어로 확장하게 되므로 질의어 확장후 재 검색을 수행하게 되면 피드백된 문서들의 검색 순위가 상승하게 되므로 검색 효율이 향상된 것 처럼 보일 수 있으므로 실제 새로이 검색된 적합 문서들에 대한 검색 향상률을 나타낼수 없다. 그러므로 실제 적합 피드백에 의해 새로운 적합 문서들이 검색되는 효과를 평가하기 위하여 본 논문에서는 Residual collection 평가 방법을 사용하였다.

Residual Collection 평가 방법에서는 사용자가 피드백한 문서는 전체 적합 문서 집단에서 제외하고 재 검색을 수행하는 방법으로 추가로 검색된 문서들만을 이용하여 검색 효율을 평가한다[8]. 즉, 원 질의어를 이용한 검색 결과에서 10위 이내에 검색된 문서를 제외한 문서 집단을 Residual Collection이라 하고, 원 질의어를 이용한 검색 효율과 적합 피드백을 이용한 검색 효율을 Residual Collection만을 이용하여 평가하였다. 이 방법은 원 질의어를 이용한 검색에서 상위 순위에 검색되었던 적합 문서들은 재 검색에서 제외되기 때문에 Standard 방법보다는 정확율-재현율이 낮게 나타나지만, 적합 피드백 방법으로 수정된 질의문에 의한 성능

변화를 정확하게 평가할 수 있다.

본 논문에서는 표 1에서와 같이 원 질의어를 이용한 초기 검색에서 상위10위 이내에 적합 문서가 1개 이상 포함되어 있는 질의어를 대상으로 실험을 하였으며 KT-set 1.0에서는 26개, KT-set 2.0에서는 42개의 질의어를 이용하였다. 또한 피드백되는 문서의 수는 질의어 하나에 KT-set 1.0에서는 평균 4.2개, KT-set 2.0에서는 4.5개를 사용하였으며, 전체 적합 문서 비율로 KT-set 1.0에서는 27.1%, KT-set 2.0 에서는 16.6%를 피드백 문서로 사용하였다.

표 1 실험에서의 적합 피드백 문서 수

	KT-set 1.0	KT-set 2.0
사용된 질의어 수	26	42
상위 10위내 문헌 수	108	189
전체 연관 문헌수	399	1137

4.3 실험 결과

표 2에서는 원 질의어를 이용하여 초기 검색을 수행할 경우 Standard 평가 방법과 Residual 평가 방법에서의 검색 결과를 나타내고 있다. 표에서는 초기 검색 결과에서 상위 10위 이내의 적합 문서들을 제외한 Residual 평가방법을 사용할 경우 KT-set 1.0에서 84.7% 재현율과 16.1%의 정확율을 보이며, KT-set 2.0에서는 73.9%의 재현율과 19.0%의 평균 정확율을 보이고 있다. 이때, 표에서의 평균 재현율은 실험에서 사용한 각 질의어들의 재현율을 합산한 후 평균으로 나타낸 것이다.

표 2 원 질의어 검색 결과 비교

	KT-set 1.0		KT-set 2.0	
	원질의문	Residual	Standard	Residual
0.0	0.70	0.15	0.67	0.18
0.1	0.69	0.16	0.65	0.24
0.2	0.61	0.19	0.60	0.25
0.3	0.63	0.20	0.51	0.26
0.4	0.53	0.19	0.45	0.26
0.5	0.52	0.18	0.35	0.23
0.6	0.46	0.18	0.32	0.23
0.7	0.40	0.17	0.32	0.22
0.8	0.32	0.15	0.24	0.13
0.9	0.25	0.13	0.14	0.07
1.0	0.10	0.08	0.05	0.03
평균 정확율	0.473	0.161	0.391	0.190
평균 재현율	0.897	0.847	0.811	0.739

표 3에서는 KT-set 1.0과 KT-set 2.0에서 Ide Dec-Hi 방법을 이용하여 역문서 빈도수(IDF)를 기준으로 확장 용어의 수를 제한할 경우의 검색 효율을 나타내고 있다. 표에서는 역 문서 빈도수(IDF)가 높은 용어순으로 질의를 확장할 경우 전체 용어들중 최소한 80% 이상 확장할 경우에만 성능이 향상되고 있음을 볼 수 있으며, 90%-100%까지는 성능에 변화가 적으며, 전체 용어를 확장할 경우 KT-set 1.0에서 55.3%, KT-set 2.0에서는 16.8%의 성능 향상이 이루어짐을 볼수 있다.

표 4에서는 KT-set 1.0과 KT-set 2.0에서 Ide Dec-Hi방법을 이용하여 특정한 기준이 없이 질의를 확장한 결과를 보여주고 있다. 표 4에서는 40% 이상의 용어들을 확장할 경우 원 질의어를 이용한 초기 검색과 비교하여 성능이 향상되고 있으며 KT-set 1.0에서는 최대 65.8%, KT-set 2.0에서는 최대 16.8%의 성능 향상을 보이고 있다. 또한 표 3과 표 4를 비교해 보면 특정한 기준이 없이 용어를 확장하는 방법이 역문서 빈도수를 기준으로 확장하는 방 보다 높은 성능을 나타내고 있는데 이러한 현상은 역문서 빈도수(IDF)를 기준으로 확장할 경우 검색되는 문서의 수가 너무 적은데서 기인하는 것으로 역 문서빈도수(IDF)를 고려하여 IDF가 높은 용어를 우선적으로 확장할 경우 전체 용어의 80% 이상을 확장해야만 성능을 향상시킬 수 있고, 특정한 기준이 없이 확장할 경우 50%이상을 확장해야만 성능을 향상 시킬 수 있음을 알 수 있다. 그러므로 Ide Dec-Hi 방법을 이용할 경우에는 특정한 기준을 이용하여 확장되는 용어의 수를 제한하거나, 특정한 기준을 이용하여 용어들을 확장하는 방법보다는 질의어로 확장될 수 있는 후보 용어들 전체를 특정한 기준이 없이 질의로 확장하는 것이 더욱 효과적이다.

표 5에서는 본 논문에서 제안하는 방법을 이용하여 일정한 기준이 없이 확장되는 용어의 수를 제한하여 확장한 결과를 보여주고 있다. 표에서는 KT-set 1.0에서는 최고 98.1%, KT-set 2.0에서는 33.7%의 성능 향상을 보이고 있으며, 표 4와 비교할 때 Dec-Hi방법에서는 전체 후보 용어들중 최소 40% 이상을 확장할 경우에 원 질의어를 이용한 검색 결과와 비교하여 성능 향상을 기대할 수 있지만 본 논문에서 제안하는 방법을 이용할 경우 최소 30%이상만 확장하면 성능을 향상시킬 수있고, 최고 향상 비율에서도 Dec-Hi 방법과 비교하여 현격한 성능 향상을 보이고 있다.

이러한 결과는 Dec-Hi 방법이 원 질의어와의 관련성을 고려하지 않고 피드백 문서에서의 가중치와 비 적합 문서에서의 가중치만을 이용하여 질의를 확장하게 됨으로 질의어로서 중요하지 않은 용어, 즉 일반적인 용어들에 높은 가중치를 부여하게 되는 경우가 많이 발생하는데 반하여, 본 논문에서 제안하는 방법을 이용할 경우 원 질의어와 질의어로서 확장될 수 있는 후보 용어간의

표 3 Ide Dec-Hi 방법, IDF 기준 확장 결과

KT-set 1.0 (Ide Dec-Hi, IDF 기준 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.15	0.17	0.17	0.24	0.28	0.26	0.27	0.23	0.31	0.34	0.34
0.1	0.16	0.15	0.14	0.24	0.28	0.25	0.26	0.24	0.32	0.35	0.35
0.2	0.19	0.15	0.14	0.18	0.24	0.21	0.25	0.21	0.30	0.29	0.29
0.3	0.20	0.15	0.13	0.14	0.22	0.21	0.24	0.21	0.28	0.31	0.30
0.4	0.19	0.13	0.13	0.14	0.17	0.17	0.20	0.19	0.25	0.29	0.29
0.5	0.18	0.08	0.05	0.05	0.12	0.16	0.18	0.19	0.25	0.28	0.28
0.6	0.18	0.09	0.06	0.05	0.12	0.15	0.16	0.17	0.22	0.25	0.25
0.7	0.17	0.08	0.06	0.05	0.10	0.10	0.13	0.13	0.18	0.19	0.20
0.8	0.15	0.08	0.05	0.05	0.10	0.10	0.11	0.12	0.16	0.18	0.18
0.9	0.13	0.08	0.05	0.05	0.10	0.06	0.10	0.10	0.14	0.15	0.16
1.0	0.08	0.02	0.02	0.02	0.02	0.04	0.04	0.07	0.10	0.10	0.10
정확율	0.161	0.108 (-32.9%)	0.091 (-43.5%)	0.111 (-31.1%)	0.158 (-1.9%)	0.155 (-3.7%)	0.177 (+9.9%)	0.169 (+5.0%)	0.227 (+41.0%)	0.248 (+54.0%)	0.250 (+55.3%)

KT-set 2.0 (Ide Dec-Hi, IDF 기준 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.18	0.16	0.30	0.31	0.33	0.33	0.37	0.37	0.43	0.48	0.48
0.1	0.24	0.17	0.16	0.21	0.23	0.28	0.30	0.30	0.36	0.41	0.39
0.2	0.25	0.14	0.15	0.17	0.18	0.25	0.25	0.27	0.32	0.34	0.34
0.3	0.26	0.13	0.13	0.13	0.16	0.19	0.23	0.24	0.26	0.27	0.26
0.4	0.26	0.11	0.11	0.10	0.13	0.18	0.22	0.23	0.25	0.25	0.26
0.5	0.23	0.08	0.07	0.07	0.09	0.12	0.16	0.15	0.17	0.18	0.19
0.6	0.23	0.07	0.05	0.05	0.07	0.12	0.16	0.14	0.16	0.17	0.17
0.7	0.22	0.06	0.04	0.04	0.05	0.11	0.15	0.12	0.14	0.15	0.13
0.8	0.13	0.03	0.02	0.01	0.03	0.09	0.10	0.09	0.09	0.08	0.08
0.9	0.07	0.02	0.01	0.01	0.02	0.08	0.07	0.07	0.08	0.07	0.07
1.0	0.03	0.01	0.01	0.01	0.01	0.05	0.04	0.04	0.04	0.04	0.04
정확율	0.190	0.0.089 (-53.2%)	0.094 (-50.5%)	0.102 (-46.3%)	0.117 (-38.4%)	0.163 (-14.2%)	0.186 (-2.1%)	0.185 (-2.6%)	0.210 (+10.5%)	0.222 (+16.8%)	0.222 (+16.8%)

표 4 Ide Dec-Hi방법, 무순위 확장 결과

KT-set 1.0 (Ide Dec-Hi, 무 순위 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.15	0.29	0.24	0.41	0.45	0.43	0.40	0.38	0.36	0.36	0.34
0.1	0.16	0.30	0.24	0.42	0.43	0.41	0.38	0.37	0.35	0.34	0.35
0.2	0.19	0.28	0.23	0.33	0.32	0.34	0.32	0.27	0.26	0.29	0.29
0.3	0.20	0.27	0.21	0.30	0.29	0.32	0.31	0.28	0.26	0.30	0.30
0.4	0.19	0.25	0.21	0.27	0.27	0.30	0.30	0.26	0.25	0.29	0.29
0.5	0.18	0.22	0.20	0.26	0.25	0.30	0.28	0.25	0.24	0.29	0.28
0.6	0.18	0.22	0.20	0.26	0.25	0.27	0.25	0.21	0.21	0.25	0.25
0.7	0.17	0.13	0.16	0.18	0.23	0.19	0.20	0.18	0.18	0.20	0.20
0.8	0.15	0.11	0.13	0.17	0.18	0.17	0.18	0.16	0.16	0.18	0.18
0.9	0.13	0.10	0.11	0.13	0.15	0.14	0.17	0.15	0.15	0.16	0.16
1.0	0.08	0.04	0.07	0.07	0.08	0.08	0.08	0.10	0.10	0.10	0.10
정확율	0.161	0.201 (+24.8%)	0.182 (+13.0%)	0.254 (+57.8%)	0.263 (+63.4%)	0.267 (+65.8%)	0.262 (+62.7%)	0.236 (+46.8%)	0.230 (+42.9%)	0.251 (+55.9%)	0.250 (+55.3%)

KT-set 2.0 (Ide Dec-Hi, 무 순위 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.18	0.35	0.37	0.41	0.45	0.48	0.50	0.47	0.48	0.48	0.48
0.1	0.24	0.27	0.27	0.31	0.34	0.37	0.39	0.38	0.40	0.39	0.39
0.2	0.25	0.18	0.21	0.27	0.30	0.32	0.33	0.34	0.33	0.34	0.34
0.3	0.26	0.16	0.19	0.21	0.25	0.27	0.27	0.27	0.27	0.26	0.26
0.4	0.26	0.13	0.16	0.19	0.21	0.24	0.25	0.25	0.26	0.26	0.26
0.5	0.23	0.12	0.14	0.16	0.16	0.18	0.19	0.20	0.20	0.20	0.19
0.6	0.23	0.11	0.12	0.13	0.15	0.15	0.16	0.17	0.17	0.17	0.17
0.7	0.22	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.13	0.15	0.13
0.8	0.13	0.06	0.07	0.07	0.07	0.08	0.07	0.07	0.08	0.08	0.08
0.9	0.07	0.04	0.05	0.05	0.06	0.06	0.06	0.06	0.07	0.07	0.07
1.0	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04
정확율	0.190	0.140 (-26.3%)	0.154 (-18.9%)	0.176 (-7.4%)	0.196 (+3.2%)	0.208 (+9.5%)	0.219 (+15.3%)	0.218 (+14.7%)	0.222 (+16.8%)	0.222 (+16.8%)	0.222 (+16.8%)

표 5 제안 방법, 무순위 확장 결과

KT-set 1.0 (제안 방법, 무 순위 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.15	0.29	0.24	0.40	0.45	0.39	0.43	0.47	0.47	0.46	0.46
0.1	0.16	0.29	0.26	0.38	0.43	0.37	0.41	0.44	0.45	0.44	0.44
0.2	0.19	0.24	0.27	0.35	0.37	0.33	0.36	0.36	0.37	0.40	0.40
0.3	0.20	0.24	0.24	0.33	0.33	0.33	0.37	0.36	0.37	0.39	0.40
0.4	0.19	0.22	0.23	0.31	0.30	0.29	0.32	0.33	0.33	0.36	0.36
0.5	0.18	0.19	0.20	0.24	0.26	0.27	0.30	0.31	0.32	0.35	0.35
0.6	0.18	0.18	0.20	0.23	0.25	0.24	0.27	0.28	0.29	0.32	0.32
0.7	0.17	0.12	0.16	0.17	0.22	0.19	0.23	0.22	0.23	0.26	0.25
0.8	0.15	0.10	0.14	0.16	0.17	0.16	0.20	0.19	0.19	0.21	0.21
0.9	0.13	0.09	0.11	0.15	0.16	0.14	0.18	0.17	0.17	0.19	0.20
1.0	0.08	0.05	0.07	0.09	0.10	0.09	0.10	0.12	0.12	0.12	0.12
정확율	0.161	0.182 (+13.0%)	0.192 (+19.2%)	0.255 (+58.4%)	0.275 (+70.8%)	0.254 (+57.8%)	0.288 (+78.9%)	0.296 (+83.9%)	0.301 (+87.0%)	0.318 (+97.0%)	0.319 (+98.1%)
KT-set 2.0 (제안 방법, 무 순위 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.18	0.38	0.39	0.40	0.42	0.42	0.43	0.47	0.50	0.51	0.52
0.1	0.24	0.30	0.28	0.32	0.35	0.36	0.39	0.40	0.42	0.43	0.45
0.2	0.25	0.23	0.24	0.28	0.30	0.31	0.33	0.35	0.37	0.38	0.39
0.3	0.26	0.20	0.24	0.27	0.28	0.28	0.29	0.31	0.32	0.33	0.33
0.4	0.26	0.15	0.21	0.23	0.25	0.25	0.25	0.27	0.28	0.28	0.28
0.5	0.23	0.14	0.17	0.18	0.19	0.20	0.21	0.21	0.23	0.24	0.24
0.6	0.23	0.12	0.15	0.17	0.17	0.17	0.18	0.18	0.21	0.20	0.20
0.7	0.22	0.11	0.13	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15
0.8	0.13	0.06	0.08	0.09	0.09	0.08	0.09	0.09	0.10	0.11	0.11
0.9	0.07	0.05	0.06	0.06	0.06	0.07	0.07	0.07	0.08	0.09	0.08
1.0	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04
정확율	0.190	0.160 (-15.8%)	0.180 (-5.3%)	0.196 (+3.2%)	0.207 (+8.9%)	0.211 (+11.1%)	0.220 (+15.8%)	0.230 (+21.1%)	0.245 (+28.9%)	0.251 (+32.1%)	0.254 (+33.7%)

표 6 제안 방법, 관련 정도 기준 확장 결과

KT-set 1.0 (제안 방법, 관련 정도 순위로 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.15	0.37	0.40	0.44	0.42	0.43	0.51	0.50	0.48	0.49	0.46
0.1	0.16	0.35	0.40	0.45	0.43	0.42	0.50	0.50	0.45	0.46	0.44
0.2	0.19	0.36	0.36	0.38	0.37	0.37	0.39	0.40	0.38	0.38	0.40
0.3	0.20	0.32	0.35	0.37	0.35	0.35	0.39	0.39	0.37	0.38	0.40
0.4	0.19	0.30	0.34	0.37	0.34	0.34	0.37	0.37	0.35	0.35	0.36
0.5	0.18	0.30	0.32	0.33	0.33	0.33	0.36	0.36	0.35	0.34	0.35
0.6	0.18	0.29	0.32	0.32	0.32	0.31	0.35	0.35	0.33	0.33	0.32
0.7	0.17	0.24	0.26	0.26	0.27	0.27	0.29	0.29	0.27	0.27	0.25
0.8	0.15	0.21	0.23	0.22	0.23	0.23	0.25	0.24	0.20	0.21	0.21
0.9	0.13	0.20	0.22	0.21	0.22	0.22	0.23	0.23	0.19	0.19	0.20
1.0	0.08	0.12	0.13	0.13	0.14	0.14	0.15	0.15	0.13	0.13	0.12
정확율	0.161	0.278 (+72.7%)	0.304 (+88.8%)	0.316 (+96.3%)	0.310 (+92.5%)	0.311 (+93.2%)	0.345 (+114.3%)	0.344 (+113.7%)	0.319 (+98.1%)	0.322 (+100.0%)	0.319 (+98.1%)
KT-set 2.0 (제안 방법, 관련 정도 순위로 확장)											
	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.18	0.47	0.48	0.51	0.52	0.52	0.51	0.53	0.53	0.54	0.52
0.1	0.24	0.38	0.38	0.40	0.41	0.43	0.41	0.43	0.45	0.46	0.45
0.2	0.25	0.33	0.36	0.35	0.36	0.38	0.36	0.39	0.39	0.40	0.39
0.3	0.26	0.32	0.33	0.32	0.34	0.34	0.32	0.35	0.35	0.35	0.33
0.4	0.26	0.27	0.28	0.26	0.26	0.26	0.26	0.28	0.28	0.29	0.28
0.5	0.23	0.19	0.21	0.22	0.22	0.22	0.22	0.22	0.23	0.23	0.24
0.6	0.23	0.18	0.19	0.19	0.19	0.19	0.20	0.21	0.20	0.21	0.20
0.7	0.22	0.15	0.16	0.17	0.15	0.14	0.15	0.15	0.15	0.15	0.15
0.8	0.13	0.11	0.12	0.11	0.10	0.10	0.11	0.11	0.10	0.11	0.11
0.9	0.07	0.07	0.08	0.08	0.08	0.08	0.09	0.09	0.08	0.09	0.08
1.0	0.03	0.03	0.05	0.05	0.05	0.05	0.05	0.04	0.03	0.04	0.04
정확율	0.190	0.227 (+19.5%)	0.241 (+26.8%)	0.241 (+26.8%)	0.244 (+28.4%)	0.246 (+29.5%)	0.243 (+27.9%)	0.254 (+33.7%)	0.255 (+34.2%)	0.261 (+37.4%)	0.254 (+33.7%)

관련성과 적합 피드백 문서들내에서의 후보 용어 가중치를 결합하여 Dec-Hi 방법보다 더욱 정확한 가중치 부여가 이루어짐으로서 성능의 향상을 이룰 수 있으며, 본 논문에서 제안하는 원 질의어-후보 용어간의 관련 정도 산정은 적합 피드백 문서내에서 용어들의 발생 빈도수(TF)만을 이용하여 간단하게 산정될 수 있음을 결과에서 보여주고 있다. 그러나 표 5에서는 높은 성능 향상과, 확장되는 용어들의 수를 줄일 수 있는 장점은 있으나, Ide Dec-Hi 방법과 비교하여 관련 정도 산정에 추가적인 계산이 필요한 단점이 있다. 그러므로 이러한 단점을 보완하기 위하여 관련 정도를 기준으로 질의를 확장하는 실험을 수행하였다.

표 6에서는 본 논문에서 제안하는 방법중 (식 1)을 사용하여 산정된 원 질의어-후보용어간의 관련 정도를 기준으로 관련 정도가 높은 용어들을 우선적으로 확장한 결과를 보여주고 있다. 표 4와 비교해 보면 Ide Dec-Hi 방법에서는 KT-set 2.0에서 후보 용어들을 30%이상 질의로 확장할 경우에 3.2% 성능 향상을 나타내고 있지만 본 논문에서 제안하는 원 질의어와의 관련 정도를 이용하여 관련 정도가 높은 용어를 우선적으로 확장할 경우에는 상위 10%만 확장을 하였을 때 19.5%의 성능 향상을 보이고 있다. 또한 최고 검색 향상에서도 원 질의어를 이용한 검색 결과와 비교하여 KT-set 1.0에서는 114.3%, KT-set 2.0에서는 37.4%의 향상을 보이고 있으며, Ide Dec-Hi방법과는 KT-set 1.0에서는 29.2%, KT-set 2.0에서는 17.6%의 성능 향상을 보이고 있다. 여기에서 KT-set 1.0과 KT-set 2.0에서의 성능 향상 비율이 큰 차이를 보이고 있는데 이러한 이유는 피드백되는 문서의 양 차이에서 비롯된 것으로, 실험에서 KT-set 1.0의 경우 질의에 대한 전체 적합 문서들 중 27.1%를 피드백한 반면, KT-set 2.0에서는 16.6%를 피드백 하였기 때문이다. 또한 본 논문에서 제안하는 방법은 관련 정도 산정에 추가적인 계산이 필요한 단점이 있는데, 표 6에서처럼 질의어와의 관련 정도를 기준으로 질의를 확장할 경우 전체 후보 용어들중 20%이상 확장할 경우 효율이 현격하게 향상되고 있으므로 확장되는 용어들의 수를 제한하여 추가적인 계산이 필요한 단점을 보완할 수 있다.

5. 결 론

본 논문에서는 원질의어와 질의어로 확장될 수 있는 후보 용어들의 관련 정도를 적합 피드백 문서내에서의 용어 발생 빈도수(TF)를 이용하여 산정하는 방법, 이러한 관련 정도와 적합 피드백 문서내에서의 후보 용어 가중치를 결합하여 용어의 가중치를 산정하는 방법에 대하여 제안하였으며, 다양한 방법을 이용하여 제안하는 방법의 검색 효율을 평가하였다.

표 3과 4에서는 Ide Dec-Hi방법을 이용할 경우 역문서 빈도수(IDF)를 기준으로 역문서 빈도수가 높은 용어들을 우선적으로 확장하는 방법은 검색되는 문서의 수가 적기 때문에 검색 효율을 높일 수 없었으며, 전체 용어들을 무작위로 확장하는 방법이 효과적임을 알수 있었고, 표 5와 표 6에서는 원 질의어와의 관련 정도를 고려하여 가중치를 부여하고 질의를 확장할 경우 높은 성능 향상을 이루어짐을 알수 있었으며, 원 질의어와의 관련 정도를 기준으로 관련 정도가 높은 용어들을 우선적으로 확장하는 방법이 더 우수한 방법임을 알 수 있었다.

결과적으로, 실험을 종합해 보면 표 5와 표 6에서 나타난 바와 같이 질의어로 확장될 수 있는 후보 용어들은 적합 피드백 문서내에서 원 질의어와의 발생 빈도 분포를 이용하여 원 질의어와의 관련 정도를 산정하고, 이러한 관련 정도와 피드백 문서내에서의 후보 용어 가중치를 결합하여 용어의 가중치를 부여할 경우 용어의 가중치를 좀 더 정확하게 산정할 수 있으며, 관련 정도를 기준으로 질의를 확장할 경우에는 원 질의를 이용한 검색과 비교하여 KT-set 1.0에서는 최고 114.3%, KT-set 2.0에서는 37.4% 검색 효율이 향상됨을 알 수 있었으며, 관련 정도를 기준으로 용어들을 확장함으로서 관련 정도 산정에 추가적인 계산이 필요한 단점을 보완할 수 있었다.

본 논문에서 사용된 방법은 기존에 확장된 용어들에 대한 정보를 타 질의어에서 활용할 수 없는 한계를 지니고 있다. 그러므로 기존에 확장된 용어들을 지속적으로 관리하여 타 질의어에서 활용할 수 있는 방법이 연구되어져야 하며, 초기 질의에 대한 검색 결과중 적합 문서에 대한 관련 정도를 반영한 가중치 산정 방법이 연구되어져야 한다. 또한 원 질의어-후보 용어의 관련 정도를 이용하여 시소러스를 자동으로 생성하는 방법이 연구될 필요가 있으며, 현재의 제안 방법을 Web에 확장하는 방법을 연구할 필요가 있다.

참 고 문 헌

[1] Salton. G, "Historical Note: The Pase thirty Years in Information Retrieval," Journal of the American Society for Information Science, Vol.38, No.5, 1987.
[2] Croft.W.B, Cook. R., and Wilder. D, "Providing

Government Information on the Internet: Experiences with THOMAS," In Digital Libraries Conference DL'95, pp.19-24, 1995.
[3] Voorhees.E, "Query expantion using lexical-semantic relations" Proceeding of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp.61-69,1994.
[4] Sparck Jones.K, "Automatic Keyword Classification for Information for retrieval," Butterworth, London. 1971.
[5] Racchio.J.J, "Relevance Feedback in Information Retrieval," Englewood Cliffs, 1971.
[6] Croft. W.B , "Experiments with Representation in a Document Retrieval System," information Technology: Research and Development, 2(1), 1-21, 1983.
[7] Robertson, S.E. and K.Sparck Jones, "Relevance Weighting of Search Terms," Journal of the American Society for Information Science, 27(3), 129-146, 1976.
[8] Harman. D, "Towards Interactive Query Expansion," Paper presented at ACM Conference on Research and Development in Information Retrieval, Grenoble, France, 1988.
[9] Salton. G. and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Science, 41(4), 228-297, 1990.
[10] Jinix Xu and W. Bruce Croft, "Query Expansion Using Local and Global Document Analysis," Proceeding of ACM SIGIR International Conference on Research and Development in Information Retrieval. pp.4-12, 1996.
[11] 김성혁 외 5인, "자동 색인기 성능 실험을 위한 Test Set 개발", 정보관리 학회지 제11권 1호, 1994.
[12] 김재군, 김영환, 김성혁, "한국어 정보 검색 연구를 위한 시험용 데이타 모음(KTSET)", 제 6회 한글 및 한국어정보처리학술대회, 1998

김 주 연
1994년 금오공과대학교 전자계산학과(학사). 1997년 금오공과대학교 전자과(석사). 1998년 ~ 현재 금오공과대학교 전자과 박사과정 재학중. 관심분야는 정보검색, 지능형 에이전트

김 병 만
1987년 서울대학교 컴퓨터공학과(학사). 1989년 한국과학기술원 전산학과(석사). 1992년 한국과학기술원 전산학과(박사). 1992년 ~ 1994년 금오공과대학교 컴퓨터공학부 전임강사. 1994년 ~ 1998년 금오공과대학교 컴퓨터공학부 조교수. 1998년 ~ 1999년 University of California, Irvine, 연구교수. 1998년 ~ 현재 금오공과대학교 컴퓨터공학부 부교수. 관심분야는 인공지능, 정보검색, 소프트웨어 검증 및 테스팅

박 혁 로
1987년 서울대학교 컴퓨터공학과(학사). 1989년 한국과학기술원 전산학과(석사). 1997년 한국과학기술원 전산학과(박사). 1994년 10월 ~ 1999년 2월 연구개발정보센터 선임 연구원. 1999년 3월 ~ 현재 전남대학교 컴퓨터정보학부 조교수. 관심분야는 정보검색, 자연언어처리, 인공지능