용어 발생 유사도와 퍼지 추론을 이용한 질의 용어 확장 및 가중치 재산정

(Query Term Expansion and Reweighting using Term Co-Occurrence Similarity and Fuzzy Inference)

김 주 연 † 김 병 만 ††
(Ju Youn Kim) (Byeong Man Kim)

요 약 본 논문에서는 사용자의 적합 피드백을 기반으로 적합 문서들에서 발생하는 용어들과 초기 질의어간의 발생 빈도 유사도 및 퍼지 추론을 이용하여 용어의 가중치를 산정하는 방법에 대하여 제안한다. 피드백 문서들에서 발생하는 용어들 중에서 불용어를 제외한 모든 용어들을 질의어로 확장될 수 있는 후보 용어들로 선택하고, 발생 빈도 유사성을 이용한 초기 질의어-후보 용어의 관련 정도, 용어의 IDF, DF 정보를 퍼지 추론에 적용하여 후보 용어의 초기 질의어에 대한 최종적인 관련 정도를 산정 하였으며, 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 Dec Hi방법, 용어 분포 유사도를 이용한 방법, 퍼지 추론을 이용한 방법들을 정확률-재현률을 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 by the relevance degree between a candidate term and initial query. The relevance degree is calculated by fuzzy inference with the information such as the term co-occurrence similarity between a candidate term and initial query, the IDF of a candidate term, and document frequency in the feedbacked documents. The terms to be actually expanded are selected by use of the relevance degree and combined with initial query to construct an expanded query. We use KT-set 1.0 and KT-set 2.0 for performance evaluation and compare our method with two methods, one Dec-Hi method and the other Term-Distribution Similarity method, based on recall and precision.

1. 서 론

정보 검색 시스템은 대용량의 데이타로부터 주어진 시간내에 원하는 정보를 발견할 수 있도록 도와주는 시스템이다[1]. 그러나, 이러한 정보 검색 시스템은 질의어로 사용한 용어가 문서에서 발생할 경우에만 검색이 가능하며, 문서를 작성할 때 사용된 용어와 의미가 유사한 동의어를 이용하여 질의어를 한 경우 질의어에 적합한 문서일 경우에도 검색을 할 수 없는 용어 불일치 문제라는 가장 기본적인 문제를 가지고 있다. 간단한 예로, 사용자가 질의어를 기술할 때 문서에서 작가가 어떤 개념을 기술하기 위하여 사용한 용어와 의미가 유사한 용어를 사용하여 질의를 한다고 가정하면, 정보 검색 시스템은 질의어로 주어진 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]. 비

† 학생회원 : 금오공과대학교 컴퓨터공학부
jykim@cespc1.kumoh.ac.kr
†† 종신회원 : 금오공과대학교 컴퓨터공학부 교수
bmkim@cespc1.kumoh.ac.kr
논문접수 : 2000년 2월 29일
심사완료 : 2000년 6월 29일

록 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 질의어에서 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 질의어를 확장하는 방법이 수동으로 구축된 시소러스를 이용하는 방법보다 더욱 효과적이라는 사실이 밝혀졌다[3].

질의어 수정을 위하여 자동으로 구문을 분석하고 확장하는 방법은 1971년 Sparck Jones[4]이 문서들에서 동시에 발생하는 용어들을 분류하여 질의어로 확장하는 방법을 제안한 이후 많이 연구되어져 왔으며, 이러한 방법들 중에서 사용자에 의한 적합 피드백을 기반으로 적합 문서내에서 확장될 용어를 선택하는 방법은 Rocchio가 용어 가중치 재산정과 질의어 확장을 조합한 질의어 수정에 관한 실험 결과를 발표하면서부터이다[5]. 적합 피드백을 기반으로 질의어를 수정하는 방법 현재까지 많이 연구되어 지고 있으며, 1)질의어 가중치 재산정 방법, 2)질의어 확장 방법, 3)질의어 가중치 재산정 및 질의 용어 확장 방법으로 구분 할 수 있다.

질의어 가중치 재산정 및 질의어 확장 방법중에서 사용자에 의해 판단된 초기 질의어에 대한 피드백 문서들을 근거로 피드백 문서내에서 발생하는 용어들을 질의어로 확장될 수 있는 후보 용어들로 선택하고, 피드백 문서들에서 후보 용어와 원 질의어들의 피드백 문서내 발생 빈도수(TF)를 이용하여 원 질의어에 대한 후보 용어의 관련 정도를 산정하는 방법[6]이 제안되었다. 그러나, 기존에 제안된 방법에서는 용어의 발생 빈도 수만을 이용하여 관련 정도를 산정하게 됨으로서 확장 용어로서 중요하지 않은 후보 용어들도 관련 정도가 높게 산정될 수 있고, 이러한 현상은 성능을 개선하는데 한계점으로 작용하고 있다.

본 논문에서는 사용자의 적합 피드백을 이용한 용어 가중치 산정 및 질의어 확장을 위해 기존의 용어 발생 분포 유사도를 이용한 관련정도 산정 방법[6]을 개선한 퍼지 추론[7]을 이용한 후보 용어-원질의어 간의 관련 정도 산정 방법에 대하여 제안한다.

후보 용어-원질의어 간의 정확한 관련 정도를 산정하기 위해 후보 용어의 역문헌 빈도수(IDF), 피드백 문서들 중에서 후보 용어가 발생한 문서 수(DF), 용어의 발생 빈도수(TF)를 이용한 후보 용어-원질의어 간의 관련 정도를 퍼지 추론에 적용하여 관련 정도를 재산정하였으며, 피드백 문서내에서 후보 용어의 가중치와 관련 정도를 결합하여 확장될 후보 용어의 가중치를 산정하였다. 성능 평가에서는 퍼지 추론에 의해 산정된 후보 용어의 관련 정도를 기준으로 관련 정도가 높은 용어들을 기준으로 확장될 용어의 수를 제한하여 성능에 미치는 영향을 실험하였으며, 상대적인 성능 평가를 위하여 적합 피드백에서 많이 사용되는 Ide Dec-Hi방법(BaseLine), 용어 발생 분포 유사도를 이용한 관련정도 산정 방법[6], 본 논문에서 제안하는 퍼지 추론을 이용한 관련 정도 산정 방법등 3가지 방법을 정확률-재현률을 이용하여 성능을 비교 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 적합 피드백에 대한 기존 연구들과 기존의 용어 발생 분포 유사도를 이용한 관련 정도 산정 방법[6]에 대하여 분석하고, 3장에서는 본 논문에서 제안하는 퍼지 추론을 이용한 관련 정도 산정 방법에 대하여 설명한다. 또한, 4장에서는 재현률과 정확률을 사용하여 제안하는 방법과 기존 방법의 성능을 다양한 방법을 이용하여 비교 평가하였다. 마지막으로 5장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구 및 문제점

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

1) 질의어 가중치 재산정 방법: 사용자의 질의와 관련이 있는 적합 문서와 관련이 없는 비적합 문서들에서 발생하는 질의어들의 중요도 분포를 계산하여 질의어의 가중치만을 재산정하고 질의 용어는 확장하지 않는 방법이다[8]. 이러한 방법은 1976년 Robertson과 Sparck Jones에 의해 처음 사용되었으며 적합, 비적합 문서에서의 질의어 분포를 바탕으로 용어의 가중치를 산정하였다[9]. 그러나 이러한 방법은 검색 시스템의 가장 기본적인 문제인 용어 불일치 문제를 해결할 수 없다.

2) 질의어 확장 방법: 수동으로 구성된 시소러스를 이용하거나 자동으로 생성된 용어를 질의어에 포함하여 검색 시 질의어를 확장하는 방법[10]으로서 이러한 연

구들의 대부분은 용어와 용어간의 관련성을 이용하거나 클러스트링 방법을 이용하여 진행되어지고 있다.

3) 질의어 가중치 재산정 및 질의 용어 확장 방법: 질의어 가중치 재산정 방법과 질의어 확장 방법을 결합한 방법으로 적합 피드백 기반의 검색에서 현재 가장 많이 연구되고 있는 분야이며, 실험에서도 3가지 방법중 평균 정확률 면에서 가장 우수한 성능을 나타냈다[11].

질의어 가중치 재산정 및 질의어 확장 방법들중에서 Salton과 Buckley는 6개의 실험 문서들을 통해서 Ide Dec-Hi, Ide Regular, Rocchio방법들을 실험하였다[11]. 이와 같은 세 가지 방법의 기본적인 연산 절차는 문서 벡터와 원래의 질의어 벡터를 병합하는 것이다. 이것은 적합 문서들에 해당 질의어의 발생으로부터 가중치를 부가하고, 비적합 문서에 대하여 가중치를 줄여줌으로서 질의어에 자동적으로 가중치가 다시 부여되도록 한다. 질의어는 원래의 질의어에 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 여섯개의 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 정확률이 증가하였다.

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. 그러므로, 이러한 Dec-Hi방법의 문제점을 개선하기 위하여 사용자의 연관 피드백 문서들에서 후보 용어들과 원 질의어와의 발생 분포 유사도를 이용하여 후보 용어-원질의어 간의 관련 정도를 산정하고, 가중치를 재산정하는 방법이 제안되었다[6].

제안된 방법[6]에서는 사용자의 연관 피드백 문서들에서 발생하는 모든 용어들을 질의로 확장될 수 있는 후보 용어들로 선택하고, 문서들에서의 용어 발생 유사도를 이용하여 후보 용어-원질의어와의 관련정도를 산정 하였으며, 산정된 관련 정도와 피드백 문서들에서의 가중치를 결합하여 후보 용어의 가중치를 재산정하였다. 이 방법에서는 관련 정도를 기준으로 질의를 확장할 경우 Dec-Hi 방법과 비교하여 KT-set 1.0에서는 29.2%, KT-set 2.0에서는 17.6%의 성능 향상을 보였으며, 최고 성능에 도달하기 위해 확장되는 용어의 수에서도 현격한 개선을 보였다.

그러나, 제안된 방법에서는 용어의 발생 빈도 수만을 이용하여 관련 정도를 산정하게 됨으로서 확장 용어로서 중요하지 않은 후보 용어들도 관련 정도가 높게 산정될 수 있고, 이러한 현상은 성능을 개선하는데 한계점으로 작용하고 있다. 그러므로, 후보 용어-원질의어 간의 관련정도를 산정할 때 용어의 발생 빈도수(TF)뿐만 아니라 용어의 역문헌 빈도수(IDF), 피드백 문서내에서의 용어 발생 문서 수(DF)등을 고려하여 관련정도를 산정할 필요가 있다.

퍼지 추론[7]은 퍼지 제어기에 도입된 의사 결정 방법으로써, 인간의 사고나 자연어의 특성과 많은 유사성을 가지고 있는 퍼지 논리에 의해 작성된 언어적 형식의 제어규칙을 이용하여 퍼지 제어 입력을 구해주는 기능을 수행하며, 전문가 시스템이나 기타 인공지능 분야에서 도입되는 추론기능보다 단순하다. 퍼지 규칙은 조건부와 출력부로 구성되어 있으며, 이러한 퍼지 규칙들은 합성연산(Composition operation)법칙을 도입한 여러 가지 추론 방법들을 통하여 추론에 이용되고, 퍼지 추론결과는 비퍼지화 작업을 통하여 명확한 비퍼지 값으로 변환한다. 본 논문에서는 여러 가지 추론 방법들 중에서 Mamdani의 min연산 방법[7]을 사용하였다.

본 논문에서는 기존의 용어 발생 분포 유사도를 이용한 관련정도 산정 방법을 개선하기 위하여 용어 발생 분포 유사도, 용어의 역문헌 빈도수(IDF), 피드백 문서내에서의 용어 발생 문서 수(DF)등의 정보들을 퍼지 추론에 이용하여 후보 용어-원질의어 간의 관련정도를 산정하였다. 이때, 본 논문의 퍼지 추론에 사용되는 각 요소들은 퍼지 이론을 적용하지 않고도 다양한 방법들에 이용될 수 있으나, 퍼지 이론을 이용할 경우 각 요소 값들을 인간의 직관적인 사고(예를 들면, "IDF가 낮

고, DF가 높은 용어일 경우 일반적인 용어일 가능성이 높다")에 반영하여 해석하기 쉽고, 인간의 직관적인 사고를 퍼지 규칙으로 작성하기 쉽다. 따라서 본 논문에서는 각 요소 값들과 인간의 직관적인 사고를 퍼지 이론을 이용하여 표현하고, 퍼지 추론을 이용하여 관련정도를 산정하였다. 또한, 본 논문에서는 Ide Dec-Hi 방법을 참조하여 초기 질의어와의 관련 정도를 고려하여 확장 용어의 가중치를 산정하였고, 관련 정도를 기준으로 확장 용어의 수를 제한하여 실험하였다. 그리고, 본 논문에서 제안하는 방법은 관련 정도를 산정할 때 용어 발생 분포 유사도를 이용하고, 확장될 용어의 가중치를 산정할 때 Ide Dec-Hi방법과 유사한 방법을 이용하여 가중치를 산정함으로 제안하는 방법의 성능을 상대 평가하기 위하여 Ide Dec-Hi방법, 용어 발생 분포 유사도를 이용한 방법을 함께 실험하였다.

3. 용어 발생 유사도 및 퍼지 추론을 이용한 질의 용어 가중치 재산정

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 후보 용어 선택 방법과 기존의 후보 용어-원질의어 간의 관련정도 산정 방법에 대하여 기술한다. 또한 산정된 관련정도를 본 논문에서 제안한 퍼지 추론에 적용하는 방법에 대하여 기술하고, 퍼지 추론 결과를 이용하여 가중치를 산정하는 방법에 대하여 기술한다.

3.1 후보 용어 집합을 생성

초기 질의어를 이용한 검색 문서들중에서 사용자에 의해 피드백된 문서들에서 발생하는 용어들중 불용어를 제외한 모든 용어들을 확장될 수 있는 후보 용어들로 생성한다. 후보 용어의 선택은 저장 공간과 실행 속도를 고려하여 용어의 수를 제한할 수 있으나, 본 논문에서는 재현률을 고려하여 적합 문서들에서 발생하는 모든 용어들은 후보 용어들로 선택하였다. 또한, 본 논문의 실험에서는 원 질의어를 이용한 검색 결과에서 상위 10위내에 검색된 문서들중 질의에 대한 적합 문서들을 피드백된 문서로 가정하여 실험하였다.

3.2 후보 용어-원질의 간의 관련정도 산정

후보 용어-원질의어 간의 관련정도는 적합 피드백 문서들에서 후보 용어와 원질의어들 간의 발생 빈도수를 이용하여 산정된다. 각 적합 피드백 문서들에서 후보 용어와 원질의어로 사용된 각 용어들과의 발생 빈도수 차를 산정하고, 이것을 전체 합산하여 관련정도를 산정하며, 후보 용어가 발생하지 않은 피드백 문서에서는 원질의어와의 관련 정도를 0으로 산정된다. (식 1)에서는 후보 용어-원질의어 간의 관련정도 산정에 사용된 식을 보여준다.

$$S_{ik}(Q, t_i) = 1 - \log_{10}(\sqrt{\sum_{j=1}^{m}(|qf_{jk} - tf_{ik}|)}) \quad \text{if } tfd_{ik} \neq 0 \qquad (1)$$

$$S_{ik}(Q, t_i) = 0 \qquad \text{if } tfd_{ik} \equiv 0$$

$S_{ik}(Q, t_i)$: 피드백 문서 k에서 후보 용어 t_i와 원 질의어들 간의 관련 정도

qf_{jk}: 피드백 문서 k에서 원질의어 j의 빈도수

tf_{ik}: 피드백 문서 k에서 후보 용어 i의 빈도수

m: 원질의어의 수

이와 같은 유사도 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.

3.3 퍼지 추론을 이용한 관련정도 산정

3.2절에서는 후보 용어-원질의어 간의 관련 정도를 피드백 문헌내에서의 용어 발생 빈도수만을 이용하여 산정하였다. 그러나 이러한 방법은 질의로서 중요하지 않은 용어들도 관련 정도가 높게 산정될 가능성이 있음으로 더욱 정확한 관련정도 산정을 위하여 3.2절에서 산정된 관련정도(S)와 후보 용어의 역문헌 빈도수(IDF), 피드백 문서중에서 후보 용어가 발생한 문서 수(DF)를 퍼지 추론에 적용하여 각 피드백 문서에서 후보 용어-원질의어 간의 관련정도를 재산정하였다. 이때, 후보 용어-원질의어 간의 관련정도를 퍼지 추론을 이용하여 산정한 이유는 퍼지 이론을 이용할 경우 각 요소 값들을 인간의 직관적인 사고에 반영하기 위하여 쉽게 해석할 수 있고, 인간의 직관적인 사고를 퍼지 규칙으로 간단하게 작성할 수 있기 때문이며, 이러한 퍼지 이론에서 제어 값의 추론은 퍼지 추론을 이용할 수 있기 때문이다. 예를 들면, "IDF가 낮고, DF가 높은 용어일 경우 일반적인 용어일 가능성이 높다" 라는 전제는 가장 보편적으로 사용되는 전제이며, 이러한 전제에서 얼마의 값을 "낮다" 혹은 "높다" 로 해석해야 할지를 기존의 정량적인 방법으로는 해석하기가 어렵지만, 퍼지 이론을 이용할 경우 불확실한 현상을 기술할 수 있음으로 쉽게 해석할 수 있다.

1) 퍼지 입출력 변수

그림 1에서는 본 논문에서 사용한 퍼지 입출력 변수

들을 나타내고 있다. 그림 1의 (a)에서 입력 변수 S는 3.2절에서 산정한 후보 용어-원질의어 간의 관련정도를 사용하고 있으며, 소속 함수는 4개를 사용하였다. 이때 Z 소속 함수를 사용한 이유는 피드백 문서에서 후보 용어가 발생하지 않을 경우 원질의어에 대한 후보 용어의 관련정도는 0으로 산정되기 때문이며, 각 소속함수들의 범위는 직관적인 값으로 설정하였다.

(b)에서는 후보 용어의 역문헌 빈도수(IDF)를 정규화한 FI, 피드백 문서중에서 후보 용어가 발생한 문서 수(DF)를 정규화한 FD 입력 변수의 소속 함수를 나타내고 있으며, 이러한 정보들은 보다 정확한 관련정도 산정을 위하여 부가적으로 사용된 정보들이다. FI는 전체 문서들에서 후보 용어의 중요도를 평가하기 위하여 사용하였으며 0.0 - 1.0의 값으로 정규화하기 위하여 (식 2)를 사용하였다. 또한 FD는 사용자의 피드백 문서들에서 후보 용어의 중요도를 평가하며, FI와 결합하여 후보 용어가 질의로서 중요하지 않은 일반 용어일 가능성을 평가하게 된다. (식 3)에서는 DF를 0.0 - 1.0으로 정규화하기 위하여 사용된 식을 나타내고 있다.

$$FI_i = \frac{IDF_i}{MIDF} \tag{2}$$

FI_i : 후보 용어 i의 퍼지 역문헌 소속함수 값

IDF_i : 후보 용어 i의 역문헌 빈도수

$MIDF$: 후보 용어들중의 최대 역문헌 빈도수

$$FD_i = \frac{FCDF_i}{TFDF} \tag{3}$$

FD_i: 후보 용어 i 의 퍼지 피드백 소속함수 값

$FCDF_i$: 피드백 문서들중 후보 용어 i가 발생한 문서 수

$TFDF$: 피드백 문서 수

(c)에서는 출력 변수 R 의 소속 함수들을 나타내고 있으며, 6개의 소속 함수들로 구성하고 소속 함수 구간의 차를 0.3으로 설정하였다. 이와 같이 출력 변수(R)의 소속 함수 구간 차를 0.3으로 설정하고, 소속 함수 값들을 정규화하지 않은 이유는 퍼지 추론에 의해 X와 XX 함수로 추론될 경우 Ide Dec-Hi방법에 의한 가중치 산정보다 부가적인 가중치를 부여하기 위함이다.

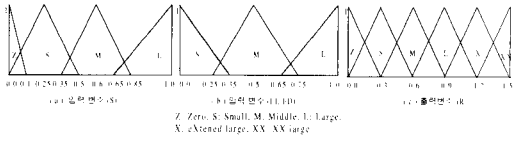

그림 1 퍼지 입출력 변수의 소속함수

2) 추론 규칙

그림 2에서는 본 논문에서 사용한 퍼지 추론 규칙들을 나타내고 있으며, 3.2절에서 산정된 후보 용어-원질의어 간의 관련 정도(S)를 우선적으로 고려하고, FI와 FD는 보조로 사용하였다. 추론 규칙은 36개의 규칙으로 구성되어 있으며, Z 소속함수 결과를 갖는 규칙이 13개, S는 4개, M은 5개, L은 6개, X와 XX는 각각 4개의 추론 규칙이 있다. 이들 규칙중 3.2절에서 산정된 후보 용어-원질의어 간의 관련정도(S)가 0.0일 경우에는 FI와 FD에 상관없이 출력이 Z 소속 함수를 가지도록 하였으며, 관련정도 S와 FI, FD로서 출력 변수의 소속함수를 결정하기 어려울 경우 출력 변수의 소속 함수는 관련정도 S의 소속함수를 가지도록 추론 규칙을 작성하였다. 규칙에서는 FI가 낮고(S) FD가 높을 경우(L) 이러한 용어들은 질의어로서 중요하지 않은 일반적인 용어일 가능성이 높으므로 관련정도에 따라 최소 Z에서 최대 M의 소속함수를 가지도록 하였으며, FI가 높고(L) FD가 높을 경우(L) 이러한 용어는 질의로서 매우 중요한 용어이므로 최소 X, 최대 XX의 소속함수를 가지도록 하였다. 또한, FI가 M이거나 혹은 L이면서 FD가 낮을 경우(S) 이러한 용어들은 특정 문서에만 발생하는 극히 제한된 용어이므로 질의로서의 중요성을 결정하기 어렵기 때문에 관련 정도의 소속 함수를 출력 함수로 가지도록 하였다.

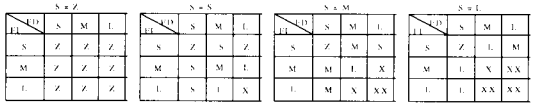

그림 2 퍼지 추론 규칙

3) 비퍼지화

퍼지 추론 규칙에 의하여 생성된 출력 변수(R)의 소속 함수 값들을 단일한 값으로 비퍼지화 하기 위하여 본 논문에서는 (식 4)와 같이 무게중심(center of gravity) 방법을 사용하였다. 무게중심법은 합성된 출력 퍼지 변수들의 무게중심을 구하여, 해당하는 제어 값을 제어입력으로 사용하는 방법이다.

$$R_{ik} = \frac{\sum_{j=1}^{n} \mu(\mu_{ij}) \times \mu_j}{\sum_{j=1}^{n} \mu(\mu_{ij})} \tag{4}$$

R_{ik}: 후보 용어 i가 피드백 문서 k에서의 관련 정도

$\mu(\mu_{ij})$: 후보 용어 i가 소속 함수 j에 소속된 정도

μ_j: 소속 함수 k의 구간 값

n: 출력 변수(R)의 소속 함수 수

3.4 후보 용어의 가중치 산정 및 질의어 확장

퍼지 추론을 이용하여 산정된 후보 용어-원질의어 간의 관련정도를 이용하여 후보 용어의 가중치를 산정하고 질의를 확장하는 방법은 Ide Dec-Hi방법을 변형한 (식 5)를 사용하였다.

$$wt_i = \sum_{k=1}^{n} (wt_{ik} * R_{ik}) \quad (5)$$

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

wt_i :후보 용어 i의 피드백된 전체 문서들에서의 가중치

wt_{ik} : 피드백 문서 k에서 후보 용어 t_i의 가중치

R_{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

k : 피드백된 문서의 수

(식 5)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. 이것은 직관적이고 간단한 방법으로 부가적인 연산이 필요 없이 전체 피드백 문서에서의 후보 용어 가중치를 산정할 수 있다.

질의어 확장은 피드백된 문서들에서 발생하는 모든 후보 용어들을 질의어로 확장한다. 후보 용어가 원 질의어에 포함되어 있을 경우 원 질의어 가중치와 (식 5)에 의해 산정된 후보 용어의 가중치를 합산하고, 만약, 원 질의어에 포함되지 않을 경우에는 후보 용어의 가중치만을 가지고 확장된다. 또한 원 질의어가 후보 용어가 아닐 경우에는 초기 가중치 값을 가진다.

3.5 퍼지 추론을 이용한 관련정도 산정 예

1) 기본 가정

사용자는 3개의 문서를 적합 피드백 하고(d1, d2, d3). 피드백 문서에서 후보 용어 t1이 1.5의 IDF 값을 가지고 8, 3, 11회 발생하였으며, 후보 용어들 중에서 최대 IDF값은 4.0으로 가정한다. 또한 원질의어 q1이 각 피드백 문서에서 3, 2, 5회 발생하고, q2 가 각 피드백 문서에서 1, 0, 2회 발생하였다고 가정한다.

2) 후보 용어-원질의어 간의 관련정도 산정

(식 1)을 이용하여 아래와 같이 각 피드백 문서내에서의 후보 용어-원질의어간의 관련 정도를 산정한다. 관련정도 산정은 기존에 제안된 용어 분포 유사도를 이용하여 산정되며, 후보 용어 t1의 관련 정도를 벡터로 표시하면 다음과 같다.

t1 = { 0.46, 0.70, 0.41 }

$S_{11} = 1 - \log 10(\sqrt{5+7}) = 0.46,$

$S_{12} = 1 - \log 10(\sqrt{1+3}) = 0.70,$

$S_{13} = 1 - \log 10(\sqrt{6+9}) = 0.41$

3) 퍼지 추론을 이용한 관련정도 산정

후보 용어의 역문헌 빈도수(IDF), 피드백 문서들에서 후보 용어가 발생한 문서 수(DF)를 퍼지 추론에 사용하기 위하여 (식 2) (식 3)를 이용하여 0.0 - 1.0의 값으로 정규화한 FI, FD를 아래와 같이 산정 한다.

$$FI_1 = \frac{1.5}{4.0} = 0.38, \qquad FD_1 = \frac{2}{3} = 0.67$$

또한, 그림 1과 (식 6)을 이용하여 관련정도(S), FI, FD의 함수에 대한 소속정도를 산정하여 아래와 같이 퍼지 집합으로 나타낸다.

S11= {0.0/Z, 0.16/S, 0.44/M, 0,0/L}, S12 = {0.0/Z, 0.0/S, 0.6/M, 0.14/L},

S13 = {0.0/Z, 0.36/S, 0.24/M, 0.0/L }

$$\mu_{A(x)} = \begin{cases} 0, & x < a1 \\ (x - a1) / (a2 - a1), & a1 \le x \le a2 \\ (a3 - x) / (a3 - a2), & a2 \le x \le a3 \\ 0, & x > a3 \end{cases}$$

FI1 = {0.0/S, 0.5/M, 0.0/L }, FD1 = {0.0/S, 0.32/M, 0.06/L } (6)

그리고, 그림 2에서의 추론 규칙을 이용하여 아래 그림 3과 같이 출력변수(R)의 소속함수 추론 규칙을 이용하여 소속값을 산정하고, (식 4)를 이용하여 아래와 같이 단일한 값으로 비퍼지화 하여 최종적으로 후보 용어-원질의어 간의 관련정도를 산정한다. 그림 3에서는 출력 변수(R)의 소속 함수 Z, S, M는 추론 결과값이 0.0이므로 그림에서 생략하였다.

$$R_{11} = \frac{(0.0*0.0) + (0.0*0.3) + (0.16*0.6) + (0.32*0.9) + (0.06*1.2) + (0.0*1.5)}{0.0 + 0.0 + 0.16 + 0.32 + 0.06 + 0.0} = 0.844$$

$$R_{13} = \frac{(0.0*0.0) + (0.0*0.3) + (0.36*0.6) + (0.24*0.9) + (0.06*1.2) + (0.0*1.5)}{0.0 + 0.0 + 0.36 + 0.24 + 0.06 + 0.0} = 0.76$$

$$R_{12} = \frac{(0.0*0.0) + (0.0*0.3) + (0.0*0.6) + (0.32*0.9) + (0.14*1.2) + (0.06*1.5)}{0.0 + 0.0 + 0.0 + 0.32 + 0.14 + 0.06} = 1.05$$

4. 성능 평가

4.1 성능 평가 환경 및 자료

본 논문의 성능 평가에 사용된 실험 자료는 한국어 테스트콜렉션인 KT-set 1.0[12]과 KT-set 2.0[13]을 사용하였다. KT-Set 1.0에는 정보과학회논문지, 1993 한국정보과학회 학술발표대회논문집, 정보관리학회지에 수록된 논문들로 구성된 1,053개의 문서들과 30개의 질

출력	퍼지 규칙			D1			D2			D3		
	S	FI	FD	함수값	Min	Max	함수값	Min	Max	함수값	Min	Max
M	S	M	M	0.16 0.5 0.32	0.16	0.16	0.0 0.5 0.32	0.0	0.0	0.36 0.5 0.32	0.36	0.36
	M	S	M	0.44 0.0 0.32	0.0		0.6 0.0 0.32	0.0		0.24 0.0 0.32	0.0	
	M	M	S	0.44 0.5 0.0	0.0		0.6 0.5 0.0	0.0		0.24 0.5 0.0	0.0	
	M	L	S	0.44 0.0 0.0	0.0		0.6 0.0 0.0	0.0		0.24 0.0 0.0	0.0	
	L	S	L	0.0 0.0 0.06	0.0		0.14 0.0 0.06	0.0		0.0 0.0 0.06	0.0	
L	S	M	L	0.16 0.5 0.06	0.06	0.32	0.0 0.5 0.06	0.0	0.32	0.36 0.5 0.06	0.06	0.24
	S	L	M	0.16 0.0 0.32	0.0		0.0 0.0 0.32	0.0		0.36 0.0 0.32	0.0	
	M	M	M	0.44 0.5 0.32	0.32		0.6 0.5 0.32	0.32		0.24 0.5 0.32	0.24	
	L	S	M	0.0 0.0 0.32	0.0		0.14 0.0 0.32	0.0		0.0 0.0 0.32	0.0	
	L	M	S	0.0 0.5 0.0	0.0		0.14 0.5 0.0	0.0		0.0 0.5 0.0	0.0	
	L	L	S	0.0 0.0 0.0	0.0		0.14 0.0 0.0	0.0		0.0 0.0 0.0	0.0	
X	S	L	L	0.16 0.0 0.06	0.0	0.06	0.0 0.0 0.06	0.0	0.14	0.36 0.0 0.06	0.0	0.06
	M	M	L	0.44 0.5 0.06	0.06		0.6 0.5 0.06	0.06		0.24 0.5 0.06	0.06	
	M	L	M	0.44 0.0 0.32	0.0		0.6 0.0 0.32	0.0		0.24 0.0 0.32	0.0	
	L	M	M	0.0 0.5 0.32	0.0		0.14 0.5 0.32	0.14		0.0 0.5 0.32	0.0	
XX	M	L	L	0.44 0.0 0.06	0.0	0.0	0.6 0.0 0.06	0.0	0.06	0.24 0.0 0.06	0.0	0.0
	L	M	L	0.0 0.5 0.06	0.0		0.14 0.5 0.06	0.06		0.0 0.5 0.06	0.0	
	L	L	M	0.0 0.0 0.32	0.0		0.14 0.0 0.32	0.0		0.0 0.0 0.32	0.0	
	L	L	L	0.0 0.0 0.06	0.0		0.14 0.0 0.06	0.0		0.0 0.0 0.06	0.0	

그림 3 퍼지 추론 규칙 적용 예

의가 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 질의어로 구성되어 있고, 질의어 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 문서들 중에서 전체 문서를 검색하지 않고 상위의 몇 개의 문서만을 검색하여 적합 피드백하게 되는 경우가 대부분이므로 상위 10위 내의 문서만을 대상으로 하였다. 또한 확장되는 후보 용어의 수가 성능에 미치는 영향을 평가하기 위하여 Dec-Hi 방법에서는 후보 용어의 가중치를 기준으로 후보 용어의 수를 백분율로 나누어 10-100%까지 변화하여 확장하는 방법(이하 비율확장)을 사용하였으며, 용어 분포 유사도를 이용한 방법은 후보 용어-원질의어 간의 관련 정도를 기준으로 후보 용어들을 비율확장 하였다. 본 논문에서 제안하는 방법에서는 저장 공간과 수행 속도를 고려하여 3.3절의 퍼지 추론에 의해 산정된 후보 용어의 관련정도를 기준으로 최대 100개까지 변화하여 확장하는 방법(이하 제한확장)을 사용하였다. 그리고 검색 효율을 평가하기 위하여 보간 기법을 이용한 고정된 재현률에 대한 정확률을 계산하고, 재현률 0.0 - 1.0까지 11개의 재현률에서의 평균 정확률을 사용하였다.

재현률(Recall)은 전체 적합 문서들중 검색된 적합 문서의 비율을 의미하고, 정확률(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 재현률에서 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 피드

표 1 실험 환경

	KTSET 1.0	KTSET 2.0
질의어 수	12	25
피드백 문서 수	47	128
전체 적합 문서 수	273	1123
평균 후보 용어 수	124	397

백한 문서는 전체 적합 문서 집단에서 제외하고 재 검색을 수행하는 방법으로 추가로 검색된 문서들만을 이용하여 검색 효율을 평가한다[10]. 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 정확률-재현률이 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 있고, 피드백된 문서를 제외한 적합 문서의 수가 10개 이상인 질의를 대상으로 실험을 하였다. 이러한 실험 방법을 선택한 이유는 적합 문서의 수가 10개 미만일 경우 11개의 재현률에서 정확률을 정확하게 산정할 수 없기 때문이다. 그러므로 본 논문에서는 KT-set 1.0에서 12개, KT-set 2.0에서는 25개의 질의를 이용하였으며, 피드백되는 문서의 수는 질의어 하나에 KT-set 1.0에서는 평균 3.92개, KT-set 2.0에서는 5.12개를 사용하였고, 전체 적합 문서 비율로 KT-set 1.0에서는 17.2%, KT set 2.0 에서는 11.4%를 피드백

표 2 Ide Dec-Hi 방법

KT-SET 1.0 (가중치 기준, 비율확장)										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.28	0.41	0.46	0.35	0.40	0.39	0.35	0.37	0.39	0.38
0.1	0.19	0.33	0.42	0.35	0.41	0.39	0.37	0.39	0.40	0.40
0.2	0.25	0.26	0.25	0.24	0.25	0.27	0.29	0.31	0.31	0.32
0.3	0.23	0.25	0.25	0.24	0.24	0.26	0.28	0.27	0.29	0.29
0.4	0.22	0.24	0.24	0.24	0.24	0.26	0.25	0.26	0.27	0.28
0.5	0.24	0.25	0.26	0.26	0.26	0.26	0.23	0.24	0.25	0.26
0.6	0.25	0.26	0.25	0.26	0.25	0.25	0.23	0.24	0.24	0.24
0.7	0.24	0.27	0.26	0.26	0.22	0.22	0.22	0.22	0.21	0.22
0.8	0.19	0.23	0.21	0.21	0.20	0.21	0.18	0.19	0.20	0.20
0.9	0.15	0.14	0.13	0.12	0.12	0.14	0.14	0.15	0.15	0.15
1.0	0.09	0.06	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.08
평균	0.211	0.245	0.255	0.238	0.244	0.247	0.236	0.245	0.253	0.256
재현률	0.903	0.910	0.919	0.924	0.929	0.935	0.940	0.940	0.940	0.949
KT-SET 2.0 (가중치 기준, 비율확장)										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.43	0.46	0.48	0.51	0.54	0.54	0.55	0.56	0.58	0.58
0.1	0.33	0.37	0.37	0.40	0.41	0.40	0.45	0.45	0.45	0.45
0.2	0.24	0.29	0.29	0.30	0.33	0.33	0.36	0.37	0.38	0.38
0.3	0.21	0.24	0.23	0.24	0.22	0.23	0.24	0.26	0.26	0.28
0.4	0.20	0.20	0.19	0.19	0.19	0.20	0.22	0.22	0.24	0.25
0.5	0.14	0.17	0.15	0.03	0.14	0.14	0.16	0.17	0.18	0.18
0.6	0.13	0.15	0.12	0.11	0.12	0.13	0.13	0.15	0.16	0.16
0.7	0.11	0.10	0.09	0.10	0.09	0.11	0.11	0.10	0.10	0.10
0.8	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.07
0.9	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05
1.0	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02
평균	0.174	0.191	0.185	0.190	0.197	0.201	0.213	0.218	0.225	0.230
재현률	0.924	0.963	0.967	0.973	0.984	0.988	0.992	0.994	0.996	0.996

문서로 사용하였다. 또한 비율확장 방법을 이용하여 100% 확장할 경우 질의어로 확장될 수 있는 후보 용어의 수는 각 질의어당 KTSET 1.0에서는 평균 124개, KTSET 2.0에서는 평균 397개의 용어가 확장될 수 있다.

4.3 실험 결과

표 2에서는 본 논문에서 상대 평가 기준으로 설정한 Dec-Hi방법을 후보 용어의 가중치를 기준으로 비율 확장한 실험 결과를 보여주고 있다. 표에서와 같이 본 논문에서 가중치를 기준으로 용어들을 확장한 이유는 무작위 순으로 확장한 경우, IDF를 기준으로 확장한 경우, 피드백 문서내에서의 DF를 기준으로 확장한 경우, 가중치를 기준으로 확장한 경우를 각각 실험하여 4가지 방법중 가중치를 이용한 방법이 가중 우수한 성능을 나타냈기 때문이다. 표 2에와 같이 Dec-Hi방법에서 후보 용어들을 비율 확장할 경우에 90%이상 확장할 경우가 재현률과 정확률에서 가장 효과적임을 보여주고 있으며, 90%로 확장할 경우 확장되는 용어는 KTSET 1.0의 경우 질의어당 평균 111.6개, KTSET 2.0에서는 357.3개가 확장된다.

표 3에서는 기존에 제안된 용어 분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정[11] 방법의 실험 결과를 보여주고 있으며, 표에서의 평균 정확률 향상 정도는 Dec-Hi방법에서 최고 정확률에 대한 향상 정도를 나타내고 있다. 표 3에서는 후보 용어 분포 유사도를 이용하여 질의를 확장할 경우 KTSET 1.0에서는 40%만 확장할 경우 Dec-Hi방법의 최고 정확률과 비교하여 5.08%의 향상을 보이고 있으며, 100%로 확장할 경우

표 3 용어 분포 유사도를 이용한 방법

KT-SET 1.0 (관련정도 기준, 비율확장)										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.40	0.35	0.50	0.49	0.56	0.56	0.56	0.56	0.55	0.49
0.1	0.37	0.35	0.51	0.50	0.50	0.50	0.50	0.47	0.48	0.45
0.2	0.32	0.28	0.30	0.30	0.31	0.29	0.30	0.31	0.33	0.36
0.3	0.30	0.29	0.29	0.31	0.32	0.31	0.33	0.34	0.33	0.37
0.4	0.24	0.28	0.29	0.29	0.28	0.28	0.28	0.29	0.28	0.32
0.5	0.21	0.25	0.26	0.25	0.28	0.28	0.27	0.28	0.27	0.31
0.6	0.21	0.25	0.24	0.24	0.24	0.25	0.25	0.25	0.24	0.29
0.7	0.21	0.22	0.21	0.22	0.21	0.22	0.22	0.21	0.22	0.25
0.8	0.18	0.17	0.16	0.17	0.17	0.17	0.18	0.18	0.18	0.20
0.9	0.17	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.17
1.0	0.10	0.08	0.07	0.07	0.08	0.08	0.09	0.09	0.08	0.08
평균	0.247 (-3.52)	0.241 (-5.86)	0.269 (-1.56)	0.274 (+5.08)	0.283 (+10.5)	0.281 (+9.77)	0.283 (+10.5)	0.284 (+10.9)	0.284 (+10.9)	0.298 (+16.4)
재현률	0.887	0.893	0.914	0.935	0.949	0.949	0.949	0.949	0.949	0.949
KT-SET 2.0 (관련정도 기준, 비율확장)										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	0.53	0.52	0.58	0.64	0.64	0.63	0.63	0.63	0.64	0.63
0.1	0.41	0.42	0.47	0.48	0.49	0.48	0.48	0.50	0.51	0.52
0.2	0.35	0.36	0.36	0.37	0.38	0.41	0.41	0.42	0.42	0.43
0.3	0.34	0.36	0.36	0.36	0.36	0.38	0.37	0.37	0.37	0.34
0.4	0.31	0.32	0.31	0.29	0.28	0.28	0.27	0.27	0.28	0.28
0.5	0.23	0.25	0.25	0.24	0.23	0.23	0.23	0.24	0.24	0.24
0.6	0.21	0.21	0.19	0.19	0.19	0.20	0.21	0.21	0.21	0.20
0.7	0.18	0.17	0.14	0.13	0.13	0.12	0.13	0.13	0.12	0.13
0.8	0.12	0.12	0.09	0.09	0.09	0.08	0.09	0.09	0.09	0.10
0.9	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.06
1.0	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
평균	0.251 (+9.13)	0.257 (+11.7)	0.257 (+11.7)	0.261 (+13.5)	0.260 (+13.0)	0.263 (+14.3)	0.265 (+15.2)	0.267 (+16.1)	0.270 (+17.4)	0.267 (+16.1)
재현률	0.971	0.972	0.975	0.988	0.991	0.992	0.996	0.996	0.996	0.996

표 4 퍼지 추론을 이용한 방법

KT-SET 1.0(관련정도 기준, 제한확장)										
	10개	20개	30개	40개	50개	60개	70개	80개	90개	100개
0.0	0.48	0.50	0.51	0.55	0.63	0.67	0.73	0.73	0.73	0.69
0.1	0.52	0.49	0.52	0.50	0.58	0.60	0.68	0.66	0.62	0.62
0.2	0.34	0.34	0.33	0.33	0.38	0.37	0.41	0.41	0.42	0.40
0.3	0.32	0.31	0.33	0.34	0.44	0.39	0.41	0.40	0.40	0.40
0.4	0.29	0.28	0.30	0.33	0.37	0.39	0.37	0.37	0.37	0.36
0.5	0.31	0.28	0.30	0.30	0.36	0.35	0.34	0.33	0.33	0.33
0.6	0.28	0.25	0.26	0.26	0.31	0.30	0.30	0.31	0.31	0.31
0.7	0.26	0.23	0.24	0.23	0.24	0.26	0.25	0.25	0.26	0.26
0.8	0.17	0.15	0.15	0.17	0.18	0.19	0.19	0.19	0.20	0.20
0.9	0.16	0.14	0.13	0.16	0.15	0.15	0.16	0.15	0.16	0.16
1.0	0.11	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09
평균	0.293 (+14.5) (+1.71)	0.279 (+8.98) (-6.38)	0.286 (+11.7) (-4.03)	0.296 (+15.6) (-0.76)	0.334 (+30.5) (+12.1)	0.342 (+33.6) (+14.8)	0.357 (+39.5) (+19.8)	0.353 (+37.9) (+18.5)	0.354 (+38.3) (+18.8)	0.347 (+35.5) (+16.4)
재현률	0.887	0.918	0.944	0.949	0.949	0.949	0.949	0.949	0.949	0.949
KT-SET 2.0(관련정도 기준, 제한확장)										
	10개	20개	30개	40개	50개	60개	70개	80개	90개	100개
0.0	0.72	0.70	0.64	0.69	0.71	0.71	0.67	0.67	0.67	0.67
0.1	0.57	0.50	0.52	0.51	0.55	0.55	0.51	0.51	0.52	0.52
0.2	0.43	0.41	0.39	0.40	0.42	0.40	0.39	0.42	0.43	0.43
0.3	0.42	0.39	0.36	0.35	0.35	0.35	0.35	0.35	0.36	0.37
0.4	0.40	0.32	0.28	0.27	0.28	0.27	0.28	0.27	0.27	0.27
0.5	0.26	0.23	0.25	0.24	0.25	0.24	0.24	0.23	0.23	0.24
0.6	0.23	0.21	0.21	0.19	0.19	0.19	0.19	0.20	0.20	0.19
0.7	0.20	0.18	0.17	0.17	0.15	0.15	0.15	0.15	0.16	0.16
0.8	0.13	0.14	0.12	0.11	0.11	0.08	0.08	0.08	0.08	0.08
0.9	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.06
1.0	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
평균	0.311 (+35.2) (+15.2)	0.288 (+25.2) (+6.67)	0.274 (+19.1) (+1.48)	0.273 (+18.7) (+1.11)	0.280 (-21.7) (-3.70)	0.274 (+19.1) (+1.48)	0.268 (+16.5) (-0.74)	0.270 (+17.4) (+0.00)	0.272 (+18.3) (+0.74)	0.274 (-19.1) (+1.48)
재현률	0.896	0.931	0.960	0.972	0.975	0.979	0.980	0.981	0.985	0.989

16.4%의 성능 향상을 보이고 있다. 또한, KTSET 2.0의 경우에는 전체 구간에서 성능이 향상되는 것을 볼 수 있으며, 특히 40%이상 질의를 확장할 경우에는 재현률과 정확률에서 Dec-Hi방법과 비교하여 높은 성능 향상을 이루었으며, 최고 정확률에서는 17.4%의 성능 향상을 보이고 있다. 그러므로 표 3에서의 실험 결과를 종합해 보면, 적합 피드백 문서내에서 용어들의 발생 빈도수(TF)만을 이용하여 원 질의어-후보 용어간의 유사도를 산정할 수 있으며, 유사도를 이용하여 용어의 가중치를 산정할 경우 Dec-Hi 방법보다 더욱 정확한 가중치 부여가 이루어짐으로서 성능의 향상을 이룰 수 있음을 볼 수 있다. 또한 유사도를 기준으로 질의를 확장하면 50% 이상 확장할 경우 가장 우수한 성능을 나타내므로, 전체 후보 용어들중 90%이상을 확장해야하는 Dec-Hi 방법과 비교하여 저장 공간과 실행 속도에서 현격한 성능 개선 효과를 기대할 수 있다.

표 4에서는 본 논문에서 제안하는 용어 발생 유사도와 퍼지 추론을 이용한 질의 용어 확장 및 가중치 재산정 방법에 대한 결과를 보여주고 있으며, 정확률 향상 정도는 Dec-Hi방법의 최고 정확률에 대한 향상 정도와 용어 분포 유사도를 이용한 방법의 최고 정확률에 대한 향상 정도를 나타내고 있다. 또한 표 4에서는 저

장 공간과 수행 속도를 고려하여 퍼지 추론에 의해 산정된 후보 용어의 관련정도를 기준으로 최대 100개까지 질의어를 확장하였다.

표에서는 KTSET 1.0의 경우 70개, KTSET 2.0의

경우 10개만을 확장할 경우 가장 우수한 성능을 나타내고 있으며, 성능 향상 정도에서도 Dec-Hi방법과 비교하여 KTSET 1.0의 경우 최고 39.5%, KTSET 2.0의 경우 35.2%의 성능 향상을 보이고 있다. 또한 표 3의 용어 발생 유사도를 이용한 방법과 비교해서는 KTSET 1.0에서는 최고 19.8%, KTSET 2.0에서는 최고 15.2%의 정확률 향상을 보이고 있으며, 확장되는 용어의 수에서도 KTSET 1.0의 경우 50개만을 확장할 경우 12.1%의 성능 향상이 이루어지며, KTSET 2.0에서는 10만을 확장할 경우에 15.2%의 성능 향상을 기대할 수 있다. 그리고 표 3에서와 같이 용어 분포 유사도를 이용한 방법은 Dec-Hi방법의 최고 정확률과 비교하여 40% 이상을 확장할 경우 성능의 향상을 기대할 수 있지만, 표 4의 퍼지 추론을 이용할 경우 어떠한 경우에도 Dec-Hi 방법보다는 성능이 월등히 향상되는 것을 볼 수 있다. 그러나, 표 4에서는 KTSET 2.0에서 후보 용어를 10개 확장할 경우 표 3과 비교하여 재현률이 미세하게 저하되는 현상이 나타나고 있는데, 이러한 현상은 그림 2의 추론 규칙에 의하여 IDF가 낮은 용어들은 일반 용어들로 가정하고 확장을 하지 않기 때문이다.

그러므로, 표 3과 표 4의 실험 결과를 종합하면, 용어 발생 유사도만을 이용하여 후보 용어의 가중치를 산정하는 방법은 Ide Dec-Hi 방법보다는 우수한 성능을 나타내고 있지만, 좀더 정확한 관련정도 산정을 할 필요가 있음을 알 수 있고, 용어의 발생 유사도를 IDF, 피드백 문서내에서의 DF와 결합하여 퍼지 추론에 이용할 경우 더욱 정확한 관련 정도 산정이 이루어짐으로써 용어 발생 유사도만을 이용한 경우보다 정확률, 저장공간, 실행 속도에서 월등히 높은 성능 향상을 기대할 수 있다. 또한 본 논문에서 제안하는 방법은 용어 분포 유사도를 이용한 방법과 비교하여 퍼지 추론에 추가적인 계산이 필요한 단점이 있는데, 표 4에서처럼 질의어와의 관련 정도를 기준으로 질의를 확장할 경우 전체 후보 용어들 중 70개 미만을 확장할 경우 성능이 현격하게 향상되고 있으므로 최고 성능을 나타내기 위해 확장되는 용어 수를 줄일 수 있으며, 이러한 결과는 검색 시간을 단축할 수 있음으로 추가적인 계산이 필요한 단점을 보완할 수 있다.

5. 결 론

본 논문에서는 원질의어와 질의어로 확장될 수 있는 후보 용어들의 발생 유사도를 적합 피드백 문서내에서의 용어 발생 빈도수(TF)를 이용하여 산정하였으며, 용어의 발생 유사도와 IDF, 피드백 문서에서의 DF를 결합하여 퍼지 추론에 의해 최종 관련정도를 산정하는 방법을 제안하였다. 또한 본 논문에서 제안하는 방법의 성능을 평가하기 위하여 다양한 방법을 이용하여 검색 효율을 평가하였다.

표 2에서는 Ide Dec-Hi방법을 이용하여 용어의 가중치를 기준으로 확장 용어 수를 변경하여 실험하였으며, 표 3에서는 용어의 분포 유사도를 이용하여 가중치를 산정하고 확장 용어의 수를 변경하여 실험하였다. 또한 표 4에서는 용어 발생 유사도, IDF, 피드백 문서내에서의 DF를 퍼지 추론에 적용하여 산정된 후보 용어-원질의 간의 관련 정도를 이용하여 가중치를 산정하고 확장 용어의 수를 변경하여 실험하였다.

결과적으로, 실험에서는 용어 발생 유사도, IDF, 피드백 문서내에서의 DF를 퍼지 추론에 적용하여 관련 정도를 산정할 경우 더욱 정확한 관련정도 산정이 이루어지며, Dec-Hi방법과 비교해서는 KTSET 1.0에서는 최고 39.5% KTSET 2.0에서는 35.2%의 성능 향상이 이루어 졌으며, 용어 분포 유사도를 이용한 방법보다 KTSET 1.0에서는 19.8%, KTSET 2.0에서는 15.2%의 성능 향상을 이루었다. 또한 최고 성능을 나타내기 위해 확장되는 용어의 수를 줄일 수 있으므로 저장 공간, 실행 속도에서도 Dec-Hi방법, 용어 분포 유사도를 이용한 방법과 비교하여 보다 우수한 방법임을 알 수 있었다.

본 논문에서는 용어 발생 유사도를 퍼지 추론에 이용하여 후보 용어-원질의 간의 관련 정도를 산정하였다. 그러나, 이러한 퍼지 추론은 퍼지 소속함수 및 값, 퍼지 규칙에 따라 성능에 많은 영향을 미치게 되며, 문헌의 특성에 따라 이들 퍼지 추론에 필요한 요소들이 변경될 필요가 있다. 그러므로, 이러한 요소들을 문헌의 특성에 따라 자동으로 생성할 수 있는 방법이 연구되어져야 한다.

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] 김주연, 김병만, 박혁로, "용어 분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정", 정보과학회논문지, 제27권 1호, 2000.
[7] Mamdani, E.H., "Application of fuzzy algorithms for control of simple dynamic plant," IEEE Proc. control & Science, Vol. 121, No. 12, pp1585-1588, Dec. 1974.
[8] Croft. W.B . "Experiments with Representation in a Document Retrieval System," information Technology: Research and Development, 2(1), 1-21, 1983.
[9] 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.
[10] Harman. D, "Towards Interactive Query Expansion," Paper presented at ACM Conference on Research and Development in Information Retrieval, Grenoble, France, 1988.
[11] Salton. G. and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Science, 41(4), 228-297, 1990.
[12] 김성혁 외 5인, "자동 색인기 성능 실험을 위한 Test Set 개발", 정보관리 학회지 제11권 1호, 1994.
[13] 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