퍼지 추론을 이용한 질의 용어 확장 및 가중치 재산정

김주연, 김병만, 신윤식
금오공과대학교 컴퓨터공학부
{jykim,bmkim}@cespc1.kumoh.ac.kr

Query Term Expansion and Reweighting by Fuzzy Infernce

Ju Youn Kim, Byeong Man Kim, Shin Yoon Sik
Dept. of Computer & Software Engineering, Kumoh National University of Technology

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 질의로 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 질의에 대한 최종적인 관련 정도를 산정 하였으며, 피드백 문서들에서의 가중치와 관련 정도를 결합하여 후보 용어들의 가중치를 산정 하였다.

1. 서론

정보 검색 시스템은 질의로 사용한 용어가 문서에서 발생할 경우에만 검색이 가능하며, 의미가 유사한 동의어를 이용하여 질의를 할 경우 질의에 적합한 문서일 경우에도 검색을 할 수 없는 용어 불일치 문제를 내포하고 있다. 이러한 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[1].

용어 불일치 문제를 해결하고 검색 성능을 향상시키기 위하여 사용자 피드백을 기반으로 질의를 수정하는 3가지 방법이 제안되었으며, 용어 가중치 재산정 및 질의 용어 확장 방법이 평균 정확도 면에서 가장 우수한 성능을 나타냈다[2].

본 논문에서는 사용자의 적합 피드백을 이용한 용어 가중치 산정 및 질의어 확장을 위해 기존의 용어 발생 분포 유사도를 이용한 관련 정도(이하 유사도) 산정 방법[3]을 개선한 퍼지 추론[4]을 이용한 후보용어 원질의간의 관련 정도 산정 방법에 대하여 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구의 문제점을 지적하고, 3장에서는 본 논문에서 제안하는 퍼지 추론을 이용한 관련 정도 산정 방법에 대하여 설명하고, 4장에서는 재현율과 정확율을 사용하여 제안하는 방법과 기존 방법의 성능을 비교 평가하였다. 마지막으로 4장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구 및 문제점

질의 용어 가중치 재산정 및 질의 용어 확장 방법들중에서 Salton과 Buckley는 6개의 실험 문서들을 통해서 Ide Dec-Hi, Ide Regular, Rocchio방법들을 실험하였다[2]. 이와 같은 세 가지 방법의 기본적인 연산 절차는 문서 벡터와 원래의 질의 벡터를 병합하는 것이다. 이것은 적합 문서들에 해당 질의어의 발생으로부터 가중치를 부가하[illegible] 비적합 문서에 대하여 가중치를 줄여줌으로서 질의어에 자동적으[illegible] 가중치가 다시 부여되도록 한다. 질의어는 원래의 질의에 없었던 어에 대하여 적합 문서에서 발생한 것인지, 아니면 비 적합 문서에[illegible] 발생한 것인가의 판단에 따라 양의 가중치와 음의 가중치가 부[illegible] 다. 또한 음의 가중치를 가지는 용어는 질의어로 확장되지 않는다.

Ide dec-hi방법은 사용자에게 보여진 집합 내에서 검색되어진 적합 문서들 전체에 대한 평가 대신에 적합 평가에 대한 최상위의 적합 문서를 사용하며, Rocchio 방법은 적합과 비적합 문서의 [illegible] 정도의 조정을 허락하였다. Salton과 Buckley의 실험 결과는 여섯[illegible] 실험문서 집합에서 거의 차이가 없었지만, Ide dec-hi 방법을 사[illegible] 을 때 가장 좋은 결과를 얻었다.

Salton과 Buckley의 실험 결과에서 가장 우수한 성능을 나타[illegible] Ide Dec-Hi 방법에서는 확장될 용어들의 가중치를 부여할 때 [illegible] 질의어와의 관련성, 질의어로서의 중요성을 반영하지 못하는 문제[illegible] 을 가지고 있다. 즉, Dec-Hi방법에서는 적합, 비적합 문서내의 [illegible] 빈도(TF)와 전체 문헌에서의 역문헌 빈도수(IDF)만을 이용하여 [illegible] 의 가중치를 산정하게 되므로 적합 문서들에서 TF가 높은 용어는 [illegible] 은 가중치를 가지게 되고, 이러한 결과는 질의어로서 중요하지 않[illegible] 용어들도 단지 적합 문서내에서만 자주 발생하게 되면 높은 가중치[illegible] 부여받게 된다는 것을 의미한다. 그러므로, 이러한 Dec-Hi방법의 [illegible] 재점을 개선하기 위하여 사용자의 연관 피드백 문서들에서 후보용[illegible] 들과 원 질의와의 발생 분포 유사도를 이용하여 후보용어-원질의[illegible] 관련 정도를 산정하고, 가중치를 재산정하는 방법이 제안되었다[3].

제안된 방법[3]에서는 사용자의 연관 피드백 문서들에서 발생[illegible] 모든 용어들을 질의로 확장될 수 있는 후보용어들로 선택하고, [illegible] 들에서의 용어 발생 유사도를 이용하여 후보용어-원질의와의 관[illegible] 도를 산정 하였으며, 산정된 관련 정도와 피드백 문서들에서의 [illegible] 치를 결합하여 후보 용어의 가중치를 재산정하였다. 이 방법에서 [illegible] 관련 정도를 기준으로 질의를 확장할 경우 Dec-Hi 방법과 비교[illegible]

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, 후보용어-원질의간의 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.

3. 퍼지 추론을 이용한 용어 가중치 재산정

이 장에서는 후보 용어 선택 방법과 기존 연구에서 산정된 후보용어-원질의간의 관련정도[3]를 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 이용하여 검색된 문서중 상위 10위 이내의 적합 문서들을 피드백 문서로 가정하고, 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.

3.2 퍼지 추론을 이용한 관련정도 산정

기존에 제안된 후보용어-원질의간의 관련 정도(이하 유사도) 산정 방법[3]은 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 기존 연구에서 산정된 유사도(S)와 후보 용어의 역문헌 빈도수(IDF), 피드백 문서중에서 후보 용어가 발생한 문서 수(DF)를 퍼지 추론에 적용하여 각 피드백 문서에서 후보용어-원질의간의 관련정도를 재산정하였다. 이때, 후보 용어-원질의간의 관련정도를 퍼지 추론을 이용하여 산정한 이유는 퍼지 이론을 이용할 경우 각 요소 값들을 인간의 직관적인 사고에 반영하여 쉽게 해석할 수 있고, 인간의 직관적인 사고를 퍼지 규칙으로 간단하게 작성할 수 있기 때문이며, 이러한 퍼지 이론에서 제어 값의 추론은 퍼지 추론을 이용하기 때문이다.

1) 퍼지 입출력 변수

그림 1에서는 본 논문에서 사용한 퍼지 입출력 변수들을 나타내고 있다. 그림 1의 (a)에서 입력 변수 S는 기존 연구에서 산정된 유사도를 사용하고 있으며, 소속 함수는 4개를 사용하였다. 이때 Z 소속 함수를 사용한 이유는 피드백 문서에서 후보용어가 발생하지 않을 경우 원질의에 대한 후보용어의 관련정도는 0으로 산정되기 때문이며, 각 소속함수들의 범위는 직관적인 값으로 설정하였다.

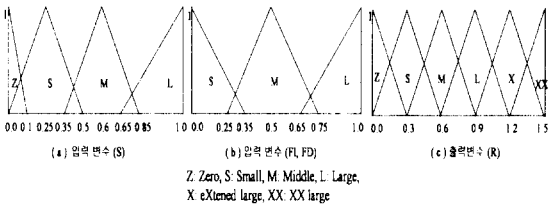

그림 1 퍼지 입출력 변수의 소속함수

(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 중요성를 평가하기 위하여 사용하였으며 0.0 - 1.0의 값으로 정규화하기 위하여 (식 1)를 사용하였다. 또한 FD는 사용자의 피드백 문서들에서 후보 용어의 중요성을 평가하며, FI와 결합하여 후보용어가 질의로서 중요하지 않은 일반 용어일 가능성을 평가하게 된다. (식 2)에서는 DF를 0.0 - 1.0으로 정규화하기 위하여 사용된 식을 나타내고 있다.

$$FI_i = \frac{IDF_i}{MIDF} \qquad \text{(식 1)}$$

FI_i : 후보 용어 i의 퍼지 역문헌 소속함수 값

IDF_i : 후보 용어 i의 역문헌 빈도수

$MIDF$: 후보 용어들중의 최대 역문헌 빈도수

$$FD_i = \frac{FCDF_i}{TFDF} \qquad \text{(식 2)}$$

FD_i : 후보 용어 i 의 퍼지 피드백 소속함수 값

$FCDF_i$: 후보 용어 i가 발생한 피드백 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, 소속함수 값들을 정규화하지 않은 이유는 퍼지 추론에 의해 X와 XX 함수로 추론될 경우 Ide Dec-Hi방법에 의한 가중치 산정보다 부가적인 가중치를 부여하기 위함이다.

2) 추론 규칙

그림 2에서는 본 논문에서 사용한 퍼지 추론 규칙들을 나타내고 있으며, 후보용어-원질의 간의 유사도(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 후보용어-원질의간의 유사도(S)가 0.0일 경우에는 FI와 FD에 관계없이 출력이 Z 소속 함수를 가지도록 하였으며, 유사도 S와 FI, FD로서 출력 변수의 소속함수를 결정하기 어려울 경우 출력 변수의 소속 함수는 유사도 S의 소속함수를 가지도록 추론 규칙을 작성하였다. 규칙에서는 FI가 낮고(S) FD가 높을 경우(L) 이러한 용어들은 질의어로서 중요하지 않은 일반적인 용어일 가능성이 높으므로 유사도에 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.

S=Z

FI \ FD	S	M	L
S	Z	Z	Z
M	Z	Z	Z
L	Z	Z	Z

S=S

FI \ FD	S	M	L
S	Z	S	Z
M	S	M	L
L	S	L	X

S=M

FI \ FD	S	M	L
S	Z	M	S
M	M	L	X
L	M	X	XX

S=L

FI \ FD	S	M	L
S	Z	L	M
M	L	X	XX
L	L	XX	XX

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 무게중심(center of gravity) 방법[4]을 사용하였다.

3.3 후보 용어의 가중치 산정 및 질의어 확장

퍼지 추론을 이용하여 산정된 후보용어-원질의간의 관련정도를 이용하여 후보용어의 가중치를 산정하고 질의를 확장하는 방법은 Ide Dec-Hi방법을 변형한 (식 3)을 사용하였다.

$$wt_i = \sum_{k=1}^{n} (wt_{ik} * R_{ik}) \quad \text{(식 3)}$$

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

wt_i : 전체 피드백 문서에서 후보 용어 I의 가중치

wt_{ik} : 피드백 문서 k에서 후보 용어 t_i의 가중치

R_{ik} : 피드백 문서 k에서 후보 용어 t_i의 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 3)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.

4. 성능 평가

4.1 실험 및 평가 방법

본 논문에서는 실험을 위하여 한국어 테스트콜렉션인 KTSET 1.0, KTSET 2.0을 사용하였으며, 원 질의어를 이용한 검색 결과에서 상위 10위 내에 검색된 문서들중 질의에 대한 적합 문서들을 피드백된 문서로 사용하였다. 또한, 재 검색에 의해 추가로 검색된 문서들의 검색 효율을 평가하기 위하여 Residual Collection 방법[5]을 사용하였으며, 검색 효율을 평가하기 위하여 재현율 0.0 - 1.0까지 11개의 재현율에서의 평균 정확율을 사용하였다.

성능의 상대적인 평가를 위하여 Dec-Hi방법[2], 용어 분포 유사도를 이용한 방법[3], 본 논문에서 제안하는 퍼지 추론을 이용한 방법들을 정확률-재현율을 사용하여 평가 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 후보용어-원질의간의 유사도를 기준으로 후보용어들을 비율확장 하였다. 본 논문에서 제안하는 방법에서는 저장 공간과 수행 속도를 고려하퍼지 추론에 의해 산정된 후보용어의 관련정도를 기준으로 최대 100개까지 변화하여 확장하는 방법(이하 제한확장)을 사용하였다.

4.2 실험 결과

본 논문에서 상대적인 평가 기준으로 선정한 Dec-Hi 방법에서는 2개의 test set에 대하여 가중치를 기준으로 전체 후보 용어의 90%이상을 확장할 경우에 재현율과 정확율에서 가장 성능이 우수하였으며, 평균 정확율에서는 최고 25.6%, 23.0%의 정확율을 보였다. 이때, 질의를 90%로 확장할 경우 확장되는 용어는 KTSET 1.0의 경우 질의어당 평균 111.6개, KTSET 2.0에서는 357.3개가 확장된다.

용어 분포 유사도를 이용한 방법에서는 유사도를 기준으로 전체 용어의 50%만을 확장할 경우 Dec-Hi방법의 최고 정확율과 비교하여 KTSET 1.0에서는 10.5%, 2.0에서는 13.0%가 향상되었다. 또한, 최고 성능에 도달하기 위하여 KTSET 1.0에서는 전체 용어의 100%, 2.0에서는 90%를 질의로 확장하여야 하며, 이때 Dec-Hi방법의 최고 정확율과 비교하여 KTSET 1.0에서는 16.4%, 2.0에서는 17.4%가 향상되었다.

본 논문에서 제안하는 퍼지 추론을 이용할 경우 전체 용어중 10개를 질의로 확장하면 Dec-Hi방법의 최고 정확율과 비교하여 KTSET 1.0에서는 14.5%, 2.0에서는 35.2%의 성능 향상을 이루었으며, 용어 분포 유사도를 이용한 방법의 최고 정확율과 비교하여 1.71%, 15.2%의 성능 향상을 보였다. 또한 KTSET 1.0에서 70개, 2.0에서 10개를 질의로 확장할 경우 가장 높은 성능을 나타냈으며, 위에서의 2가지 방법과 비교하여 KTSET 1.0에서는 39.5%, 19.8%, 2.0에서는 35.2%, 15.2%의 성능 향상을 나타냈다.

5. 결론 및 향후 연구 과제

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.

결과적으로, 실험에서는 용어 발생 유사도, IDF, 피드백 문서내에서의 DF를 퍼지 추론에 적용하여 관련 정도를 산정할 경우 더욱 정확한 관련정도 산정이 이루어지며, Dec-Hi방법과 비교해서 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 후보용어-원질의간의 관련 정도를 산정하였다. 그러나, 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 연구되어야 한다.

6. 참고 문헌

[1] 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.

[2] Salton. G. and C. Buckley, "Improving Retrieval Performance by Relevance Feedback", Journal of the American Society for Information Science, 41(4), 228-297, 1990.

[3] 김주연, 김병만, 박혁로, "용어분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정", 정보과학회논문지, 게재예정(2000년 3월)

[4] 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.

[5] Harman. D, "Towards Interactive Query Expansion", Paper presented at ACM Conference on Research and Development in Information Retrieval, Grenoble, France, 1988.