퍼지 논리를 이용한 질의어 확장과 문서 분류

은희주○†, 이기영††, 김용성†
†전북대학교 컴퓨터과학과
††원광보건대학 정보통신과

Query Extending and Document Classification Using Fuzzy Logic

Hye-Ju Eun○†, Ki-Young Lee††, Yong-Sung Kim†
†Dept. of Computer Science, Chonbuk National University
†† Dept. of Information & Communications, Won Kwang Health Science College

요약

본 연구에서는 인터넷 상의 많은 문서들 중에서 사용자에게 보다 적합한 문서를 제공하기 위해 퍼지 관계성을 이용하여 검색 결과 집합의 문서에서 추출한 키워드간의 유사 클래스를 생성한다. 또한, 기존의 키워드 직접 매칭에 의한 검색 방법의 단점이라 할 수 있는 의미적 관계를 가지는 문서에 대한 검색 방법도 제안한다. 생성된 유사 클래스는 사용자의 질의를 확장하여 사용자의 관심도를 보다 많이 반영하게 되고, 그 질의어가 포함된 단어나 구의 발생 빈도수가 높은 문서에 대해 의미적으로 서로 연결시켜 분류한다. 본 연구에서 제안한 알고리즘에 의해 문서를 사용자 관심 정도로 분류, 카테고리를 생성하여 검색 효율을 증대시키고 사용자의 요구에 적합한 결과를 제공하고자 한다.

1. 서론

현대사회는 컴퓨터의 사용이 보편화되고 네트워크 기술의 발달과 인터넷의 보급으로 인하여 다양하고 많은 정보를 손쉽게 얻을 수 있고 이용·활용하게 되었다. 하지만 방대한 데이터 저장소에서 기존의 검색 엔진을 통해 검색된 자료 집합은 항상 사용자의 관심도가 높은 정보만을 제공하지는 못한다.

또한, 현재 많이 사용하고 있는 검색엔진은 단순히 사용자의 질의어가 문서 내에서 동일한 용어의 발생 여부를 반영하여 문서를 검색하므로 사용자의 원하는 정보와 무관한 문서까지 검색결과로 보여주게 된다.

따라서, 본 연구에서는 이러한 질의어와의 직접 매칭으로 인한 검색의 문제 해결 방법으로 먼저, 사용자 관심을 표현하는 질의어와 유사한 키워드끼리 구성되는 유사 클래스와 호환 클래스를 생성하여 사용자에게 보다 적합한 문서를 검색 결과 집합으로 제공하고자 한다.

또한, 동일한 문서뿐만 아니라 다른 문서에서 어떤 단어나 구가 동시에 발생하는 빈도수가 높을수록 그 단어를 포함한 문서끼리는 서로 의미적으로 연결된다는 사실에 기반을 두고 질의어와 직접 매칭 되지 않은 문서까지 검색할 수 있도록 한다.

2. 관련연구

본 연구의 질의 확장에 관련된 연구로는 사용자 개인의 관심도와 선호도를 반영하기 위해 개인, 그룹의 프로파일을 작성하여 질의를 입력하면 작성된 프로파일이나 시소러스를 참조하여 질의를 확장하는 기법이 있다[1][3].

또한, 문서를 분류하기 위해 많이 사용되는 벡터 유사도 방법은 문서의 키워드들로 구성된 문서 벡터와 카테고리의 색인어로 구성된 벡터를 이용하여 두 벡터 사이의 각이 작을수록 높은 유사도를 갖도록 하는 기법이다[1].

이 기법은 키워드간의 동의어와 불용어 처리가 어렵고

사용자의 관심을 표현하기 위한 방법이 역문헌 빈도에 의한 키워드 추출에 한정되어 있을 뿐만 아니라, 학습과정이 비교적 단순하여 사용자의 관심도를 충분히 반영하지 못한다는 문제점이 있다.

3. 퍼지 관계(Fuzzy relations)

본 절에서는 질의 확장을 위해 사용자 질의와 유사한 키워드끼리 구성되는 유사 클래스와 문서의 카테고리 집합 생성에 기반이 되는 퍼지 관계와 연산에 대하여 언급한다.

전체집합 A의 임의 퍼지 집합 X, Y에 대해 X의 임의 원소 x와 Y의 임의 원소 y 사이에 관계는 순서쌍(x, y)로 표시하고 (x, y)의 모임을 관계 R로 표시한다. 따라서 관계 R의 소속함수는 $\mu_R: X\times Y\rightarrow[0,1]$이다[2][4].

3.1 유사 관계(similarity relation)(≅)

퍼지 집합 X에 정의된 퍼지관계 $R\subseteq X\times X$가 임의의 $x,y,z\in X$에 다음과 같이 반사, 대칭, 이행 관계를 만족하면 유사 관계(≅)라 한다[2][4].

$$\mu_{\cong}(x,x)=1,\ \mu_{\cong}(x,y)=\mu_{\cong}(y,x),$$

$$\mu_{\cong}(x,z)\geq\min\{\mu_{\cong}(x,y),\ \mu_{\cong}(y,z)\}$$

3.2 호환 관계(tolerance, compatibility) (≈)

호환관계(≈)는 다음과 같이 반사와 대칭 성질을 만족하고 이행관계는 만족하지 않는다[2][4].

$$\mu_{\approx}(x,x)=1,\ \mu_{\approx}(x,y)=\mu_{\approx}(y,x)$$

퍼지 호환 관계 R이 임의의 집합 X에 주어지면, 호환 관계를 만족하는 부분 집합들로 분할이 되는데 이렇게 얻어진 부분 집합들을 퍼지 호환 클래스라 한다.

3.3 α-cut

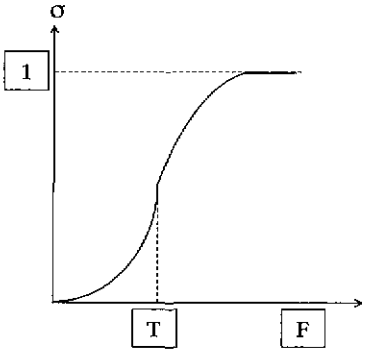
일반적으로 어떠한 데이터에서 일정수준 이상의 데이터를 선별하고자 할 때 α-cut을 적용한다[2][3]. 본 연구에서 질의 확장에 이용된 유사 클래스와 호환 클래스는 전체 키워드 집합 중에서 퍼지 유사 관계와 호환 관계에 의한 유사도 값이 일정수준(α수준) 이상이 되는 키워드를 찾아 유사 클래스를 생성할 수 있다.

3.4 Sigmoid 함수

sigmoid 함수는 일반적인 정수 값을 퍼지 소속 정도로 사상시키기 위한 소속함수로써 다음과 같은 속성을 만족한다[4][5].

$$\sigma: R^{+}\rightarrow[0,1]$$
$$\sigma(F_1)\ >\ \sigma(F_2)\Leftrightarrow F_1\ >\ F_2$$
$$\frac{d^2(\sigma)}{dF2}\geq 0\Leftrightarrow F\leq T_F \text{ and } \frac{d^2(\sigma)}{dF2}\leq 0\Leftrightarrow F\geq T_F$$

시그마(σ)는 S 형태의 곡선이 되고 키워드의 발생 빈도수가 문서의 중요도 정도를 나타낸다.



[그림 1] 빈도수에서 퍼지 소속 정도에의 사상 함수

4. 질의 확장과 문서 분류의 알고리즘

본 논문에서 제안하는 퍼지 관계성을 이용하여 질의 확장과 문서 분류의 알고리즘은 다음과 같다.

Step 1. 사용자 질의에 의한 결과 집합의 문서에서 키워드 추출

Step 2. 유사 클래스 생성

1. 추출한 키워드간의 유사도 값 측정
   /*키워드와 키워드간의 관계 R을 구한다. */
2. α-cut을 이용하여 α값 이상의 키워드 추출

Step 3. 유사 클래스에 의한 질의 확장

Step 4. 문서에서 확장된 질의어의 빈도수 계산

Step 5. 카테고리 생성

1. Sigmoid 함수를 이용한 퍼지 소속 정도 계산
2. α-cut을 이용하여 α값 이상의 문서 추출

5 실행 시나리오

본 절에서는 위의 4절의 알고리즘을 적용하여 질의를 확장하고, 확장된 질의가 문서에서 동시에 존재하는 발생 빈도수를 퍼지 소속 정도로 사상시켜 질의에 대한 직접 매칭뿐만 아니라 의미적으로 연결된 문서들을 분류하고자 한다.

5.1 질의 확장 시나리오

다음은 'Altavistar' 검색엔진을 이용하여 검색 결과 집합의 문서를 가지고 실행한 예이다.

먼저 질의 입력에 "인공지능" 키워드를 입력했을 때 결과 집합의 문서에서 추출한 키워드는 "전문가 시스템, 신경망, 학습, 정보처리, 퍼지논리, 정보검색"이다.

사용자의 관심을 나타내는 "인공지능"의 질의와 일정 수준(α-cut)이상이 되는 키워드들의 유사클래스를 생성하기 위해서 각각의 키워드들에 대하여 2.1절에서 소개한 유사관계를 이용하면 다음 [표 1]과 같은 유사도 측정값을 얻을 수 있다.

[표 1] 키워드들간의 유사도 값

	신경망	전문가 시스템	퍼지 논리	정보 처리	학습	정보 검색
신경망	1	0.8	0.2	0.4	0.2	0
전문가 시스템	0.8	1	0.2	0.4	0	0
퍼지 논리	0	0	1	0	1	0.5
정보 처리	0.4	0.4	0	1	0	0
학습	0	0	1	0	1	0.5
정보 검색	0	0	0.5	0	0.5	1

만약 α = 0.5에서 퍼지 집합을 분할하면 0.5 정도 이상으로 유사한 키워드들로 구성된 유사 클래스는 {신경망, 전문가 시스템}, {정보처리}, {퍼지논리, 학습, 정보검색}이 된다. 또한, 퍼지 집합을 이루고 있는 각 원소 즉, 키워드와 유사한 집합을 이룰 수 있다.

예를 들어 "정보처리"와 유사 관계를 맺고 있는 유사 클래스는 {(신경망, 0.4), (전문가 시스템, 0.4), (정보처리, 1)}이다. 따라서 사용자 질의인 "인공지능"에 의해 검색된 결과 집합의 문서에서 추출한 키워드에 대한 유사한 클래스를 생성함으로써 보다 사용자의 관심을 적절하게 표현하기 위해 질의를 확장할 수 있다.

5.2 문서 분류의 실행 시나리오

1. 유사 클래스를 구성하고 있는 키워드들에 의해 검색된 문서의 결과 집합에서 불용어 및 상대적 불용어 즉, 발생 빈도수는 높지만 의미의 중요도가 낮은 용어를 제외한 키워드(단어, 구)를 추출한다.
2. 추출한 키워드(단어, 구)에 대하여 문서에서 동시 발생 빈도수를 구한다.
3. 문서와 키워드간의 관계를 정의하기 위해서 행(raw)은 키워드를 열(colum)은 검색결과에 대한 키워드 빈도수를 나타내는 키워드-문서 행렬을 구한다.
4. 위의 4에서 구한 키워드-문서간의 빈도수를 sigmoid 함수를 이용하여 문서에서의 퍼지 소속 정도로의 변환을 수행한다.

   다음 [표 2]는 본 연구에서 추출한 키워드(단어, 구)의 빈도수(F)에 대한 시그마(σ), 즉 퍼지 소속 정도를 정의한 것이다.

   [표 2] 빈도수에 대한 σ 값

F	0	1	2	3	4	5	6	7	8
σ	0	0.1	0.3	0.5	0.8	0.92	0.94	0.97	1

5. 위의 5에서 구한 퍼지 소속 정도를 가지고 일정한 수준(α-cut) 이상이 되는 문서를 찾아 전체 문서 집합 중 유사 문서 집합을 구한다.

6. 결론 및 향후 연구과제

본 논문에서는 검색 엔진을 사용하여 검색된 결과집합에서 추출된 문서들의 키워드에 대해 퍼지 관계성을 이용, 유사클래스를 생성하고 질의어를 확장하였다.

또한, 확장한 질의를 통해 검색된 문서에 대하여 키워드(단어, 구)를 추출하여 문서들에서 동시 발생 빈도수를 구해 질의에 직접 매칭 되지 않은 문서까지 분류할 수 있도록 하였다. 본 논문의 향후 연구과제는 사용자 질의어와 문헌내의 색인어 사이의 퍼지 매트릭스를 구성하여 분류된 카테고리의 색인어를 자동으로 추출할 수 있도록 하는 것이다.

[참고 문헌]

[1] 하얀, 최봉진, 김용성, 김순기, "2단계 필터링을 이용한 문서 선별 및 순위", 한국정보과학회 봄 학술발표논문집(B) 제26권 제1호, 1999. 4.

[2] 이관형, 오길록, "퍼지이론및응용",홍릉과학출판사, 1991

[3] 최봉진, 하얀, 황용주, 김용성, Fuzzy Logic을 기반으로 한 SDI 서비스 설계" 한국정보과학회 가을 학술발표논문집(I), 제25권, 제2호, 1998. 10

[4] László T Kóczy, "Information retrieval by fuzzy relations and hierarchical co-occurrence", 1997

[5] László T. Kóczy, Tamás D. Gedeon, Judit A. Kóczy, "The construction of fuzzy relational maps in information retrieval", IETR98-01.