사용자의 선호도를 반영한 확장 퍼지 정보 검색 시스템의 설계

Design of a Extended Fuzzy Information Retrieval System using Users' Preference

김대원†, 이광형
한국과학기술원 전자전산학과 전산학전공, AITrc, 인공지능 연구실
Daewon Kim†, Hyung Lee-Kwang
Depart. of Electrical Engineering and Computer Science
Division of Computer Science, AITrc, KAIST

요 약

정보 검색 시스템의 목표는 사용자가 원하는 정보를 빠른 시간 내에 효율적으로 검색하는 것이다. 이를 위해 불리언 모델, 벡터 모델을 비롯한 기존의 많은 검색 모델들과 퍼지 이론에 기반한 퍼지 검색 모델들이 제안되어져 왔다. 그러나 기존의 모델들은 관련 문서를 검색하는 데 있어서 사용자의 선호도를 반영하지 못하는 한계점을 지닌다. 본 논문에서는 기존의 퍼지 검색 모델의 단점을 보완하기 위해서 확장 퍼지 검색 모델을 제안하고 설계하였다. 제안하는 모델은 색인어와 문서 가중치의 유사도를 결정하는 데 있어서 사용자의 선호도를 반영할 수 있도록 설계하였다.

I. 서론

정보검색(information retrieval)과 관련된 연구는 최근 인터넷을 통한 네트워크의 발달로 말미암아 그 중요성이 갈수록 높아지고 있다. 정보 검색은 검색의 대상이 되는 객체가 단순한 데이터(data)의 모음이 아니라, 사용자가 원하는 정보(information)라는 측면에서 전통적인 데이터베이스 검색과는 구별된다. 이와 관련된 연구는 방대한 정보의 효과적인 저장을 위한 정보축적(information repository)과 저장된 정보를 효율적으로 검색하는 정보검색의 두 분야에서 활발히 진행되고 있으며, 본 논문에서는 효율적인 정보 검색에 관한 새로운 방법론을 제안한다.

정보검색을 위한 방법론 및 모델링에 관한 연구는 정보검색 초창기부터 지금까지 대표적으로 연구되고 있는 분야이다. 대표적인 모델로는 불리언 모델, 벡터 모델, 베이지언 넷을 이용한 추론 모델 등이 있다. 불리언 검색 모델은 가장 많이 사용되는 모델로서, 모델링의 단순화와 빠른 검색 속도를 보여주는 장점을 지닌다. 그러나 검색되는 문서를 관련 문서와 그렇지 않은 문서라는 이진 논리를 가정함으로써, 검색의 정확률(precision)과 재현률(recall)의 성능 평가에서 다른 모델에 비해 다소 낮은 결과를 보여준다. 이를 보완하고 보다 정확한 정보검색을 위해 벡터 검색 모델이 제안되었다. 벡터 검색 모델에서의 문서는 색인어 항목과 그 항목의 가중치 집합으로 표현된다. 그리고 주어진 문서 집합과 질의 어간에 내적을 계산함으로써 관련 문서를 순위화(ranking)하여 결과를 제공하는 방법이다. 따라서 일반적으로 불리언 모델의 이진 논리보다 향상된 성능을 보여주게 된다. 이 외에도 베이지언 넷을 이용하여 추론 과정을 통한 검색 모델도 제안되었다.

본 논문[1]에서는 기존의 퍼지 이론에 기반한 퍼지 모델의 단점을 보완하는 확장 퍼지 모델을 제안하고 이를 설계, 구현하였다. 현재까지 제안된 퍼지 모델들(확장 불리언 모델로도 알려져 있다)은 불리언 모델이 가지는 단점을 보완하기 위한 것으로서, MMM 모델, PAICE 모델 등이 제안되어져 왔다. 퍼지 모델에서는 문서 가중치(document weight)의 개념이 사용된다. 문서 가중치란 특정 색인어 항목에 의해서

1) 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

결정되어지는 문서의 관련 정도 값을 나타내는 것이다. 주어진 질의어에 대해 관련된 문서를 검색하기 위해서는 모든 문서 집합에 존재하는 문서들에 대해서 질의어와 문서간의 유사도(similarity)를 계산하고, 이를 순위화하여 사용자에게 결과를 알려준다. 그러나 기존의 퍼지 모델들은 색인어 가중치의 사용에 있어서 각 색인어가 가지는 애매하고 모호한 특성을 정확히 반영하지 못하는 한계가 있다. 다시 말해, 사용자가 색인어에 가중치를 부여하고 검색을 시작할 경우 색인어 각각에 대한 가중치도 중요하지만, 전체적인 질의어 관점에서 색인어 가중치의 분포를 고려하는 것 또한 중요하다. 따라서 본 논문에서는 색인어 가중치의 설정에 사용자의 선호도를 반영할 수 있는 확장된 퍼지 정보 검색 시스템을 설계한다.

II. 기존의 검색 모델

본 절에서는 기존의 정보검색 연구에서 사용된 검색 모델을 살펴본다. 전통적인 불리언 모델과 벡터 모델에 관한 내용은 생략하며, 퍼지 집합 모델에 바탕을 둔 MMM 모델과 PAICE 모델을 살펴본다[1].

가. MMM 모델

MMM(Min-Max Model) 모델은 퍼지집합 이론에 근거한 것으로서, 각 문서는 색인어와 문서-색인어 가중치 쌍을 원소로 갖는 퍼지 집합으로 표현된다. 문서 가중치는 색인어에 대한 소속함수 값이 되며, 이 모델의 기본 개념은 다음과 같다. 문서 D가 주어진 경우, 색인어가 $A_1, A_2, \ldots, A_n$이며 각각의 가중치가 $d_{A1}, d_{A2}, \ldots, d_{An}$일 때, 질의어 형태가

$$Q_{or} = A_1 \ or \ A_2 \ or \ \ldots \ or \ A_n$$

$$Q_{and} = A_1 \ and \ A_2 \ and \ \ldots \ and \ A_n$$

위와 같으면, 질의어에 대해서 각 문서의 유사도 값은 식 (1)과 (2)에 의해서 계산되며, 그 유사도 값을 순위화하여 검색 결과를 제공하게 된다.

$$SIM(Q_{or}, D) = C_{or1} \times max(d_{A1}, \ldots, d_{An}) + C_{or2} \times min(d_{A1}, \ldots, d_{An}) \quad (1)$$

$$SIM(Q_{and}, D) = C_{and1} \times min(d_{A1}, \ldots, d_{An}) + C_{and2} \times max(d_{A1}, \ldots, d_{An}) \quad (2)$$

위 식에서 C_{or1}, C_{or2}, C_{and1}, C_{and2}는 각 연산자에 대한 계수 가중치이다 ($C_{or1} > C_{or2}$ and $C_{and1} > C_{and2}$). 이와 같은 기본적인 개념의 확장 불리언 모델에서는 각 문서가 색인어에 대한 가중치(문서 가중치)를 가지며, 이것을 문서-질의어 유사도 값의 결과 검색에 이용하는 것이다.

나. PAICE 모델

PAICE 모델은 앞 절에서 살펴본 MMM과 기본 방법론은 매우 유사하다. MMM 모델의 경우 유사도를 계산할 때, 색인어에 대한 문서 가중치의 최대값, 최소값만을 이용한다. 이에 반해 PAICE 모델은 이를 일반적으로 확장하여 색인어에 대한 모든 문서 가중치를 고려하도록 제안되었다. 식 (3)은 PAICE 모델에서의 유사도 값 모델을 나타낸 것이다.

$$SIM(Q, D) = \sum_{i=1} r^{i-1} d_i \ / \ \sum_{i=1} r^{i-1} \quad (3)$$

III. 확장 퍼지 검색 모델

본 절에서는 먼저 기존의 검색 모델들이 가지는 제한 사항을 예제를 통해 기술하고, 이를 해결하기 위하여 본 논문에서 제안하는 확장된 퍼지 검색 모델에 대해서 살펴본다.

가. 기존 모델의 한계점

앞 절에서 살펴본 퍼지 모델들은 기존의 불리언 모델의 단점을 보완하고, 문서 가중치와 색인어 가중치의 개념을 도입하여 검색된 문서의 순위화를 제공할 수 있도록 제안되었다. 그러나 색인어 가중치의 사용에 있어서 각 색인어가 가지는 애매하고 모호한 특성을 제대로 반영하지 못한 한계가 있다. 사용자가 색인어에 가중치를 부여하고 검색을 시작할 경우 색인어 각각에 대한 가중치도 중요하지만, 전체적인 관점에서의 색인어 가중치의 분포를 고려하는 것 또한 중요한 문제이다. 다음의 예를 살펴보자. 질의어가 아래와 같이 주어진 경우, (질의어는 {색인어, 색인어 가중치}의 집합으로 주어진다)

Query (and) :
(*petri*, 0.8), (*system*, 0.7), (*korea*, 0.3), (*author*, 0.2)

위 질의어는 색인어와 만족도(membership degree)의 쌍을 원소로 갖는 다음과 같은 퍼지집합으로 표현될 수 있다.

Query =
{(*petri*, 0.8), (*system*, 0.7), (*korea*, 0.3), (*author*, 0.2)}

그리고 아래와 같이 퍼지 집합으로 표현된 네 가지의 문서가 문서 집합에 있는 경우를 생각해보자. 이 경우 주어진 질의어에 대해, 검색된 관련 문서의 결과 순위는 어떻게 나타날 것인가? 직관적으로,

$Document_1$: {(*petri*, 0.8), (*system*, 0.7)}	*문서집합*
$Document_2$: {(*petri*, 0.2), (*system*, 0.2), (*korea*, 0.3), (*author*, 0.2)}	
$Document_3$: {(*korea*, 0.7), (*author*, 0.8)}	
$Document_4$: {(*petri*, 0.8), (*system*, 0.7), (*korea*, 0.3), (*author*, 0.2)}	

$Document_4$ 가 가장 관련 있는 문서이며, $Document_3$ 이 가장 관련성이 적은 결과를 나타낼 것은 명백하다. 그러면 $Document_1$ 과 $Document_2$ 의 순위는 어떠한가? 지금까지 제안된 여러 가지 검색 모델에서는 두 문서의 순위가 결정적으로 정해지지 않는다. 기존의 불리언, 벡터 모델과 퍼지 모델에 적용시킬 경우, 두 문서의 관련 유사도 값은 크게 차이나지 않는다. 이것은 유사도 결정 모델에서 관련 유사도 값을 계산할 때, 각 색인어의 가중치를 단조 합산하는 방향으로 전체 값이 결정되기 때문이다. 즉, 가중치가 큰 두 개의 색인어를 가진 $Document_1$ 과 낮은 색인어 가중치 여러 개를 갖는 $Document_2$ 가 비슷한 유사도 값을 산출하게 되는 것이다.

본 논문에서 제안하는 점은 이와 같은 경우 사용자의 선호도(preference), 즉 관심을 색인어 가중치에 반영함으로써 보다 명확한 검색 순위화 결과를 제공하는 것이다. 위 예제에서 사용자는 네 개의 색인어가 모두 검색되지만 각각의 가중치가 낮은 $Document_2$ 보다, 비록 질의어와 정합(matching)되는 색인어의 수는 적지만 그 만족도 값이 높은 $Document_1$ 이 높은 유사도 값을 가지기를 원할 수 있다. 이와 같은 모델은 색인어와 색인어 가중치가 가지는 애매 모호성을 사용자의 선호도를 유사도 모델에 반영함으로써 보다 명확한 검색 결과를 주고자 하는 것이다.

나. 확장 퍼지 모델의 제안

기존의 불리언, 벡터 및 퍼지 모델은 색인어와 색인어 가중치를 이용한 유사도 비교에 사용자의 선호도를 반영하지 못하는 한계점을 지닌다. 이에 본 논문에서는 사용자의 선호도를 반영할 수 있는 기 제안된 유사도(similarity) 비교 알고리즘을 응용하여, 퍼지 정보검색 시스템 설계에 이용하였다[2].

정보검색에서 사용되는 질의어와 문서는 {색인어, 가중치}의 쌍이 {원소, 만족도}의 쌍으로 표현되는 퍼지 집합으로 나타난다. 따라서 질의어와 문서의 유사도 비교는 문서와 질의어에 해당하는 두 퍼지 집합의 유사도 값을 비교하는 것으로 모델링할 수 있다.

유사도 모델링은 전체 문서 공간에 퍼지집합으로 표현된 N개의 문서가 존재할 경우, 질의어로 표현된 퍼지집합과 문서 공간의 N개의 집합 중에서 가장 가까운 집합을 찾는 문제이다. 지금까지 알려진 유사도 비교 방법론들은 완전하지 못하다. 또한 퍼지집합은 앞에서 언급한 바와 같이 보통 집합처럼 명확한 값으로 표현된 것이 아니라, 각 원소들의 가능성 분포를 이용하여 표현된 애매모호한 집합이다. 따라서 두 퍼지집합의 유사도를 비교할 때 전체적인 집합의 가능성 분포를 고려하여야 하며, 문서의 검색시 비교하는 두 집합간에 사용자의 선호도를 반영할 수 있어야 한다. 이것은 퍼지집합을 비교하는 관점, 다시 말해 사용자의 선호도에 따라 다른 비교 결과가 나올 수 있기 때문이다.

따라서 유사도를 이용한 퍼지집합의 비교 척도로서 보다 정확한 결과를 제시하며, 사용자의 선호도와 관심을 반영할 수 있는 새로운 방법론이 필요하다[2]. 기 제안된 알고리즘에서는 사용자의 선호도를 반영하기 위하여, 비교하는 두 집합의 도메인에 대한 선호도(domain preference, $f_{domain}(x)$)와 만족도에 대한 선호도(membership-degree preference, $f_{MV}(y)$)의 개념을 도입하였다 [그림1]. 이렇게 함으로써 사용자의 의도를 퍼지집합 비교에 반영하여 비교하고자 하는 부분에 좀 더 가중치를 둘 수 있게 하였다. 여기서 유사도 비교 값은 도메인 선호도와 만족도 선호도의 적분을 이용한 합성 값으로 계산된다. 본 방법론은 기존의 유사도를 측정하는 방법의 일반화라는 점과 사용자의 선호도를 퍼지집합의 비교에 반영하여 보다 관련성 있는 문서를 검색할 수 있다는 점에서 의미가 있다고 할 수 있다.

다. 제안된 모델의 유사도 척도

도메인 선호도 함수는 식 (4)와 같이 비교하는 두 퍼지집합의 도메인 축에 변화를 주는 선호도이다. 비교하는 도메인 상에서 좀 더 비중을 두고 싶은 영역에 가중치를 두게 된다. 도메인 선호도는 보다 일반적인 퍼지 집합의 유사도 비교를 위해 제안되었으

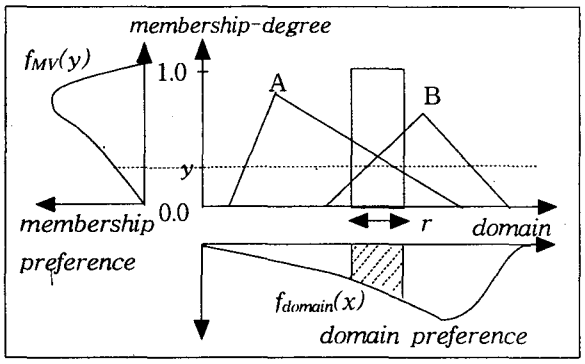

그림 1. 만족도 선호도 함수의 적용

나, 현재 정보검색 시스템에서는 각 색인어에 대한 가중치가 명확히 주어지므로, 사실상 이 선호도 함수는 상수(uniform) 함수의 형태를 가진다. 식 (5)는 도메인 선호도 함수를 이용한 계산 모델을 보여주는 것으로 본 논문에서는 상세한 설명을 생략한다. (그림 1에서 A는 질의어 Q를, B는 문서 D로 가정한다)

$$d\text{Domain} = f_{Domain}(x)dx \tag{4}$$

$$f(y) = \begin{cases} \int_{Domain} \mu_{Q,D}(x,y)\, d\text{Domain} \\ \int_{Domain} \mu_{Q,D}(x,y) f_{Domain}(x)dx \end{cases} \tag{5}$$

만족도 선호도 함수는 식 (6)에 보인 바와 같이 비교하는 두 퍼지 집합의 만족도 축에 변화를 주는 선호도이다. 비교하는 만족도 축([0.0, 1.0]) 상에서 좀 더 비중을 두고 싶은 영역에 가중치를 둘 수 있다. 그림 1에서의 만족도 함수는 만족도 1.0에 가까운 영역에 사용자의 비교 가중치를 둔 것이다. 검색 시스템에서 이와 같이 선호도를 설정할 경우, 질의어와 문서의 색인어 가중치가 높은 문서들에게 검색 결과에서 높은 순위를 주게되는 것이다.

$$d\text{MV} = f_{MV}(y)dy \tag{6}$$

도메인 선호도 적용의 결과 값 $f(y)$를 만족도 축에 대해서 만족도 선호도 값과 적분을 하게 되면 전체 유사도 값이 결정된다. 식 (7)과 (8)에 대한 상세한 설명은 [2]를 참고하기 바란다.

$$SIM_{Q,D} = \int_{MV} f(y)d\text{MV} \tag{7}$$

$$SIM_{Q,D} = \int_{y=0}^{1} f(y) f_{MV}(y)\, dy \tag{8}$$

IV. 실험 및 검증

본 논문에서 구현한 퍼지 정보검색 시스템은 앞 절에서 제안한 유사도 척도 알고리즘을 사용하여 개발하였다. 개발한 시스템은 기존의 불리언, 벡터 모델 그리고 MMM, PAICE 모델을 선택적으로 검색할 수 있도록 구현되었다. 실험 집합으로는 4414개의 문서와 질의어 50개를 가지는 KT-SET을 이용하였으나, 본 논문에서는 지면 관계상 앞 절에서 예로 든 문서

순위	*Boolean*	*Vector*	*MMM*	*PAICE*
1	Doc 2 Doc 4	Doc4 (0.94)	Doc4 (0.38)	Doc4 (0.80)
2		Doc1 (0.87)	Doc3 (0.24)	Doc1 (0.80)
3		Doc2 (0.71)	Doc1 (0.24)	Doc2 (0.80)
4		Doc3 (0.60)	Doc2 (0.23)	Doc3 (0.30)

표 1 기존 검색 모델에서의 실험 결과

순위	*Very High*	*High*	*Median*
1	Doc4 (1.9)	Doc4 (5.0)	Doc4 (4.5)
2	Doc1 (1.6)	Doc1 (4.0)	Doc1 (3.5)
3	Doc2 (0.4)	Doc2 (2.0)	Doc2 (2.0)
4	Doc3 (0.2)	Doc3 (1.0)	Doc3 (1.0)

표 2 제안한 시스템에서의 실험 결과

집합에 대해서 다른 모델과 간략한 비교를 수행하였다. 표 1은 기존의 대표적인 4가지의 검색 모델에 대한 결과를 순위대로 배열한 것이다. 앞 절에서 설명한 바와 같이 $Document_1$과 $Document_2$의 순위는 명확하지 않다.

표 2는 제안한 시스템에서 같은 데이터로 실험한 결과로서, 만족도 선호도 함수의 형태를 세 가지로 구분하여 실험한 결과이다. 만족도 함수(*Very High*)는 그림 1과 같이 만족도 1.0에 많은 가중치를 두고 유사도를 계산하는 모델이다. 따라서 색인어와 문서의 가중치 정합 만족도가 높은 $Document_1$이 보다 높은 관련성을 가지는 것으로 나타난다.

V. 결론

본 논문에서는 사용자의 선호도를 반영하는 퍼지 집합의 유사도 척도를 이용하여, 효율적인 정보 검색을 수행하는 시스템을 설계하고 구현하였다. 제안한 퍼지 모델은 기존의 퍼지 모델을 확장한 것으로서, 사용자의 선호도를 색인어의 가중치 분포에 반영할 수 있도록 하였다. 이것은 기존의 문서가 가지는 문서 가중치 외에도, 사용자가 검색하고자 하는 색인어에 보다 관심과 가중치를 조절할 수 있다는 점에서 의의가 있다.

V. 참고문헌

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Morern Information Retrieval", *Addison-Wesley*, 1999.

[2] 김대원, 신준범, 정성원, 이광형, "소속함수에 사용자의 선호도를 반영한 퍼지 집합의 비교 알고리즘", *퍼지 및 지능 시스템 학회 논문지*, Vol. 10, No. 1, pp. 12-15, 2000.

[3] Jee-Hyung Lee, Comparison, Ranking and Determination of the Ranks of Fuzzy Numbers based on Satisfaction Function, *Ph.D Thesis*, KAIST, 1999.

[3] 이지형, 이광형, "사용자의 관심도를 반영하는 퍼지숫자의 정렬방법", *퍼지 및 지능 시스템 학회 논문지*, Vol. 8, pp. 14-20, 1998.