정보 검색에서 확장 퍼지 개념 네트워크를 이용한 문서 순위 결정 방법

Document Ranking Method using Extended Fuzzy Concept Networks in Information Retrieval

손 현 숙, 정 환 묵
대구효성가톨릭대학교 전자정보 공학부

Son Hyun Sook, Chung Hwan Mook
Faculty of Electronics & info. Engineering
Catholic Univ. of Taeguhyosung

ABSTRACT

정보 검색은 사용자가 원하는 요구에 가장 적합한 정보를 검색할 수 있도록 되어야 한다. 질의어가 문서에 대하여 어느 정도의 유사성을 가지고 존재하느냐를 기준으로 문서를 순서화 한다. 실제 순서화된 문서들을 보면 질의어와는 다른 문서들이 순서화 되는 경우를 볼 수 있다. 본 논문에서는 순서화 되는 문서들 중에서 그 문서들이 질의어와 어느 정도 가까운지를 확장 퍼지 개념 네트워크에 근거한 문서 검색을 위한 퍼지 순위 처리를 위한 방법을 제시한다. 확장 퍼지 개념 네트워크에는 개념들 사이에 4가지의 퍼지 관계를 사용한다. 퍼지 양의 관계, 퍼지 음의 관계, 퍼지 일반화, 및 퍼지 세분화등이 있다. 확장 퍼지 개념 네트워크는 관계 행렬과 관련 행렬로 모델화 한다.

1. 서론

최근 인터넷 이용자 수의 계속적인 성장으로 인해 네트워크 사용인구의 폭발적인 증가가 네트워크 상에 있는 정보량도 기하급수적으로 증가하는 추세이다. 이러한 '정보의 홍수'속에서 사용자의 요구에 해당하는 정보만을 추출하여 사용자에게 제공하여 주는 정보 검색 시스템이 널리 사용되고 있다. 그러나, 방대한 정보들 중에서 사용자가 원하는 중요한 목적 중의 하나는 단순히 사용자 질의어를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의어를 만족하는 정도에 따라 검색된 문서들에게 순위를 부여함으로써

사용자들이 필요한 문서를 얻는데 소모되는 시간을 최소화하는 것이다.[1]

본 논문에서는 사용자의 질의어를 만족하는 문서들에게 순위를 주어 질의어에 가장 적합한 문서를 상위레벨에 유지시키는 방법으로서, 기존의 Boolean 방법은 불확실한 정보를 나타낼 수 없기 때문에 퍼지이론에 근거한 퍼지 정보 검색을 이용한다.

2. 개념과 문서의 표현

2.1 개념 네트워크

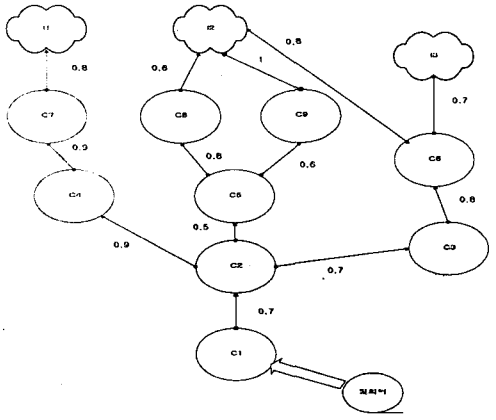

그림 1 개념 네트워크

퍼지 정보 검색[2]을 위한 개념 네트워크를 제안하고, 개념 네트워크는 노드와 방향성 링크로 이루어져 있으며 각 노드는 개념이나 문서를 나타내고, 각 방향성 링크는 한 개의 개념 C_i로부터 한 개의 문서 I_j로 또는 두 개의 개념으로 연결하고, 0 과 1 사이의 값을 부여한다. $C_i \xrightarrow{\mu} C_j$ 는 개념 C_i 로부터 개념 C_j까지 관련 정도가 μ라는 것을 나타낸다. 만일 개념 C_i로부터 개념 C_j까지 관련 값이 $F(C_i, C_j)$이고, 개념 C_j로부터 개념 C_k까지의 관련 값이 $F(C_j, C_k)$라면, 개념 C_j부터 개념 C_k까지의 관련 값은 다음과 같은 표현으로 구할 수 있다.

$$F(C_i, C_k) = \min(F(C_i, C_j), F(C_j, C_k))$$

여기서 문서 I_j에 대하여 문서들은 다음 표현식에 의해 개념 집합의 퍼지 부분 집합으로 정의된다.

$$I_j = \{(C_i, f_{dj}(C_i)) \mid C_i \in C\}$$

$f_{dj}(C_i), f_{dj} : C \rightarrow [0,1]$는 개념 C_i에 관련한 문서 I_j의 관련 정도를 나타낸다.

2.2 확장 퍼지 개념 네트워크

확장 퍼지 개념 네트워크[3]는 제시된 퍼지 개념 네트워크보다 일반적이다. 확장된 퍼지 개념 네트워크의 개념들 사이에 4가지의 퍼지관계가 있다. 양의 조합, 음의 조합, 일반화, 세분화 등이다.

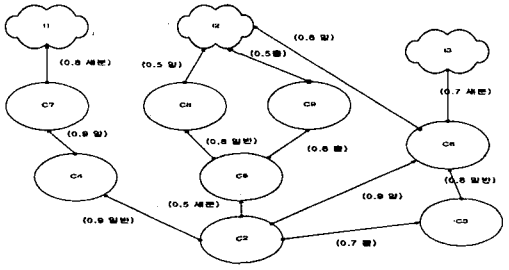

그림 2 확장 개념 네트워크

개념들 사이의 퍼지관계는 아래와 같이 정의한다.

[정의 2.1]

① *양의 조합*
- 퍼지 유사 관계로서 반사적, 대칭적, 이행적 관계를 가진다.

② *음의 조합*
- 퍼지 여(complement)관계로서 비반사적, 대칭적, 반이행적 관계를 가진다.

③ 일반화
- 비반사적, 비대칭적, 이행적 관계를 가진다.

④ 세분화
- 비반사적, 비대칭적, 이행적 관계를 가진다

확장 퍼지 개념 네트워크에서, 개념 c_i 와 개념 c_j 사이의 관련 정도가 μ_{ij} 이고, $\mu_{ij}\in[0,1]$, 또한 개념 c_j 와 개념 c_k 사이의 관련 정도가 μ_{jk} 이고, $\mu_{jk}\in[0,1]$면, 개념 c_i 와 개념 c_k 사이의 관련 정도 μ_{ik} 는 아래와 같이 구할 수 있다.

$$\mu_{ik}=\min(\mu_{ij},\mu_{jk})$$

표 1 퍼지관계 조합

F_{ij} / F_{jk}	양	음	일반	세분
양	양	음	일반	세분
음	음	양	음	음
일반	일반	음	일반	양
세분	세분	음	양	세분

확장 퍼지 개념 네트워크에서는, 개념 c_i 와 개념 c_j 사이의 퍼지관계는 F_{ij} 이고, c_j 와 c_k 사이의 퍼지관계가 F_{jk} 라면, 개념 c_i 와 개념 c_k 사이의 퍼지관계 F_{ik} 는 표1에서 구할 수 있으며, 여기서 양, 음, 일반, 및 세분은 각각 양의 조합, 음의 조합, 일반화, 및 세분화를 의미한다.

3. 관계행렬과 관련행렬

확장된 퍼지 개념 네트워크를 모델화 하는데 관계 행렬과 관련 행렬의 정의[5]를 한다.

[정의3.1] 관련 행렬 A는 퍼지 행렬이며, 원소 $A(c_i,c_j)$ 는 개념 c_i 와 c_j 사이의 관련 정도로 $A(c_i,c_j)\in[0,1]$를 나타내고 있다. A를 관련 행렬로 놓으면

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

[정의 3.2] 관계 행렬 B 는 퍼지 행렬이고, 원소 $B(c_i,c_j)$은 개념 c_i와 개념 c_j 사이의 퍼지관계를 나타내고, 표1과 같으며 개념들 사이에 퍼지 관계가 확실하지 않는 「불확실」을 첨가한다. B을 관계 행렬로 놓으면,

$$B=\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}$$

$b_{ij}\in$[양, 음, 일반, 세분, 불확실]로 나타낸다.

4. 정보 검색을 위한 문서 순위 결정 방법

문서를 표현하기 위하여 문서 관련 행렬 및 문서 관계 행렬[6]을 이용한다.

[정의4.1] I를 문서 집합, $I=\{i_1,i_2\cdots,i_m\}$이고, C를 개념 집합으로 놓으면, $C=\{c_1,c_2\cdots,c_m\}$ 이다. 문서 관련 행렬 D 는 다음과 같다.

D	c_1	c_2	$\cdots$	c_n
i_1	v_{11}	v_{12}	$\cdots$	v_{1n}
i_2	v_{21}	v_{22}	$\cdots$	v_{2n}
$\cdots$			$\cdots$	
i_m	v_m	v_m	$\cdots$	v_m

[정의4.2] I를 문서 집합, $I=\{i_1,i_2\cdots,i_m\}$이고, C를 개념 집합으로 놓으면, $C=\{c_1,c_2\cdots,c_m\}$ 이다. 문서 관계 행렬 M은 아래와 같다.

M	c_1	c_2	$\cdots$	c_n
i_1	r_{11}	r_{12}	$\cdots$	r_{1n}
i_2	r_{21}	r_{22}	$\cdots$	r_{2n}
$\cdots$			$\cdots$	
i_m	r_m	r_m	$\cdots$	r_m

문서 관련 행렬 D 와 문서 관계 행렬 M 에서, 개념과 문서 사이의 관련 정도와 퍼지 관계는 전문가에 의하여 제시된다. 사용자의 질의 Q는 질의 관련 벡터 $\overline{qa}$ 와 질의 관계 벡터 $\overline{qb}$로 표현될 수 있다. 사용자의 질의는 다음과 같다.

$$Q=\{(c_1,\langle x_1,y_1\rangle,(c_2,\langle x_2,y_2\rangle,$$

$$\cdots(c_n,\langle x_n,y_n\rangle\}$$

$$\overline{qa}=\langle x_1,x_2\cdots x_n\rangle$$

$$\overline{qr}=\langle y_1,y_2\cdots y_n\rangle$$

(x,s)와(y,t)를 두 쌍의 값으로 놓고, $x\in[0,1]$, $y\in[0,1]$, $s\in$ {양, 음, 일반, 세분}

그리고 $t\in$ {양, 음, 일반, 세분}

이면, (x,s)와(y,t)사이에 유사도의 정도를 U에 의해 계산할 수 있다.

$$U(\langle x,s\rangle\langle y,t\rangle)=0 \qquad \text{if } s\neq t$$

$$U(\langle x,s\rangle\langle y,t\rangle)=x-y \qquad \text{if } s=t$$

본 논문에서는 유사도 측정을 위하여 유크리드 거리 측정법을 사용하여, 문서 I_i 가 사용자의 질의를 만족시키는 유사도를 다음 식으로 계산 할 수 있다

$$UD(I_i)=\frac{1}{\sqrt{\sum_{\overline{qa(j)}\neq-\text{ and }\overline{qb(j)}\neq-\text{ and }j=1,\cdots n}U^2}}$$

여기서 $UD(I_i)\in[0,1]$이고, 유크리드 거리가 가까울수록 유사도가 높음을 의미하므로, 일반적으로 유크리드 거리의 역수를 취하여 유사도를 얻는다. $UD(I_i)$ 의 값이 크면 클수록, 문서 I_i 가 사용자의 질의사이의 유사도는 더 커진다. 유사도 측정 결과를 토대로 사용자에게 제공될 문서를 선택하는 방법으로 문서의 일정한 비율만큼의 문서를 유사도가 높은 순으로 선택하는 방법과 유사도를 $\alpha-cut$이상인 문서만을 선택하는 방법이 있다. 그러나 $\alpha-cut$를 이용한 방법은 적당한 $\alpha-cut$를 어떻게 정할것인지에 대한 문제가 있다.

5. 결론

Boolean논리와 비교할 때 퍼지논리에 의한 것은 훨씬 유동적이었다. 확장 퍼지 개념 네트워크는 개념들 사이에 4가지의 퍼지관계를 정의하여 사용자로 하여금 좀더 유동적인 방법으로 질의를 수행할 수 있도록 해준다. 그러나 문서의 개수가 증가함에 따라 유사도 측정 계산은 많은 양의 계산과정을 필요로하는 단점이 있었다.

6. 참고문헌

[1] Pattie Maes, "Agents that reduce work and information overload", Communication of ACM Vol.37, No7

[2] G. T. Her and J. S. Ke, "A fuzzy information retrieval system model," in *Proc. 1983 National Computer Symp.*, Taiwan, R.O.C., 1983, pp. 147~151.

[3] Y. J. Homg and S. M. Chen, "Document retrieval based on extended fuzzy concept networks," in Proc. 4th No.1. Conf. Defense Management, Taipei, Taiwan, R.O.C., 1996, vol. 2, pp. 1039~1050.

[4] I. Itzkovich and L. W. Hawkes, "Fuzzy extension of inheritance hierarchies," Fuzzy Sets Syst., vol. 62, no. 2, pp. 143~153, 1994.

[5] S. M. Chen and J. Y. Wang, "Document retrieval using knowledge-based fuzzy information retrieval techniques," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, pp. 793~803, May 1995.

[6]M. Kamel, B. Hadfield, and M. Ismail, "Fuzzy query processing using clustering techniques," *Inf. Process. Manage.*, vol. 26, no. 2, pp. 279~293, 1990.