퍼지 이론을 이용한 웹 에이전트의 정보 분류 평가에 관한 연구

A Study on Information Classification Evaluation of Web Agent Using Fuzzy Theory

김두완*, 정구범**, 정환묵*
대구가톨릭대학교*, 상주대학교**

Doo-Ywan Kim* · Gu-Beom Jeong** · Hwan-Mook Chung*
Catholic University of Daegu*, Sangju University**
kimdy@cu.ac.kr · jgb@sangju.ac.kr · hmchung@cu.ac.kr

요 약

인터넷의 급격한 보급으로 다양하고 많은 종류의 유용한 정보를 이용할 수 있게 되었다. 이와 같은 정보의 바다에서 원하는 정보를 검색하고 이를 관리하고 사용하는 것은 매우 어렵다. 이러한 문제를 해결하기 위해 검색엔진, 메타검색 엔진, 스파이더, 지능 에이전트 혹은 웹 에이전트와 같은 여러 종류의 시스템들이 개발되고 있다. 이와 같은 시스템들은 지능 에이전트로써 정보의 과부하를 피하기 위해 사용되어지고 있다. 소프트웨어 에이전트들을 효율적으로 개선하기 위해서는 검색된 데이터를 표현하고 분류하는 것이 필요하다. 또한, 분류기를 생성할 수 있는 지능 에이전트들의 성능을 개선하기 위해 퍼지 이론을 적용하여, 웹으로부터 다른 검색 정보와의 적합성을 평가하고, 사용자에게 가장 적합한 정보를 분류하기 위한 방법을 제안한다.

Key Words : 퍼지 거리, 정보 분류, 에이전트

1. 서 론

인터넷이 개발된 이후 우리의 생활은 많은 변화를 가져왔다. 과거 오프라인에서 발생하는 모든 행동들은 인터넷의 발달로 인해 온라인 상태에서 즉시 발생한다. 예를 들어, 정보검색, 쇼핑, 교육, 항공사 혹은 열차 버스 예약 등 많은 사항들이 온라인에서 이루어지고 있다. 또한 연결된 사용자에 의해 가격, 시간표, 각종 예약표의 의견을 묻는 것이 가능하다. 그러므로 월드 와이드 웹의 개발은 다양한 종류의 유용한 정보를 발견하는 것이 가능하도록 하였다. 그러나 이와 같은 정보들은 주로 데이터베이스에 저장된 매우 방대한 양의 정보들이기 때문에 저장된 정보를 검색하고 사용하는 것은 심각한 문제가 있다. 이러한 문제들 중 몇 가지를 요약하면, 사이트의 수는 많은 종류의 다양하고 유용한 정보가 기하급수적으로 제공되고, 발견된 문서(데이터)의 수는 거대하며, 회사들에 의해 제공된 정보가 시간에 따라 변화할 수 있으며, 동일한 종류의 정보에 대해 다른 표현들로 사용될 수 있다는 것이다[1,2].

이 문제를 해결하기 위하여 다양한 접근 방법들이 연구되어지고 있다. 먼저, 가장 인기있는 검색 엔진들은 소프트봇, 스파이더 혹은 웜즈와 같이 소프트웨어 시스템들을 사용한 것으로 이들 시스템들은 웹상에 존재하는 모든 정보를 필요에 따라 다시 모으는 것이다. 이들 시스템들은 정보를 요구하는 사용자들에게 원하는 정보를 제공할 수 있도록 완전하고 동적인 데이터베이스를 구축한다. 두 번째로 메타검색 엔진이 있다. 이 시스템은 검색 엔진의 집합을 사용한 시스템들로 검색 작업을 개선하는 또 다른 작업들을 수행한다. 세 번째로 지능 소프트 에이전트는 에이전트 개념을 사용한 접근 방법이다. 여기서 에이전트들은 소프트웨어 응용들을 개발하기 위한 새로운 파라다임이다. 마지막으로 멀티에이전트 시스템은 문제들을 해결하기 위해 협력하도록 지능 에이전트 그룹들의 상호작용을 다루는 응용이다. 멀티에이전트 시스템은 거대한 문제를 해결할 수 있기 때문에 성공적이며, 각각의 에이전트들과 협력할 수 있고, 지식과 지식 등을 공유할 수 있다[3-4].

분석과 분류를 통하여 정보의 과부하를 관리하려고 하는 과거의 모든 시스템들은 웹으로부터 얻어진 정보를 검색, 필터, 표현할 때, 불안정한 표시를 사용자 또는 다른 에이전트들에게 알려줄 필요가 있다. 이를 기반으로 웹에서 다른 사이트로부터 정보를 검색할 수 있는 전문 에이전트들을 구축한 시스템들을 개발할 수 있다. 일반적으로 웹에 저장된 정보를 다루는 멀티에이전트 시스템은 에이전트들의 두 가지 서로 다른 형태로 구성되어 있다. 하나는 웹에서 특별한 사이트로부터 정보를 검색하고 여과하고 저장하도록 전문화된 에이전트이다. 또 다른 에이전트는 문제를 해결하기 위해 전자의 에이전트가 검색한 지식을 재사용하거나 플래닝 혹은 학습과 같은 인공지능에 전통적 기법들을 이용해서 추론할 수 있다.

본 논문에서는 퍼지 이론을 기반으로 하여 에이전트들의 행동을 적절하게 비교할 수 있고, 평가 및 분류할 수 있다. 퍼지 지식 기반의 프로토타입은 서로 다른 에이전트들의 행동 사이에 거리를 평가하기 위해 제안하였다.

2. 퍼지 거리

퍼지 대상간의 거리는 실수값을 주는 경우와 퍼지수를 주는 경우 두 가지 방법이 있다.

그리고, 각각에 대하여 몇 가지의 방법들을 살펴보면, 먼저 중심간 유클리드 거리 측정 방법으로 퍼지 대상 O_n, $O_{n'}$의 중심간 유클리드 거리는 식 (1)과 같이 나타낸다[5].

$$d(O_n, O_{n'}) = | a_n - a_{n'} | \quad (1)$$

또한 퍼지수로 주어지는 퍼지 거리 측정 방법으로 퍼지 대상 O_n, $O_{n'}$간의 퍼지 거리 $\tilde{d}(O_n, O_{n'})$ 를 다음의 소속 함수로 정의하는 것이 가능하다.

$$\mu_{d(O_n, O_{n'})}(\delta) = \sup_{\delta = x-y} \min\{ x-y \quad x \epsilon [\mu_{O_n}(x), \mu_{O_{n'}}(y)]\} \quad (2)$$

일반적으로 이 거리는 퍼지 거리로써 알려져 있다. 또한 구간 퍼지 수에서 주어지는 퍼지 거리 측정으로는 두 가지 방법이 있다.

① $\alpha-cut$ 집합간의 구간 퍼지 거리

퍼지 대상 O_n, $O_{n'}$ 집합 간의 최대와 최소 거리는 식 (3), (4)와 같다.

$$d^{\alpha}_{sup(nn')} = sup\{ | a_n - a_{n'} | \ | \ x \in [O_n]_{\alpha}, y \in [O_{n'}]_{\alpha}\} \quad (3)$$

$$d^{\alpha}_{inf(nn')} = inf\{ | a_n - a_{n'} | \ | \ x \in [O_n]_{\alpha}, y \in [O_{n'}]_{\alpha}\} \quad (4)$$

$\alpha-cut$을 이용하여 구간 퍼지수로 정의한다.

$$\tilde{d}^{\alpha}(O_n, O_{n'}) = [O^{\alpha}_{inf(nn')}, O^{\alpha}_{sup(nn')}] \quad (5)$$

이 거리를 구하기 위해서는 복잡한 계산이 요구된다.

② $\alpha-cut$ 데이터 벡터 집합간의 구간 퍼지 거리

$$d^{\alpha}_{max(nn')} = max\{ | a_n - a_{n'} | \ | \ x \in [D_n]_{\alpha}, y \in [D_{n'}]_{\alpha}\} \quad (6)$$

$$d^{\alpha}_{min(nn')} = min\{ | a_n - a_{n'} | \ | \ x \in [D_n]_{\alpha}, y \in [D_{n'}]_{\alpha}\} \quad (7)$$

$\alpha-cut$을 이용해서 구간 퍼지수로 정의한다.

$$\tilde{d}^{\alpha}(D_n, D_{n'}) = [d^{\alpha}_{min(nn')}, d^{\alpha}_{max(nn')}] \quad (8)$$

3. 웹 에이전트의 정보 분류에 대한 퍼지 평가 시스템

3.1 퍼지 평가 시스템

정보 분류 방법으로 두 벡터들 사이의 거리를 측정해야하는 방법이 자주 사용되고 벡터들은 정보를 나타낸다. 하나의 벡터가 다른 벡터와의 관계 정도를 나타내기 위해 거리를 측정하는 방법들이 사용되어왔다. 그러나 문제의 원 정보가 부정확할 경우, 유클리디안 거리와 같은 일상적 거리를 사용할 경우 정확한 결과를 얻기 어렵다. 왜냐하면 전통적인 거리 측정 방법은 데이터에 관하여 알려진 모든 의미론적 정보를 고려하지 않기 때문이다. 그러므로 퍼지 거리와 같은 비전통적인 거리 측정 방법을 사용하는 것은 전통적 기술들로부터 얻어진 결과들을 개선시킬 뿐만 아니라, 웹 에이전트들의 효율성을 개선시킨다.

퍼지 이론을 적용한 소프트웨어 에이전트를 개발하기 위해서는 먼저 퍼지 알고리즘 위에서 적용될 수 있도록 구성하는 것이 필요하다. 그런 다음, 검색되고 여과된 정보 데이터를 특성화하는 벡터로 구성한다. 벡터는 태스크에 관련된 효용성에 관한 일반적인 특성으로 그룹화 된 특성들의 집합을 나타낸다. 정보는 웹에서 특성화 한 값의 집합으로 $(F_1, F_2, ..., F_n)$로 표현될 수 있다. 각 $f_i \in F_i$에서 f_i의 가능한 수치 값의 치역이 F_i가 된다.

3.2 평가 결과와 퍼지 수

에이전트가 평가 결과값을 고려하여 정보를 선택하고자 할 때, 전반적으로 우수한 결과의 정보일지라도 중요한 특정 항목의 결과가 사용자를 만족하지 못할 경우가 발생한다. 이와 같은 경우는 선택의 결과가 다르게 나타날 것이다. 일반적인 경우 전반적인 평가 결과값 만으로 정보를 분류하였다. 하지만 특정 항목의 최소 요건을 충족하면서 가장 유사한 정보를 분류하는 방법이 중요하다.

퍼지집합에 포함된 평가 결과값들 중에서 일정한 가능성 이상 포함된 결과값들만 구성된 보통 집합을 만들 수 있다. 또한 사용자에게 적합한 각 항목에 대한 한계값을 다음과 같이 나타낸다.

$$C_n = x \in U \ \tilde{f}_i(x) \geq \alpha \quad (9)$$

여기서 C_n는 사용자의 한계 정도를 나타낸다. 퍼지 평가 결과값에 대한 $\alpha-cut$을 나타내면

$$\tilde{F}_i = \begin{cases} \tilde{f}_i & \tilde{f}_i \geq \alpha \\ 0 & \tilde{f}_i < \alpha \end{cases} \quad (10)$$

각 항목에 대한 $\alpha-cut$을 적용한 퍼지 평가

결과값을 나타내면 다음과 같이 나타낸다.

$$U_{\tilde{F}_i} = \{\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_n\} \quad (11)$$

여기서 $\tilde{F}_1$, $\tilde{F}_2$, ..., $\tilde{F}_n$은 $\alpha-cut$을 적용하여 사용자의 한계값을 나타낸 결과이다.

기준 값에 대한 최소상계와 최대하계를 다음과 같이 나타내고, 각 기준 값을 0과 1사이의 중간 값으로 설정하여, 최소상계와 최대하계 사이를 0과 1 사이의 퍼지 구간 값으로 나타냈다. 또한 최소상계 이상의 값이면 경계 값을 벗어난 값으로 간주하여 1로 설정하였으며, 또한 최대하계 이하의 값이면 0으로 설정하였다.

$$INFs_i = \{b \mid \tilde{f}_i \geq b\} \quad (12)$$

$$I_{Lsize} = \frac{\tilde{f}_i - inff_i}{p} \quad (13)$$

$$I_{Usize} = \frac{supf_i - \tilde{f}_i}{p} \quad (14)$$

여기서, I_{Lsize}는 기준 값 아래의 구간 크기이고 I_{Usize}는 기준 값 위의 구간 크기이며, p는 구간의 수를 나타낸다. 따라서 각 소속함수는 다음과 같이 정의된다.

$$I_i = \{0.0/I_0, \ldots, 0.5/I_5, \ldots, 1.0/I_{10}\} \quad (15)$$

각 항목의 평가 결과 값을 식 (15)에 대응시켜 퍼지수로 나타낸다. 또한 가중치 $\tilde{W}_i$는 전문가가 평가 항목의 중요도에 따라 적절한 가중치를 부여한다.

위의 결과 값 $\tilde{F}_i$와 각 항목들에 대한 가중치를 퍼지수로 나타내면 표 1과 같다.

표 1. 각 항목에 대한 가중치와 퍼지수

no.	가중치	정보A	정보B	…	정보N
1	$\tilde{W}_1$	$\tilde{F}_{1A}$	$\tilde{F}_{1B}$	…	$\tilde{F}_{1N}$
2	$\tilde{W}_2$	$\tilde{F}_{2A}$	$\tilde{F}_{2B}$	…	$\tilde{F}_{2N}$
3	$\tilde{W}_3$	$\tilde{F}_{3A}$	$\tilde{F}_{3B}$	…	$\tilde{F}_{3B}$
⋮	⋮	⋮	⋮	…	⋮
n	$\tilde{W}_n$	$\tilde{F}_{nA}$	$\tilde{F}_{nB}$	…	$\tilde{F}_{nN}$

각 정보의 전체 퍼지수는 $S(\cdot)$로 표시된다.

$$S(A) = \tilde{W}_1 \otimes \tilde{F}_{1A} \oplus \cdots \oplus \tilde{W}_n \otimes \tilde{F}_{nA} \quad (16)$$

$$S(B) = \tilde{W}_1 \otimes \tilde{F}_{1B} \oplus \cdots \oplus \tilde{W}_n \otimes \tilde{F}_{nB} \quad (17)$$

$$\vdots$$

$$S(N) = \tilde{W}_1 \otimes \tilde{F}_{1N} \oplus \cdots \oplus \tilde{W}_n \otimes \tilde{F}_{nN} \quad (18)$$

여기서, $S(A), S(B), \ldots, S(N)$는 삼각퍼지수를 나타낸다.

적합성 정도는 $D_\lambda(A)$, $D_\lambda(B)$ …, $D_\lambda(N)$으로 표시하며, λ는 정보의 적합성 정도를 나타낸다. λ값이 적은 것은 적합성의 정도가 높음을 나타낸다.

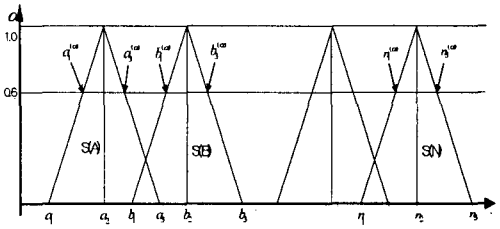

그림 1. $\alpha-cut$을 실행한 $S(\cdot)$

$$a_1^{(\alpha)} = \alpha(a_2 - a_1) + a_1 \quad (19)$$

$$a_2^{(\alpha)} = -\alpha(a_3 - a_2) + a_3 \quad (20)$$

$$b_1^{(\alpha)} = \alpha(b_2 - b_1) + b_1 \quad (21)$$

$$b_2^{(\alpha)} = -\alpha(b_3 - b_2) + b_3 \quad (22)$$

$$\vdots$$

$$n_1^{(\alpha)} = \alpha(n_2 - n_1) + n_1 \quad (23)$$

$$n_2^{(\alpha)} = -\alpha(n_3 - n_2) + n_3 \quad (24)$$

여기서 크리스프 한 값을 구하기 위하여 $\alpha-cut$을 적용하면, $S(A)$, $S(B)$, …, $S(N)$의 $\alpha-cut$은 $[a_1^{(\alpha)}, a_3^{(\alpha)}]$, $[b_1^{(\alpha)}, b_3^{(\alpha)}]$, …, $[n_1^{(\alpha)}, n_3^{(\alpha)}]$ $\alpha \in [0, 1]$로 표시할 수 있으며, 그림 1과 같다.

각 평가 항목별로 평가한 결과가 사용자의 주관적인 가중치에 의해서 특정 정보에 대하여 비관적으로 평가할 수도 있고 낙관적으로 평가할 수도 있다.

또한 λ의 값이 작은 것은 "낙관"의 정도가 높음을 나타내고, λ의 값이 큰 것은 "낙관"의 정도가 낮음을 나타내며, 각 정보간의 적합성 퍼지수 값은 식 (25)~(27)과 같이 정의할 수 있다.

$$D_\alpha^\lambda(A) = \lambda a_1^{(\alpha)} + (1-\lambda)a_3^{(\alpha)} = P_A \quad (25)$$

$$D_\alpha^\lambda(B) = \lambda b_1^{(\alpha)} + (1-\lambda)b_3^{(\alpha)} = P_B \quad (26)$$

$$\vdots$$

$$D_\alpha^\lambda(N) = \lambda n_1^{(\alpha)} + (1-\lambda)n_3^{(\alpha)} = P_N \quad (27)$$

각 정보의 평가항목별로 평가된 점수가 각 항목별 A, B 정보 전체의 평가 점수에 대하여 차지하는 비율이 높은 정보가 상대적으로 적합한 정도를 나타내고 있으며 식 (28)~(30)과 같이 나타낸다.

$$N_\lambda^\alpha(A) = \frac{P_A}{P_A + P_B + \cdots + P_N} \quad (28)$$

$$N_{\lambda}^{\alpha}(B) = \frac{P_B}{P_A + P_B + \cdots + P_N} \quad (29)$$

$$\vdots$$

$$N_{\lambda}^{\alpha}(N) = \frac{P_N}{P_A + P_B + \cdots + P_N} \quad (30)$$

$N_{\lambda}^{\alpha}(A)$, $N_{\lambda}^{\alpha}(B)$, ···, $N_{\lambda}^{\alpha}(N)$의 값들은 모든 정보들이 평가받은 점수에 대한 각각의 점수 비율을 나타내는 것이며, 가장 큰 값이 사용자에게 가장 적합한 정보를 나타낸다. 여기서, $N_{\lambda}^{\alpha}(A)$, $N_{\lambda}^{\alpha}(B)$, $N_{\lambda}^{\alpha}(N) \in [0, 1]$이 된다.

4. 시스템 평가 및 결과 분석

평가된 된 정보들 중 사용자에게 가장 적합한 정보를 선정하기 위해 각각의 항목에 따라 측정된 결과에 대한 각 정보의 퍼지수와 각 항목에 따른 가중치를 설정한다. 여기서 평가 된 결과값과 가중치는 삼각퍼지수로 표현된 것이다. 퍼지수와 에이전트에 의해 설정된 각 항목의 가중치를 설정하였다. 또한 측정된 결과값을 바탕으로 $\alpha-cut$의 값을 0.7로 설정하여 적용한 각 정보의 값을 사용하였다.

표 2. $\alpha-cut$ 에 따른 결과의 적합성 정도

	$\alpha-cut$을 적용한 적합성 정도	$\alpha-cut$을 적용하지 않은 적합성 정도
정보 A	29.73	29.73
정보 B	28.06	30.46

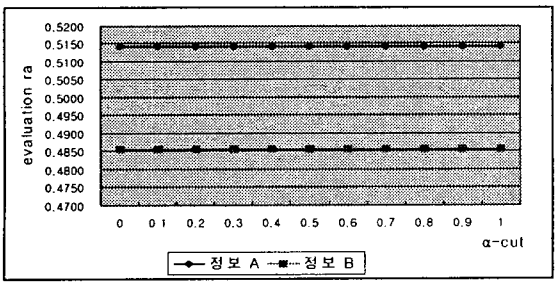

그림 2. $\alpha-cut$을 적용한 정보간의 평가비율

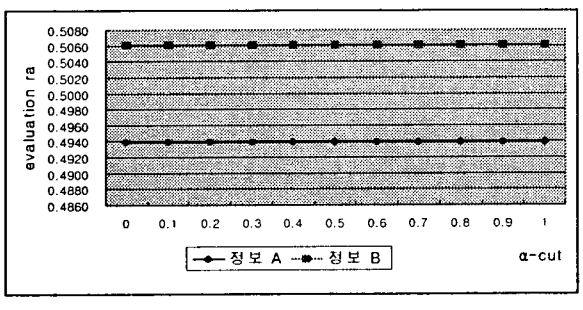

그림 3. $\alpha-cut$을 적용하지 않은 정보간의 평가비율

$\alpha-cut$을 적용한 경우의 적합성과 평가비율은 $D_{\alpha}^{\lambda}(A) = 29.73$, $N_{\lambda}^{\alpha}(A) = 0.5144$, $D_{\alpha}^{\lambda}(B) = 28.06$, $N_{\lambda}^{\alpha}(B) = 0.4855$이다. 반면 $\alpha-cut$을 적용하지 않은 경우는 $D_{\alpha}^{\lambda}(A) = 29.73$, $N_{\lambda}^{\alpha}(A) = 0.4939$, $D_{\alpha}^{\lambda}(B) = 30.46$, $N_{\lambda}^{\alpha}(B) = 0.5060$이라는 사실을 알 수 있다. 여기서 $\alpha-cut$을 0.7로 적용한 결과 정보 A가 더 적합하게 평가되었지만, $\alpha-cut$을 적용하지 않았을 경우는 정보 B가 더 적합하게 평가되었다. 이 결과는 일반적으로는 특정 정보가 적합하게 나왔더라도 특정 사용자의 환경에는 적용되지 않는다는 사실을 알 수 있다.

5. 결론

이 시스템은 주로 정보간의 비교 평가에 사용된다. 사용자로부터 양질의 웹 서비스 요구가 증대되고 있는 웹 시스템은 다양한 정보로 구성되어 있으며, 정보들을 최적으로 분류하기 위해서는 각 정보간의 적합도를 평가하게 된다. 이를 위하여 퍼지 이론을 이용하게 되며, 비교적 객관적이고 사용자에게 보다 정확한 평가 자료를 제시할 수 있다. 또한 에이전트들이 정보간의 분류를 위해 항목간의 가중치를 조정할 수 있으며, 검색자의 최소 요구사항을 고려하여 분류할 수 있다.

5. 참고문헌

[1] Camacho D., Hernández C., Molina J. M., INFORMATION CLASSIFICATION USING FUZZY KNOWLEDGE BASED AGENT, Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference, 2001.

[2] Camacho D., Hernández C., Molina J. M., Information Classification in Web Agents using Fuzzy Knowledge for Distance Evaluation. WSES International Conference on : Fuzzy Sets & Fuzzy Systems(FSFS -01), Canary Islands Spain, Feb. 2001.

[3] Etzioni O. Moving Up the Information Food Chain, In AI Magazine, Vol. 18, (2), pp.11-18, summer 1997.

[4] Selberg E., Etzioni O. The MetaCrawler Architecture for Resource Aggregation IEEE Expert, vol. 12, no. 1, pp8-14, 1997.

[5] 유동선, 이교원, 기초 퍼지이론, 교우사, 1998.

[6] Camacho D., Molina J.M., Borrajo D., Aler R.. MAPWEB:Cooperation between Planning Agents and Web Agents. Information & Security: An International Journal. Volume 7. 2001.