논문 2009-4-19

퍼지를 이용한 도메인 검색용어 중요성의 표시

An Expresson of Domain Searching Term Weight using Fuzzy

진현수*, 홍유식**

Hyun-Soo Jin, You-Sik Hong

요 약 최근의 여러분야에서 검색되어지고 있는 인터넷 도메인 용어의 전문성의 표시화는 온톨리지를 통한 지식의 축적의 목표가 되고 있다. 도메인 용어의 중요성을 표시화 한다면 기계가 온톨리지를 이용하여 정보의 관리및 해석을 스스로 하는것이 가능할 것으로 본다. 본 논문에서는 온톨로지의 중요성 (weight)을 구성하는 속성을 확장된 퍼지를 사용하여 기존 웹문서의 구조정보로부터 추출하는 알고리즘을 제안하였다. 특히 속성정보로 구성된 도메인 지식을 표시화함으로써 속성추출 알고리즘을 개선하고, 추출결과의 품질을 향상시킨다. 5만문서를 대상으로 제안된 알고리즘을 적용한 결과 약 94%의 신뢰도의 속성정보를 추출할수 있었다.

Abstract The leveling of technical internet domain term with its aim to accumulate knowledge that machine can comprehend, which has been used widely in recent years. If stratify domain term weight, we believe that machine can manage and analyze in formation on its own using the ontology. In this paper, we propose an algorithm that allows us to extract properties of ontology weight from structured information already existing in web documents. In particular by stratification of the domain knowledge that is composed of property information, we were able to make the algorithm better,and improve the quality of extraction results.In our experiments with 50 thousands targeted documents, we were able to extract property information with 94% confidence.

Key Words : 온톨로지, 도메인지식, 속성, 표시화

I. 서론

최근의 웹사회는 자기 스스로 정보를 생산해내는 블로그 문서뿐만 아니라 정보회사에서 저장된 메시지를 송출하는 싱크로 웹문서등 정보의 홍수속에서 살아도 무방할정도의 문서의 홍수속에서 살아가고 있다. 지난 시간에는 원하는 정보를 어떻게 찾아내는가에 달려 있었는데 현재의 사회에서는 어떻게하면 사용자가 원하지 않는 정보를 걸러내는가에 달려 있다. 정보를 찾아내는 사회를 탐색사회 (searching)라 한다면 정보를 걸러내는 사회를 (filtering)검색사회라 할수 있다. 실상은 현재의 사회가 정보량이 증가하고 있는 추세에 정보검색및 정보 탐색 기술이 정보량의 폭주에 따라가지 못하고 있는 실정이다. 이러한 상황은 현재의 웹문서의 기본 골격인 HTML구조 때문에 그렇다. HTML구조의 웹환경은 정보의 가시적인 표현만 기술할 뿐으로 정보의 의미를 구조화 혹은 체계화하는 기능이 없기 때문에 웹문서로 표현된 정보의 해석이나 활용은 순전히 사람에 의해 수행될수 밖에 없다. 이런 어려운 점을 해결하기 위해서 WWW consottium(W3C)은 기계가 이해할 수 있는 의미구조를 가진 시멘틱 웹을 제안하고 있다. 선택된 결정들 중에서 최적의 상위어 용어를 결정하여 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 더 정확하게 용어를 결정한다[1][2].

*정회원, 백석대학교 정보통신학부
**정회원, 상지대학교 컴퓨터 정보공학부
접수일자 2009.7.27, 수정완료 2009.8.5

II. 관련연구

2.1 온톨로지

Gruber가 정의한 “온톨로지는 명확한 관심사에 대한 공유된 개념의 정형화된 규정”이라는 정의가 사용되고 있다 [3] . 또한 온톨로지의 특징은 개념간의 관계와 용어간의 관계를 분리하여 해당 관계영역을 파악하고 관계영역내에서 일관성있고 명확하게 개념을 정의하고 새로운 지식을 추론하고 지능적인 정보처리에 적용될수 있다. 그림 1 은 온톨로지와 유사한 시소러스와 온톨로지의 차이점을 나타낸다.

시소러스는 웹의 기본 근간인 온톨로지를 구축하기 위해 사용되고 있다.

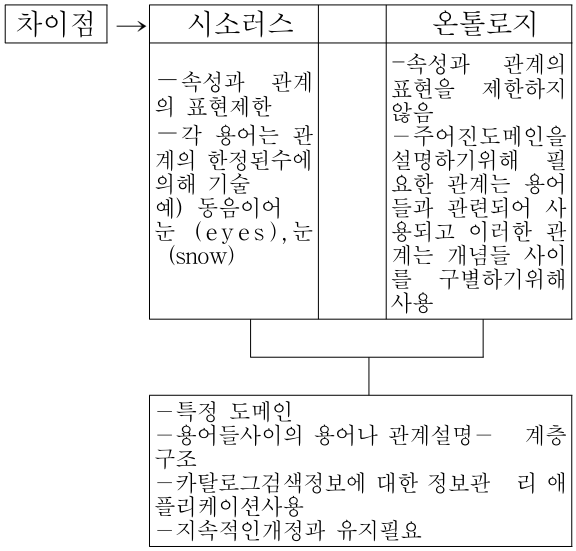

그림 1. 온톨로지, 시소러스의 관계
Fig. 1. Sysurus and Ontology

할수 있고 여러전문가들의 작업결과가 일관성을 가지게 된다. 그림 2 는 온톨로지 학습단계를 나타낸다.

학습단계에서 가장 기본 단계인 “Term"단계에서는 온톨로지 구축을 위한 대상용어를 추출하고 선정 하는데 관계 용어인 ”Disease""illness" "hospital" "operation(수술) “사용하여 대상을 추출하고 ”synonym"단계에서는 선정한 용어들 사이의 동의어를 그룹핑하고 따라서 illness와 disease가 동의어로써 추출되어지고 “concepts" 단계에서는 그룹핑된 용어들은 개념으로 표현하고 따라서 Disease를 정형화한 표시도구인I(Illness) E(END),L(Location)철자로서 표현한다. ”Hierachies"단계에서는 개념들 사이의 상하위 관계를 설정하고 개념들 계층관계에서 서로 상관관계를 유지해주는 Is_A(Doctor person)으로 연결고리가 되어준다.

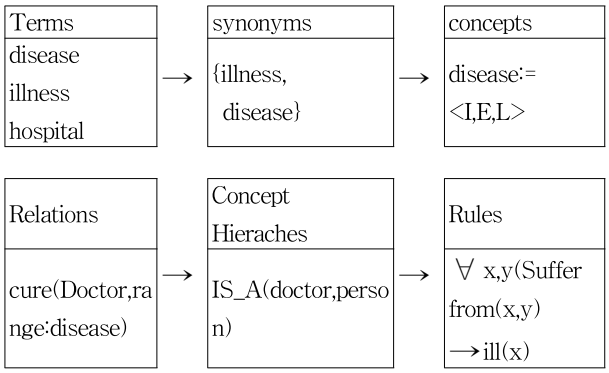

그림 2. 온톨로지 학습단계
Fig. 2. Ontology learning step

Relations단계에서는 개념간 상속관계를 표현하기 때문에 지능형 시스템에서 상하위어 관계탐색을 통한 추론 기능을 제공한다. Rules단계에서는용어들 사이의 계층관계를설정하여 조직화 시킨 것으로 계층구조에 포함된 모든 용어는 한개 이상의 용어와 계층관계를 가진다.

III. 도메인 온톨로지 용어의 전문성과 의미 유사도

본 논문에서는 “an ontology is an explicit formal specification of shared conceptulization of a domain of interest"라고 정의하기 때문에 다음과 같이 확정지을수 있다. ”어휘들에 대해서 일정 영역의 개념적 예들을 한곳으로 집합시킨 하나의 독립된 집합체“이고 해당 도메인의 정의는 다음과 같다

3.1 도메인 용어의 정의

데이터는 임베디드 시스템에서 실험과 연구에 대한 Imbedi.com과 kwork.com과 과 같은 experiment website 상에 망라된 임베디드 고장괴 시스템 연결상태를 고려한 온톨로지에서 도메인 용어의 전문성과 의미 유사도를 나타낸다. ‘experiment success of web servant'의 해당 도메인에 대하여 ‘embedid system' 'computer memory' 'power supply'와 같은 embedid system에 관한 관련된 domain용어가 있다.

이 해당도메인을 이용하여 동력문구 패턴 응용학습주에 전체-부분관계를 구글 (power fault of embedid system,약 3,990,000문서) 에서 8개의 용어 (hard spare,memory stick,computer software,power builder,OBbear,disk failure,virus check,printer error)의 부분을 추출했으며 각 개념은 하위 개념들의 집합을 가진다. 예를 들면 computer power는 printer error,disk builder,OB bear,disk failure,virus check,printer error"와 같은 퍼지용어를 표1에 나타낸다[4][5].

표 1. 전문적으로 제시된 도메인 용어
Table 1. Proposal Domain term of technician

term	member-shipfunction	explanation of term
hard spare	(20,40,60,80)	free disk space free memory space
memory space	(100,30,50,60)	virus information of
computer S/W	(3,2,7,9)	trend micro status of computer
power builder	(20,40,60,80)	hardware power remain
OB bear	(20,40,50,70)	operative condition of
disk failure	(20,40,60,70)	the program server operative
virus check	(20,50,70,90)	condition
printer error	(80,70,100,90)	cpu loading

파견자 전문가 집단 소속 2는"RLC Element,Network Element,memory check,Hardware checker,domain hacker"를 제안하여 표2에 나타내었다

표 2. 파견자 전문 집단소속2의 제안
Table 2. proposal of expert group 2

term	membership function	explanation of term
RLC Element	(60,70,100)	Network loading
Network	(70,80,90)	cpu loading
Element		Free memory
memory	(100,300,500)	space
check		Rank of
Hardware	(1,3,5)	dangerous for
checker		server
domain	(1,3,7)	status
hacker		of computer hardware

두 그룹이 제시한 도메인 용어중'memory stick'과 'memory check'의 용어는 'memory stick'용어로 의견을 통합하고'domain hacker'와 'hardware checker'를 채택하여 표3에 나타낸다. 통합된 제안을 바탕으로 서버고장에 대한 퍼지용어들의 온톨로지 관계를 그림3에 나타낸다.

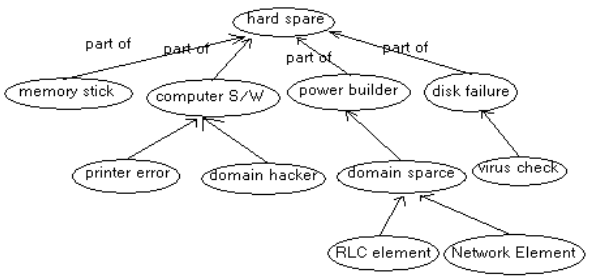

그림 3. 온톨로지에서 퍼지용어간의 관계
Fig.3 Relation of fuzzy terms in ontology

3.2 도메인 용어 전문성

도메인 용어의 퍼지 전문성은 퍼지 용어가 포함하는 전문적인 정보의 량을 정량적으로 표현한 것이다. 타기준 의사결정 문제인 퍼지용어의 전문성을 나타내기 위하여 'disk failure of printer server'도메인을 이용한다. 이것은 의사결정 기준의 집합에 의한 대안 집합의 평가로 사용된다[6][7].

표 3. 쌍비교 행렬의 구조
Table 3. structure of pair-wise comparison

	w_1	w_2	• • •	w_i	w_n	GM	μ_i
ω_1	1					GM_1	μ_1
w_2		1				GM_2	μ_2
• • •			1			GM_3	μ_3
w_i				1		GM_i	μ_i
w_n					1	GM_n	μ_n

$GM_i=[\pi_j^n v_{ij}]^{\frac{1}{n}}$ 단 i=1,n,w_{ii}=1

$$w_i=\frac{GM_i}{(GM_1+GM_2+\bullet\ \bullet\ GM_n)} \quad (1)$$

표5는 'disk failure of printer server'도메인에서 채택된 도메인 용어 10개를 쌍비교 행렬로 표현하여 비교하고 각 도메인 용어의 기하평균과 가중치를 계산한다.

퍼지 용어 C_1은hard spare C_2는 memory stick C_3는 computer error C_4는 power builderC_5는 OB bear, C_6는 disk failure, C_7는 virus check, C_8는 printer error C_9는Networkelement,C_{10}은 RLC element를 의미한다.표 3에 나타난 쌍비교 행렬의 값은 1부터 5까지 분류된다. 다른 퍼지용어들을 비교함으로써C_1, C_2, C_3, C_4,C_5,C_6,C_7,C_8,C_9,

C_{10}의가중치는각각 0.728,0.0179,0.1059,

0.1782,0.0459,0.0782,0.0782,0.0875,

0.0162,0.1973,0.7182으로 계산된다. 퍼지용어들의 상대적 중요성은 C_5, $C_7 > C_4 > C_2$, $C_3 > C_8$, $C_{10} > C_1 > C_6$, C_9 순이다[8].

표 4. 쌍비교 행렬
Table 4. pair-wise comparison matrix

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	GM	wi
C_1	1	1/3	3	1/4	1/3	1/4	1/5	1/3	1/3	1/5	0.21	0.07
C_2	1/2	1	1/4	1/5	1/4	1/3	1/2	1/3	1/2	1/5	0.23	0.08
C_3	1/3	1/2	1	1/4	1/5	1/6	1/3	1/2	1/2	1/4	0.24	0.03
C_4	1/2	1/3	1/2	1	1/4	1/5	1/3	1/2	1/3	1/2	0.25	0.04
C_5	1/4	1/2	1/3	1/4	1	1/4	1/4	1/3	1/2	1/5	0.24	0.05
C_6	1/3	1/2	1	1/4	1/5	1	1/3	1/2	1/2	1/4	0.24	0.03
C_7	1/4	1/2	1/3	1/4	1/4	1/3	1	1/3	1/2	1/5	0.24	0.05
C_8	1	1/3	3	1/4	1/3	1/4	1/2	1	1/3	1/5	0.21	0.07
C_9	1/2	1	1/4	1/5	1/4	1/3	1/2	1/3	1	1/5	0.23	0.08
C_{10}	1/3	1/2	1	1/4	1/5	1/6	1/3	1/2	1/2	1	0.24	0.03

위의 결과를 토대로 퍼지용어의 계층구간을 표5와 그림4에 나타낸다

표 5. 퍼지용어의 계층구간
Table 5. level interval of fuzzy terms

퍼지용어	구간
C_6,C_9	0.05~0.07
C_1	0.07~0.09
C_8,C_{10}	0.09~0.1
C_2,C_3	0.1~0.21
C_4	0.12~0.18
C_5,C_7	0.16~

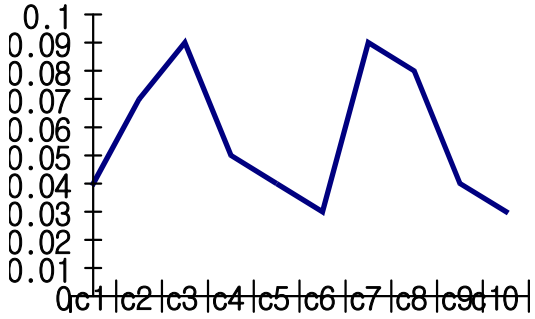

그림 4. 퍼지용어 중요도 구간
Fig.4 Level Interval graph fuzzy term

3.3 퍼지정보서 퍼지용어 의미유사도

실제 웹문서에서 두 정보의 퍼지정도가 유사함을 검색을 통해 그림5에 얻어낼 수가 있었고 관계정립을 위해 유사정 관계치를 구하였다.

그림 5. 실시간 검색 용어관련도메인
Fig.5 Realtime relative domain term

3.3.1 Min연산자 의미유사도

'disk failure of printer server' 도메인에서 축소된 AHP를 이용해 퍼지 용어전문성을 구한다음 퍼지집합의 Min연산자를 이용하여 퍼지용어 의미유사도를 비교한다. 퍼지집합의 성질은 다음과 같이 정의한다.

$$f_{A \cup B}=\max[f_{A(x)}, f_B(x)]$$
$$f_{A \cap B}=\min[f_{A(x)}, f_B(x)] \quad (2)$$
$$f_{A'}(x)=1-f_A(x)$$

퍼지용어의 관계행렬은 퍼지용어의 문맥정보로서 두 퍼지용어에 대한 소속도를 나타낸다. C_1(HD space)와

C_2(memory space)의 문맥정보벡터는 각각 C_1 =(0.4,0.7,0.8,0.9,0.7,0.6, 0.3,0.2,0.8,0.7), C_2=(0.6,0.7,0.4,0.5,0.3,0.7,0.4,0.5,0.8,0.9) 이고 퍼지집합의 Min연산자를 사용한 두벡터사이의 유사도 구간은 표7에 나타낸다.

표 7 의 결과에 의하면 퍼지용어의 의미 유사도 계층은 C_5>C_4,C_9>C_1,C_7>C_2,C_8>C_3,C_{10}>C_6 순으로 구성된다. 그러나 퍼지집합의 사용은 Min,Max가중치를 부여해야만 하고 검색된 문헌의 순위 부여능력이 모든 검색어에 민감하지 못하는 단점을 가지고 있다.

$$(C_1,C_2)=\frac{2/\sum(x_i-y_i)}{\sum x_i^2+\sum Y_i^2} \qquad (3)$$

여기서 X= $\sum_{i=1}^{n} x_i$와 Y= $\sum_{i=1}^{n} Y_i$는 각가 두 퍼지용어 C_1과C_2의 특징에 대하여 가중치를 나타내는 벡터이다. 표 9 는 다이스 계수를 사용한 퍼지용어 의미 유사도의 결과를 구간별로 나타낸다.

IV. 결론

본논문은 문서에서 추출된 퍼지용어 정보를 바탕으로 한 온톨로지 구조를 카테고리화 하여 퍼지용어의 전문성을 이용하여 주어진 퍼지용어의 상위어 후보를 레벨화한 후 퍼지용어 의미 유사도를 계산하여 선택된 후보들 중에서 최적의 상위어 후보를 결정한다. 즉 퍼지용어간의 전문성을 레벨화하기 위한 확장된 AHP방법은 가중치가 부여된 다수 평가자의 평가치를 통합한후 퍼지 용어의 쌍비교를 통해서 가중치나 상대적 중요성을 결정한다. 그리고 퍼지용어 의미 유사도는 퍼지집합의 Min연산자와 다이스 계수 Min+ 다이스계수를 비교한다. 이러한 방법들은 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 보다 더 정확하게 문서를 분류할수 있고 자연적 처리등에 많이 활용될수 있을 것이다. 향후 퍼지 용어의 의미유사도를 다치형태 확장의 연구가 요구된다.

참고문헌

[1] M.J.Matric,"Behavior_based control:Examples from navagation, learning,and group behavior, "Journal of Experimental and Theoretical Artificial intelligence, vol.9,no.2,pp.323-336, 1997

[2] 진현수, “퍼지제어를 이용한 온톨로지 행동학습과 평형 알고리즘에 관한 고찰”, 퍼지및 지능시스템학회, 춘계컨퍼런스, vol3.no2,pp.428-430,1998

[3] Satty,R,W., "The Analytic Hierachy Process-what it is and how it is used,"Mathematical Modeling, pp.161-176,1987.9.

[4] H.J.Zimmermann and P.Zysno,"Decision and Evaluation by Hierarchical Aggregation of Information", Fuzzy Sets and Systems Vol.10, pp31-36, 1983

[5] 진현수, “퍼지 지능기법을 이용한 교차로 시스템의 제어에 관한 연구”, 퍼지및 지능시스템 학회 춘계학술대회, vol4,no 3,pp321-316,1998

[6] Toshio Fukada,"Multi-Sensor tegration System with Fuzzy Inference and Neural Network", IEEE Fuzzy Int.conf.vol13,no22,pp234-239, 1992

7] Gilles Mauris, "The aggregation of information by examples via fuzzy sensors",IEEE third Int.Conf. system, Orlando, USA, pp.67-72,june

[8] E.Benoit,L.Foulloy et.al, "Fuzzy sensor for the perception of Colour", Submitted to the Third IEEE Int.Conf. on Fuzzy,USA,pp.28-45,june 1994

저자 소개

진 현 수 (정회원)

- 1986년 서울시립대학교 전자과 학사 졸업.
- 1991년 서울시립대학교 전자과 석사 졸업.
- 2001년 서울시립대학교 전자과 박사 학위
- 2008년 현재 백석대학교 정보통신학부 교수.

<주관심분야 : 인터넷,인공지능, 웹메니지먼트>

홍 유 식 (종신회원)

- 제 9 권 3 호 참조
- 현 상지대학교 컴퓨터정보공학부 교수

<주관심분야 : 퍼지시스템, 전문가시스템, 신경망, 교통제어>