한국지능시스템학회 논문지 2009, Vol. 19, No. 2, pp. 212-217

정보시스템에서 퍼지용어의 확장된 AHP를 사용한 레벨화와 유사성측정

A Leveling and Similarity Measure using Extended AHP of Fuzzy Term in Information System

류경현 · 정환묵

Kyung-Hyun Ryu and Hwan-Mook Chung

대구가톨릭대학교 컴퓨터정보통신공학부

요 약

특정 분야의 용어를 표현하는 전문용어 사이의 계층관계를 학습하는 방법은 규칙기반학습방법, 통계기반학습방법 등이 있다. 본 논문에서는 문서에서 추출된 퍼지용어 정보를 바탕으로 한 온톨로지 구조를 카테고리화하여 퍼지용어의 전문성을 이용하여 주어진 퍼지용어의 상위어 후보를 레벨화한 후 퍼지용어 의미유사도를 계산하여 선택된 후보들 중에서 최적의 상위어후보를 결정한다. 즉, 퍼지용어의 전문성을 레벨화하기 위한 확장된 AHP방법은 퍼지용어사이의 비교를 통해 가중치나 상대적 중요성을 결정한 후 퍼지집합의 Min연산자와 다이스계수, Min+다이스계수방법들을 비교한다. 이 방법들은 퍼지용어 의미유사도에 따라 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 보다 더 정확하게 문서를 분류할 수 있고 자연어처리 등 많은 분야에 활용될 수 있을 것이다.

키워드 : 퍼지정보시스템, 온톨로지, 퍼지용어, 유사도, 계층분석과정

Abstract

There are rule-based learning method and statistic based learning method and so on which represent learning method for hierarchy relation between domain term. In this paper, we propose to leveling and similarity measure using the extended AHP of fuzzy term in Information system. In the proposed method, we extract fuzzy term in document and categorize ontology structure about it and level priority of fuzzy term using the extended AHP for specificity of fuzzy term. the extended AHP integrates multiple decision-maker for weighted value and relative importance of fuzzy term. and compute semantic similarity of fuzzy term using min operation of fuzzy set, dice's coefficient and Min+dice's coefficient method. and determine final alternative fuzzy term. after that compare with three similarity measure. we can see the fact that the proposed method is more definite than classification performance of the conventional methods and will apply in Natural language processing field.

Key Words : Fuzzy Information System, Ontology, Fuzzy Term, Similarity, Analytic Hierarchies Process

1. 서 론

정보시스템(Information System)은 많은 양의 정보를 효과적으로 검색하고 저장하고 모델화하기 위해 설계된 시스템으로, 비구조화된 정보(텍스트)의 관리와 구조화된 구조(정형화된 데이터를 표현하는 사실을 근거한 정보)는 정보검색 시스템(IRS)과 데이터베이스 관리시스템(DBMS) 두 가지로 구분된다. 퍼지정보시스템은 DBMS나 IRS에서 부정확하고 불확실성의 표현 및 검색과 관련된 주요한 문제를 연구한다[10][15]. 정보시스템에서 문서 분류를 위한 대표적인 모델은 다음과 같다. 첫째, 규칙기반학습방법은 학습 문서들에서 나타나는 범주간의 구별된 규칙을 이용하여 전문가가 찾아주거나 학습을 통해 추출된 규칙을 이용하여 문서를 분류하는 방법으로 정확률은 높지만 재현율이 낮다[2]. 여기에는 어휘구문패턴기반학습방법[11][12][13], 정의문패턴기반학습방법[1], 수직관계기반학습방법이 있다. 둘째 통계기반학습방법은 학습 문서에서 특징을 추출하여 확률적으로 접근하고 문맥 정보 사이의 유사도 계산을 전제로 하고 있으며 기존의 통계적인 자연언어처리방법, 계층관계학습에 적용되는 방법으로 정확률은 낮지만 재현율은 높다[3].

본 논문은 특정분야의 문서에서 추출된 퍼지용어 정보를 바탕으로 온톨로지 구조를 카테고리화하여 확장된 AHP를 이용하여 퍼지용어의 전문성을 계산하여 주어진 퍼지용어의 상위어 후보를 레벨화한 후 Min 연산자, 다이스계수, Min+다이스계수 방법을 사용하여 퍼지용어 의미유사도를 계산한다. 마지막으로 선택된 후보들 중에서 최적의 상위어 후보를 결정하여 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 더 정확하게 문서를 분류하는 것을 알아본다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련연구를 통하여 온톨로지에 대한 배경지식을 알아보고 제 3장에서

접수일자 : 2008년 11월 1일
완료일자 : 2008년 12월 31일

는 퍼지용어의 전문성과 퍼지용어 의미유사도에 대하여 살펴보며 마지막으로 제 4장에서는 결론을 제시한다.

2. 관련 연구

2.1 온톨로지

Gruber가 정의한 "an Ontology is an explicit formal specification of a shared conceptualization of a domain of interest"가 일반적으로 사용되고 있다[14]. 또한 온톨로지의 특징은 개념간의 관계와 용어간의 관계를 분리하여 해당주제영역을 파악하고 주제영역내에서 일관성 있고 명확하게 개념을 정의하고 관계를 구조화하여 해당 주제영역의 특성을 보다 분명하게 반영할 수 있고 구조화된 지식으로 새로운 지식을 추론하고 지능적인 정보처리에 적용될 수 있다. 표 1은 시소러스와 온톨로지의 차이점과 유사점을 나타낸다[5].

표 1. 시소러스와 온톨로지
Table 1. Thesaurus and Ontology

	시소러스	온톨로지
차이점	-속성과 관계의 표현 제한 -각 용어는 관계의 한정된 수에 의해 기술 (계층(BT/NT)과 비계층(RT) 예) 동음이어(homonymous term) tiredness(psychology),tiredness(physics)	-속성과 관계의 표현을 제한하지 않음 -주어진 도메인을 설명하기 위해 필요한 관계는 용어들과 관련되어 사용되고 이러한 관계는 개념들 사이를 구별하기 위해 사용(지식획득에 중요)
유사점	특정도메인	
	용어들 사이의 용어나 관계 설명	
	계층구조	
	카탈로그나 검색정보에 대한 정보관리 애플리케이션사용	
	지속적인 개정과 유지 필요	

기존의 온톨로지들은 대부분 전문가의 수작업에 의존하고 있어 시간 및 인적 제약 때문에 실용적인 온톨로지를 구축하기 어렵다. 앞으로 온톨로지에서 표현되는 여러 가지 관계 중에서 가장 핵심인 개념간 계층관계를 자동으로 추출하는 방법은 전문가의 수작업을 최소화할 수 있고 여러 전문가들의 작업결과가 일관성을 가지게 된다. 그림 1은 온톨로지 학습단계를 나타낸다[2][6].

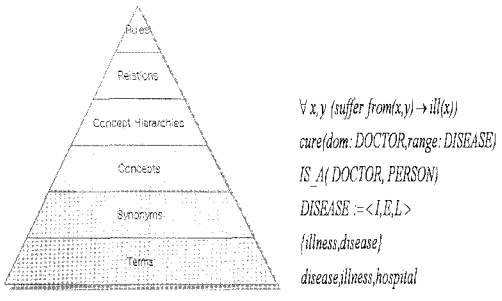

그림 1. 온톨로지 학습단계
Fig. 1. Ontology learning step

온톨로지 학습에서 가장 기본 단계인 "Terms" 단계에서는 온톨로지 구축을 위한 대상 용어를 추출하고 선정하며 "Synonyms" 단계에서는 선정한 용어들 사이의 동의어를 그룹핑하고 "Concepts" 단계에서는 그룹핑된 용어들을 개념으로 표현하고 "Concept Hierarchies" 단계에서는 개념들 사이의 상하위 관계를 설정하고 "Relations" 단계에서는 상하위어 관계 이외의 다양한 관계를 표현하며 마지막으로 "Rules" 단계에서는 개념 사이의 관계를 논리 형태로 표현한다. 전체 학습 단계에서 "Concept Hierarchies"는 개념들을 조직화하는 가장 기본적이고 필수적인 단계이다[2][6]. 개념간 상하위 관계는 개념간 상속관계를 표현하기 때문에 지능형 시스템에서 상하위어 관계 탐색을 통한 추론기능을 제공한다. 용어 계층 구조는 용어들 사이의 계층 관계를 설정하여 조직화시킨 것으로 계층구조에 포함된 모든 용어는 한 개 이상의 용어와 계층 관계를 가진다. 계층관계는 IS-A, PART-OF, INSTANCE-OF 등의 관계를 포함한다. 온톨로지의 연구분야로는 전문분야 온톨로지와 데이터 및 메타데이터 온톨로지, 웹온톨로지, 의미망 온톨로지 등이 있다.

2.2 확장된 AHP

계층분석과정(Analytic Hierarchy Process :AHP)은 1970년대 초 T. L. Saaty에 의하여 개발된 방법론[16]으로 주어진 의사결정문제를 계층화 형태로 표현하여 상위계층에 있는 한 요소관점에서 하위계층에 있는 한 요소들의 상대적 중요도 또는 가중치를 쌍비교에 의해 측정하는 방법으로 의사결정자의 오랜 경험과 직관 등을 평가의 기본으로 하여 수치로 표현할 수 있는 정량적 평가기준과 의사결정문제에서 다루기 힘들지만 다루어야하는 정성적인 기준에 대한 평가기준도 처리하기 쉽다는 장점을 가지고 있다. 그러나 평가들간의 평가능력이 동등하게 설정되어 다수평가가자에 대한 평가결과에 차별을 두지않는다는 단점이 있다. 따라서 본 논문에서는 가중치가 부여된 다수평가자의 평가치를 통합하는 확장된 AHP를 제안한다.

3. 퍼지용어의 전문성과 의미유사도

본 논문은 온톨로지를 "어휘들에 대해서 일정 영역의 개념적 예들을 한곳으로 집합시킨 하나의 독립된 집합체"로 정의하고 문제도메인을 다음과 같이 정의한다.

3.1 문제 도메인 정의

데이터는 웹애플리케이션에서 고장의 원인과 보급에 대하여 cnet.com과 eweek.com과 같은 technology website상에 리스트된 웹사이트 정전사건과 시스템 고장의 케이스연구를 조사하여 수집한다. 'possible failure of Web Server'의 문제 도메인을 고려한 온톨로지에서 퍼지용어의 전문성과 의미유사도를 나타낸다. 퍼지용어는 사용자가 일상적으로 사용하는 애매한 용어로써 동일한 퍼지용어도 사용되는 분야에 따라서 서로 다른 특징집합을 가지고 퍼지용어 의미유사도는 분야 의존적인 성질을 가진다. 'possible failure of web server'의 문제 도메인에 대하여 'computer virus', 'disk space', 'memory space'와 같은 서버 고장에 관한 여러 가지 관련된 퍼지용어가 있다.

이 문제도메인을 이용하여 어휘구문패턴기반학습중에 전

체-부분관계를 구글 (possible failure of web server, 약 3,880,000 문서)에서 10개의 용어(HD space, memory space, computer virus, H/W damage, power failure, program bugs, OS failure, CPU loading, Network loading, Hacker)의 "부분"을 추출했으며 각 개념은 하위개념들의 집합을 가진다. 예를 들면 Network loading는 network congestion, configuration errors, router failures 등이다.

과거 경험이나 관련 배경이 다른 전문가1은 이 문제 도메인에 대하여 "HD space, memory space, computer virus, H/W damage, power failure, program bug, OS failure, CPU loading"과 같은 퍼지용어를 표 2에 나타낸다.

표 2. 전문가1의 제안
Table 2. proposal of expert1

용어	소속함수	용어설명
HD space	(10,30,50)	Free disk space
Memory space	(100,300,500)	Free memory space
Computer Virus	(1,3,5)	Virus information of trend micro
H/W damage	N/A	Status of computer hardware
Power failure	(10%, 30%, 50%)	Power remain
Program bug	N/A	Operative condition of the program
OS failure	N/A	Server operative condition
CPU loading	(80%, 90%, 100%)	CPU loading

전문가2는 "Network loading, CPU loading, memory space, Hardware failure, hacker"를 제안하여 표 3에 나타낸다.

표 3. 전문가2의 제안
Table 3. proposal of expert2

용어	소속함수	용어설명
Network loading	(50%,70%,100%)	Network loading
CPU loading	(70%,80%,90%)	CPU loading
Memory space	(100,300,500)	Free memory space
Hacker	(1,3,5)	Rank of dangerous for server
Hardware failure	N/A	Staus of computer hardware

두 전문가가 제시한 퍼지용어 중 'H/W damage'와 'Hardware failure'의 용어는 'H/W damage' 용어로 의견을 통합[8]하고 'CPU loading'과 'Network loading'을 채택하여 표 4에 나타낸다.

통합된 제안을 바탕으로 서버 고장에 대한 퍼지용어들의 온톨로지 관계를 그림 2에 나타낸다.

표 4. 통합된 제안
Table 4. Integrated proposal

용어	소속함수	용어설명
HD space	(10,30,50)	Free disk space
Memory space	(100,300,500)	Free memory space
Computer Virus	(1,3,5)	Virus information of trend micro
H/W damage	N/A	Status of computer hardware
Power failure	(10%, 30%, 50%)	Power remain
Program bug	N/A	Operative condition of the program
OS failure	N/A	Server operative condition
CPU loading	(75%, 85%, 95%)	CPU loading
Network loading	(50%,70%,100%)	Network loading
Hacker	(1,3,5)	Rank of dangerous for server

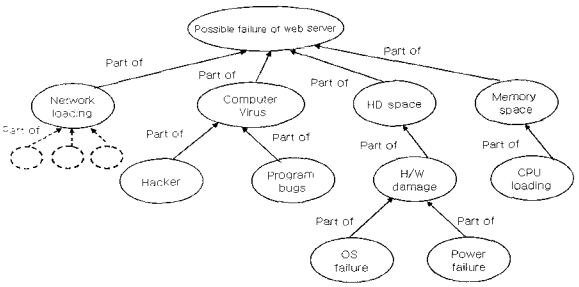

그림 2. 온톨로지에서 퍼지용어들의 관계
Fig. 2. relation of fuzzy terms in Ontology

3.2 퍼지용어 전문성

퍼지용어의 전문성은 퍼지용어가 포함하는 전문적인 정보의 양을 정량적으로 표현한 것이다[9]. 다기준 의사결정 문제인 퍼지용어의 전문성을 나타내기 위하여 'possible failure of web server' 도메인을 이용한다. 이것은 의사결정 기준의 집합에 의한 대안 집합의 평가로 사용된다. 다수의 다기준 의사결정 방법들로 가중치 합 모델(the weighted sum mode), 가중치 곱 모델(the weighted product model), 계층분석과정(AHP)등이 제안되어왔다[7]. 퍼지용어의 전문성을 레벨화하기 위해 확장된 AHP 방법은 가중치가 부여된 다수평가자의 평가치를 통합한 후 기준의 쌍비교를 통해서 기준의 가중치나 상대적 중요성을 결정함으로써 의사결정 문제들을 다룬다. 쌍비교 행렬로부터 각 기준의 상대적 가중치를 유도하기 위해 다양하게 사용되는 방법은 eigenvecter(고유벡터)와 기하 평균이 있는데 본 논문은 고유벡터 분석의 left-right 고유벡터 비대칭과 퍼지용어들 중에서 상대적 측정에 대한 종속 문제를 해결하기 위해 기하 평균과 가중치를 사용한다. 표 5는 쌍비교 행렬 구조 예로서 기하평균은 식 (1), 가중치는 식 (2)에 나타낸다.

표 5. 쌍비교행렬의 구조
Table 5. structure of pair-wise comparison matrix

	V_1	V_2	⋯	⋯	V_n	GM	W_i
V_1	1	V_{12}	⋯	⋯	V_{1n}	GM_1	W_1
V_2	V_{21}	1	⋯	⋯	V_{2n}	GM_2	W_2
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
V_i	V_{i1}	V_{i2}	⋯	⋯	V_{in}	GM_i	W_i
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
V_n	V_{n1}	V_{n2}	⋯	⋯	1	GM_n	W

$$GM_i = \left(\prod_j^n V_{ij} \right)^{\frac{1}{n}} \quad (1)$$

단, $i = 1, n, V_{ii} = 1$

$$W_i = \frac{GM_i}{(GM_1 + GM_2 + \ldots + GM_n)} \quad (2)$$

표 6은 'possible failure of web server' 도메인에서 채택된 퍼지용어 10개를 쌍비교 행렬로 표현하여 비교하고 각 퍼지용어의 기하평균과 가중치를 계산한다. 퍼지용어 C_1은 HD space, C_2는 Memory space, C_3은 Computer virus, C_4는 H/W damage, C_5는 Power failure, C_6은 Program bugs, C_7은 OS failure, C_8은 CPU loading, C_9는 Network loading, C_{10}은 Hacker를 의미한다. 표 3에 나타낸 쌍비교 행렬의 값은 1부터 5까지 분류된다. 여기서 1은 동등한 중요성을 의미하고 2는 중간적 중요성을 의미하고 3은 강한 중요성 4는 매우 강한 중요성 5는 극히 강한 중요성을 의미한다. 그의 반대는 역수 1/2, 1/3, 1/4, 1/5 로 나타낸다. 다른 퍼지용어들을 비교함으로써 C_1, C_2, C_3, C_4, C_5,C_6,C_7,C_8,C_9, C_{10}의 가중치는 각각 0.0626, 0.1065, 0.1047, 0.1253, 0.1673, 0.0451, 0.1606, 0.0875, 0.0518, 0.0886으로 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.

표 6. 쌍비교행렬
Table 6. 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	W_i
C_1	1	1/4	3	1/3	1/2	1/2	1/4	1/2	3	1/2	0.6871	0.0626
C_2	4	1	1/3	1/3	1/4	4	1/3	2	4	4	1.1684	0.1065
C_3	1/3	3	1	1/2	1/3	4	3	1/2	2	2	1.1487	0.1047
C_4	3	3	2	1	1/2	1	1/2	4	4	1/3	1.3741	0.1253
C_5	2	4	3	2	1	3	4	3	1/2	1/2	1.8346	0.1673
C_6	2	1/4	1/4	1	1/3	1	1/4	1/2	1/3	1/2	0.4941	0.0451
C_7	4	3	1/3	2	1/4	4	1	3	3	4	1.7617	0.1606
C_8	2	1/2	2	1/4	1/3	2	1/3	1	2	3	0.9603	0.0875
C_9	1/3	1/4	1/2	1/4	2	3	1/3	1/2	1	1/3	0.5676	0.0518
C_{10}	2	1/4	1/2	3	2	2	1/4	1/3	3	1	0.9716	0.0886

위의 결과를 토대로 퍼지용어의 계층구간을 표 7과 그림 3에 나타낸다.

표 7. 퍼지용어의 계층구간
Table 7. level interval of fuzzy terms

퍼지용어	구간
C_6, C_9	0.04~0.06
C_1	0.06~0.08
C_8, C_{10}	0.08~0.1
C_2, C_3	0.1~0.12
C_4	0.12~0.16
C_5, C_7	0.16~

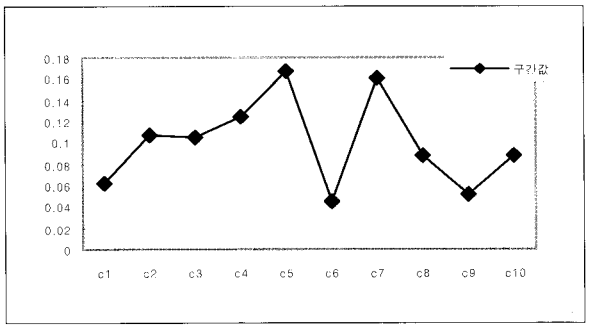

그림 3. 퍼지용어 계층구간
Fig 3. level interval of fuzzy terms

3.3 퍼지정보시스템에서 퍼지용어 의미유사도

퍼지정보시스템에서 두 퍼지용어가 비슷한 문맥에서 사용되는 경우 의미적으로 유사하고 사전적 정의문이 유사한 경우 두 퍼지용어가 유사하다고 판단할 수 있다. 의미적으로 유사한 퍼지용어를 정의할 때 비슷한 단어를 이용하여 정의하기 때문에 정의문에 나타나는 단어를 비교하여 유사한 정도를 판단할 수 있다. 그리고 특징 집합이 완전히 일치하거나 포함관계에 있거나 부분적으로 겹치는 관계에 있거나 또는 전혀 겹치지 않는 경우도 포함한다. 동일한 퍼지용어도 사용되는 분야에 따라서 서로 다른 특징 집합을 가진다. 따라서 퍼지용어 유사도도 분야 의존적인 성질을 가진다.

3.3.1 Min 연산자 의미유사도

'possible failure of web server' 도메인에서 확장된 AHP를 이용해 퍼지용어 전문성을 구한 다음 퍼지집합의 Min 연산자를 이용하여 퍼지용어 의미유사도를 비교한다. 퍼지용어 의미유사도는 퍼지용어의 특징 집합 사이의 포함관계의 정도를 정량적으로 표현한 것이다. 퍼지집합은 하나의 문헌이 특정한 색인어로 표현되는 개념 범주에 속하는 정도를 나타내는 부분 소속도로 A AND B 쿼리에 대한 문헌 X에 나타난 두 용어 A와 B의 소속 함수값을 비교한다. 퍼지집합의 성질은 다음과 같이 정의한다.

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

퍼지용어의 관계행렬은 퍼지용어의 문맥정보로서 두 퍼지용어에 대한 소속도를 나타낸다. C_1(HD space)와 C_2(Memory space)의 문맥정보벡터는 각각
$C_1 = (0.6, 0.6, 0.4, 0.8, 0.8, 0.4, 0.7, 0.5, 0.7, 0.5)$
$C_2 = (0.6, 0.5, 0.4, 0.7, 0.8, 0.3, 0.6, 0.5, 0.7, 0.4)$. 이고 퍼지집합

의 Min 연산자를 사용한 두 벡터 사이의 유사도구간은 표 8에 나타낸다.

표 8. Min연산자를 사용한 유사도 구간
Table 8. similarity interval using min operation

퍼지용어	구간	퍼지용어	구간
C_1	[0.1-0.4]	C_6	[0.1-0.1]
C_2	[0.1-0.3]	C_7	[0.1-0.4]
C_3	[0.1-0.2]	C_8	[0.1-0.3]
C_4	[0.1-0.5]	C_9	[0.1-0.5]
C_5	[0.1-0.6]	C_{10}	[0.1-0.2]

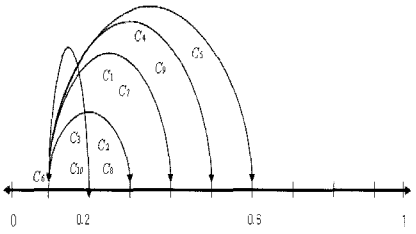

그림 4. Min연산자를 사용한 유사도
Fig 4. similarity using Min operation

표 8의 결과에 의하면 퍼지용어의 의미유사도 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, Max 가중치를 선택하기 때문에 색인어에 가중치를 부여해야만하고 검색된 문헌의 순위 부여 능력이 모든 검색어에 민감하지 못하는 단점을 가지고 있다.

3.3.2 다이스계수(dice's coefficient) 의미유사도

문헌과 질의에 N 차원 벡터 값을 표시하여 각 퍼지용어의 벡터 간 유사도를 산출한다. 두 개의 벡터가 각각 유사도를 비교하고자 하는 두 개의 퍼지용어의 문맥정보를 대표한다고 할 때 이 계산 방법에서는 두 문맥정보의 관련성 척도를 나타낸다. 다이스 계수 유사도 계산 방법은 식(3)과 같이 표현된다.

$$S(C_1, C_2)=\frac{2\,|\sum(X_i - Y_i)\,|}{\sum X_i^2 + \sum Y_i^2} \qquad (3)$$

여기서 $X=\sum_{i=1}^{n} x_i$와 $Y=\sum_{i=1}^{n} y_i$는 각각 두 퍼지용어 C_1과 C_2의 특징에 대하여 가중치를 나타내는 벡터이다. 표 9는 다이스계수를 사용한 퍼지용어 의미유사도의 결과를 구간별로 나타낸다.

표 9. 다이스계수 사용한 유사도 구간
Table 9. similarity interval using dice's coefficient

퍼지용어	구간	퍼지용어	구간
C_1	[0.1-0.9]	C_6	[0.3-1.1]
C_2	[0.1-0.8]	C_7	[0.1-1.0]
C_3	[0.2-0.9]	C_8	[0.1-0.7]
C_4	[0.0-1.0]	C_9	[0.0-1.0]
C_5	[0.1-1.1]	C_{10}	[0.2-0.7]

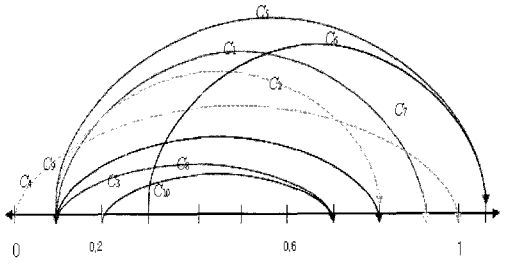

그림 5. dice's coefficient를 사용한 유사도
Fig 5. similarity using dice's coefficient

표 9의 결과에 의하면 퍼지용어의 의미유사도 계층은 C_5, $C_4,C_9>C_7>C_1,C_6>C_2,C_3>C_8>C_{10}$ 순으로 구성된다. 그러나 두 벡터 X, Y 사이의 유사도 계산 방법인 타니모토 계수(Tanimoto coefficient), 다이스 유사계수(dice's coefficient), 자카드 유사계수(Jaccard's coefficient), 중복도 계수(Overlap coefficient)들은 문서분류의 정확도를 어느 정도 보장하지만 미리 규칙을 위한 학습과정이 필요하며 그에 따른 학습 데이터가 반드시 필요하다.

3.3.3 Min+다이스계수 의미유사도

Min 연산자와 다이스계수를 결합하여 퍼지용어의 의미유사도를 구간별로 나누면 표 10과 같이 나타난다. 표 10의 Min+다이스계수를 사용한 결과에 의하면 퍼지용어의 의미유사도는 $C_{10}>C_5>C_4$, $C_9>>C_2$, $C_8>C_7>C_1$, C_3,C_6 순으로 구성된다.

표 10. Min+다이스계수를 사용한 유사도 구간
Table 10. similarity interval using Min+dice's coefficient

용어	구간	양수	용어	구간	양수
C_1	[0.1-1.5]	3	C_6	[0.9-2.3]	0
C_2	[0.1-2.1]	4	C_7	[0.1-2.0]	6
C_3	[0.5-1.9]	1	C_8	[0.1-2.2]	5
C_4	[0.1-2.2]	7	C_9	[0.2-2.3]	8
C_5	[0.2-2.6]	9	C_{10}	[0.1-2.6]	2

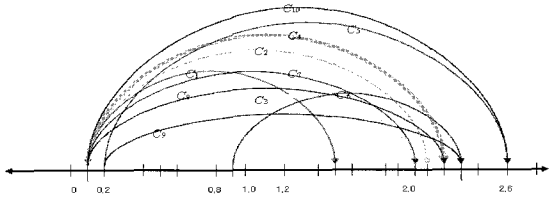

그림 6. Min+dice's coefficient를 사용한 유사도
Fig 5. similarity using Min+dice's coefficient

위의 결과로부터 본 논문은 서버 고장의 문제도메인에서 퍼지 용어 전문성 C_5, C_7 레벨에서 퍼지집합의 Min 연산자 의미유사도는 $C_5>C_4,C_9>C_1$, $C_7>C_2,C_8>C_3,C_{10}>C_6$ 순의 유사도를 가지고 다이스계수 의미유사도는 $C_5,C_4,C_9>C_7>C_1,C_6>C_2,C_3>C_8>C_{10}$ 순의 유사도를 가지며 Min+다이스계수 의미유사도는 $C_{10}>C_5>C_4$, $C_9>C_2$, $C_8>C_7>C_1$, C_3,C_6 순으로 관계를 가진다는 것을 알았다. 약 3,880,000 문서를 실행한 결과를 표 11에서 나타낸다.

표 11. 평가 결과
Table 11. result of evaluation

단계	Min연산자	다이스계수	Min+다이스
1	45,900	2,700	34,500
2	2,700	2,120	16,800
3	10,700	735	867
4	9,670	482	696

4. 결 론

본 논문은 문서에서 추출된 퍼지용어 정보를 바탕으로 한 온톨로지 구조를 카테고리화하여 퍼지용어의 전문성을 이용하여 주어진 퍼지용어의 상위어 후보를 레벨화한 후 퍼지용어 의미유사도를 계산한 후 선택된 후보들 중에서 최적의 상위어 후보를 결정한다. 즉, 퍼지용어간의 전문성을 레벨화하기 위한 확장된 AHP방법은 가중치가 부여된 다수평가자의 평가치를 통합한 후 퍼지용어의 쌍비교를 통해서 가중치나 상대적 중요성을 결정한다. 그리고 퍼지용어 의미유사도는 퍼지집합의 Min연산자와 다이스계수, Min+다이스계수를 비교한다. 이러한 방법들은 문서들이 가지는 의미론적 내용과 관계의 식별을 바탕으로 보다 더 정확하게 문서를 분류 할 수 있고 자연어처리 등에 많이 활용될 수 있을 것이다. 향후 퍼지용어의 의미유사도를 다치 형태 확장의 연구가 요구된다.

참 고 문 헌

[1] ISO, "Terminology work-principle and methods", *ISO 704 second edition*, 2000.

[2] 최기선, 류법모, "온톨로지의 구축과 학습 : 상하위 관계", *정보과학회지 제 24권 제 4호*, pp. 24-30, 2006.

[3] Lee, L., "Measures of Distributional Similarity", *Proceedings of ACL*, pp. 25-32, 1999.

[4] 옥철영, "한국어정보처리와 온톨로지", 2004 *한국어정보처리연구회 동계 튜토리얼 자료집*.

[5] Chang Chun, Lu Wenlin, "From Agricultural Thesaurus to Ontology", *Agricultural Information and Knowledge Management Papers*, 2002.

[6] Lassila, O., McGuinness, D., "The Role of Frame-Based Representation on the Semantic Web", *Technical Report KSL 01-02, Knowledge System Laboratory, Stanford University*, 2001.

[7] Yuxia Huang, Ling Bien, "A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet", *Expert System with Applications*, 2007.

[8] 류경현, 정환묵, "MFAC를 사용한 근접관계의 분류", *한국퍼지및지능시스템학회논문지*, Vol. 18, No. 1, pp.139-144, 2008.

[9] Learhoven. P. J. M., & Pedrycz, W., "A Fuzzy extension of Satty's priority theory", *Fuzzy Sets and Systems*, 11, 1983.

[10] Donald H. Kraft, Frederick E. Petry, "Fuzzy information systems : managing uncertainty in databases and information retrieval systems", *Fuzzy Sets and Systems* 90, pp. 183-191, 1997.

[11] Hearst, M., A., "Automatic Acquisition of Hyponyms from Large Corpora", *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 1992.

[12] Caraballo, S. A., "Automatic construction of a hypernym-labeled noun hierarchy from text", *Proceedings of ACL*, 1999.

[13] Berland, M., Charniak, E., "Finding Parts in Very Large Corpora", *Proceedings of ACL*, 1999.

[14] Gruber, T. R., "A Translation approach to portable ontology specifications", *Knowledge Acquisition*, 5(2), pp. 199-220, 1993.

[15] Abraham Silberschatz, Henry F. Korth, S. Sudarshan, *Database System Concepts* 5th edition, McGraw-Hill, 2005.

[16] Saaty, T. L., *The Analytic Hierarchy Process*, McGraw-Hill, 1980.

저 자 소 개

류경현(Kyung-Hyun Ryu)

1990년 : 대구가톨릭대학교 전산통계학과 학사

1992년 : 대구가톨릭대학교 전산통계학과 석사

2006년 : 대구가톨릭대학교 컴퓨터정보통신공학부 박사과정

2006년~현재 : 대구가톨릭대학교 컴퓨터와디지탈정보 강의전담교수

관심분야 : 지능에이전트, 데이터마이닝, 의사결정, 온톨로지

E-Mail : r11047@cu.ac.kr

정환묵(Hwan-Mook Chung)

18권 1호 참조

Phone : +82-53-850-2741

Fax : +82-53-850-2741

E-mail : hmchung@cu.ac.kr