온톨로지 계층관계를 이용한 서비스 발견 알고리즘

최원종[O] 양재영 최중민
한양대학교 컴퓨터공학과 지능시스템 연구실
{wjchoi[O], jyyang, jmchoi}@cse.hanyang.ac.kr

조현규 조현성 김경일
한국전자통신연구원 지능형웹기술연구팀
{hkcho, hsc, zbmon}@etri.re.kr

Wonjong Choi[O] Jaeyoung Yang Joongmin Choi
Intelligent Systems Lab,
Dept. of Computer Science and Engineering, Hanyang University

Hyeonkyu Cho Hyeonsung Cho Kyungil Kim
Intelligent Web Technology Team

요 약

인터넷망의 지속적인 발달과 더불어 웹서비스가 차지하는 비중은 매우 커지고 있다. 이와 관련해 서비스 발견을 위한 다양한 노력들이 진행되었으며, 그 중에서도 DAML-S문서로 기술된 매치메이커에서 제시한 알고리즘은 서비스 발견자와 서비스 제공자사이의 서비스 발견에 대한 유사도 측정의 한 방법을 제시하고 있다. 하지만 온톨로지상의 관계표현에 있어 네 가지 규칙만을 적용하여 정밀한 유사도측정이 불가능하다는 단점이 있다. 따라서 본 논문에서는 기존의 알고리즘의 개선을 위해 두 가지 유사도 측정함수 1) 계층구조함수 2) 계층계수함수 를 정의하고, 이에 기반한 새로운 서비스 발견 알고리즘을 제시하고자 한다.

1. 서 론

오늘날 인터넷망의 지속적인 발달은 기존의 오프라인형태의 기업서비스들에서 비즈니스차원의 웹서비스로의 통합 및 재창출의 계기로 작용하고 있다.
이에 W3C에서는 웹서비스 표준안으로서 WSDL, SOAP, UDDI를 제창하였다. WSDL과 SOAP은 웹서비스의 개발 및 구축을 담당하며, UDDI의 경우 서비스의 통합적인 관리 및 발견이라는 측면을 담당하고 있다. 최근에는 DAML+OIL의 장점을 이용한 서비스기술언어의 하나인 DAML-S가 등장했다 DAML-S의 경우 기존의 UDDI에 비해 뛰어난 표현력을 가지고 있으며, 특히 서비스자체의 홍보적인 요소가 풍부하다. 이러한 서비스홍보에 대한 표현력은 서비스발견에 있어서 좀더 효과적인 검색을 가능하게 하며, 실제로 이러한 서비스발견에 대한 다양한 형태의 방법들이 제안되고 있다. 이 중 DAML-S에서 정의된 서비스요소간의 온톨로지관계를 이용한 매칭방법이 가장 대표적이다. 본 논문의 구성은 다음과 같다. 2장에서 온톨로지 정보를 이용한 기존의 매칭방법을 살펴보고, 3장에서는 기존의 매칭방법을 개선한 형태의 방법을 소개할 것이다. 4장에서는 3장의 내용에 대한 간단한 결론과 향후연구에 대해 언급하겠다.

2. 종래의 매치메이커 시스템에 대한 요약

종래의 매치메이커시스템은 DAML-S로 기술된 profile문서의 서비스 입력과 출력부분의 내용을 분석하여, 서비스 제공자와 서비스발견자 사이의 유사도를 측정한다.
유사도 측정을 위한 규칙은 서비스제공자와 서비스 발견자가 제공하는 각각의 문서에 정의되어 있는 서비스에 대한 입출력부분이 참조하는 온톨로지노드집합간의 계층관계에 대해 정의하고 있다..

Exact : if request equivalent to advertisement.
Plug in : if advertisement subsumes request.
Subsume : if request subsumes advertisement.
Fail : failure occurs when no subsumption relation

표 1 매치메이커 시스템의 유사도 측정규칙

유사도 측정규칙은 표1에서처럼 "exact", "plug in", "subsume", "fail"의 네 가지 계층관계로 구분되어지며, "exact"가 가장 높은 유사도를, "fail"로 갈수록 낮은 유사도를 가진다.

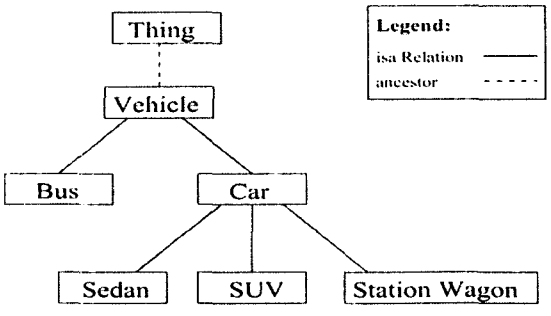

그림 1 Vehicle 온톨로지

S_R = 서비스발견자가 제공한 문서에 매칭되는 온톨로지노드 집합
S_A = 서비스제공자가 제공한 문서에 매칭되는 온톨로지노드 집합
M_{RA} = 유사도 측정규칙에 의한 결과값들의 집합

표 2 축약표현의 정의

참고로 앞으로의 설명의 편의를 위해 표2의 축약표현을 사용하도록 하겠다.

가령 그림1의 vehicle 온톨로지에서 각각의 집합이 S_R ={Car}이고, S_A ={Car, Sedan, Vehicle}일 경우 M_{RA} ={exact, subsumes, plug in}이 된다. 따라서 " exact" 에 해당하는 " Car" 와 매칭되는 서비스제공자의 내용을 서비스발견자가 찾는 것에 가장 유사한 것으로 본다. 마찬가지로 S_R 은 동일하고, S_A ={Sedan, Vehicle}일 경우 M_{RA} ={subsume, plug in}이 된다. 따라서 " plug in" 에 해당하는 " Vehicle" 과 매칭되는 서비스제공자의 내용을 가장 유사하다고 인정한다. 이는 계층관계에서 상위개념이 하위개념보다 더욱 일반적인 개념이므로, 상위노드에 해당하는 서비스가 하위노드에 해당하는 서비스보다 더욱 더 포괄적인 서비스를 제공한다는 것을 의미한다. 따라서 매치메이커 시스템의 유사도 측정규칙은 일반적인 계층구조의 특징을 그대로 사용하고 있음을 알 수 있다,

하지만 이러한 유사도 측정 규칙은 온톨로지의 계층구조 관계를 단지 네 가지 규칙으로만 판단한다는 문제점을 가지고 있다. 가령 상위노드의 관계의 경우 " 증조할아버지" , " 할아버지" , " 아버지" 등의 관계로 다시 세분화할 수 있으며, 이는 하위노드의 관계에서도 " 증손자" , " 손자" 등으로 세분화 될 수 있다. 즉 이러한 계층관계는 다시 세분화된 형태로 표현이 가능함을 보여준다.

가령 S_R ={Vehicle}이고, S_A ={Car, SUV}일 경우 M_{RA} ={subsume, subsume}이 된다. 따라서 결과적으로 " Car" 와 " SUV" 모두 동일한 유사도를 가진다고 판단한다. 하지만 실제로 서비스발견자의 입장에서는 " Car" 쪽을 추천하는 것이 더 합당하다. 이러한 불합리성은 S_A 집합의 개수가 많아질수록, 동일한 랭킹의 증가에 대한 문제를 야기시킨다. 물론 종래의 매치메이커알고리즘은 DAML-S로 기술된 서비스제공자의 profile부분의 입력과 출력부분 외에 provider부분의 내용을 분석함으로서, 이러한 문제점을 해결할 수 있다고 말하고 있다. 하지만 이것 역시 동일한 provider를 가진 다른 서비스집합이나 혹은 provider내용이 존재하지 않는 경우, 여전히 동일한 문제를 발생시킨다. 따라서 이러한 유사도 측정규칙과는 다른 온톨로지노드사이의 거리관계에 기반한 새로운 유사도 측정 알고리즘을 제안하고자 한다

3. 노드간의 유사도 측정을 위한 새로운 알고리즘

본 논문에서 제시하는 유사도 측정알고리즘은 종래의 유사도측정규칙과는 다르게 온톨로지 노드사이의 거리관계와 계층구조모두를 고려한다. 그리고 온톨로지 노드사이의 계층구조파악의 용이성을 위해 각각의 노드들에 고유한 ID를 부여하고 있다. 이러한 ID부여는 유사도 측정함수에서 노드사이의 거리관계를 따지는데 있어 매우 유용하며, 특정 노드가 어떠한 계층관계를 가지고 있는가를 파악하는 데 중요하다.

3.1 온톨로지 노드사이의 계층구조에 대한 ID정의

각각의 온톨로지에 존재하는 노드는 고유한 ID를 부여받게 되는데, ID자체는 그 노드의 상위로부터의 어떠한 경로를 거쳐왔는지에 대한 정보라고 할 수 있다. 즉 각각의 자식 노드가 가지고 있는 ID는 각각 바로 윗단계의 부모노드의 ID를 포함하고 있으며, 이는 자신의 부모가 누구인지를 구분하는 동시에 자신의 시블링노드에 대한 정보도 포함하고 있음을 알 수 있다.

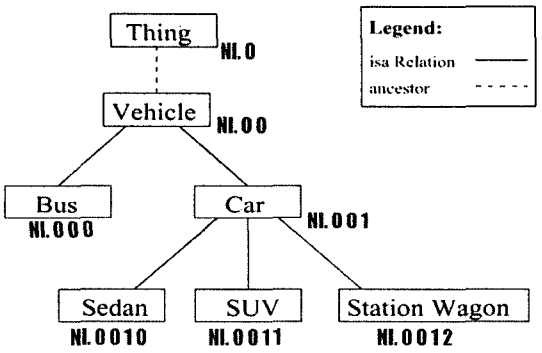

그림 2 노드 ID가 부여된 vehicle 온톨로지

그림2는 실제 노드의 ID가 어떻게 부여되는가를 표현한 것이다. 최상위 노드에 해당하는 " Thing" 의 경우 " 0" 이라는 ID값을" Vehicle" 의 경우 " 0 0" 이라는 ID값을 가지고 있다. 항상 노드의 마지막 상수는 현재 노드의 동일한 시블링들 사이의 순서를 의미하며, 나머지 ID값들은 현재노드에 도달하기 위한 상위노드들의 위치정보를 의미한다. 따라서 " Car" 와 " Sedan" 은 " Vehicle" 의 후손임을 알 수 있다. 그밖에 중요하게 봐야 할 점이 각 노드의 ID들간의 길이관계는 온톨로지 노드사이의 거리관계를 표현하고 있다는 점이다. 즉 노드사이의 거리차이가 1인 경우 ID의 길이도 1만큼의 차이를 가진다. 이러한 ID의 길이정보는 상대적이며, 온톨로지 노드의 거리를 고려한 유사도 계산에 사용함으로서 종래의 유사도 측정 규칙에 비해 좀 더 세밀한 유사도 계산을 가능하게 함을 알 수 있다.

3.2 유사도 측정 함수의 정의

노드 사이의 유사성측정은 앞에서 언급했던 ID값을 이용하여 이루어진다.단 전제조건으로 서비스발견자와 서비스제공자에 매칭된 각각의 온톨로지노드에 대한 ID정보는 전처리과정을 거쳐 집합의 형태로 제공된다고 가정한다.

$$SF_{Onto.Node}(R,A)=\begin{cases} R.index \textbf{ includes } A.index \Rightarrow 1-DS_{Onto.Node}(R,A) \\ R.index \textbf{ belongs to } A.index \Rightarrow NF-DS_{Onto.Node}(R,A) \\ R.index \textbf{ equals in } A.index \Rightarrow 1 \\ \textbf{Other relation} \Rightarrow 0 \end{cases}$$

수식 1 계층구조함수의 정의

$$DS_{Onto.Node}(A,B)=NF*\frac{|Length(A.Index)-Length(B.index)|}{max\ Length(A.Index, B.index)}$$

수식 2 계층계수함수의 정의

기본적으로 유사도 측정함수는 크게 수식1의 " 계층구조함수와 수식2의 " 계층계수함수" 로 구성된다.
계층구조함수는 노드사이의 기본적인 관계를 정의하는데 사용되며, 서비스발견자와 서비스제공자사이의 포함관계를 나타낸다. 그리고 각각의 관계에 따라 적절한 계산법을 적용한다. 이는 기본적으로 온톨로지노드의 부모노드가 자식노드보다 더 일반적이고 더욱 포괄적인 개념을 가지고 있다는 것에 착안한 계산법이다.
다음으로 계층계수함수의 경우 실제 온톨로지노드사이의 거리차이를 0에서 *NF*사이의 값으로 표현한다. 가령 두 노드사이에 대한 계층계수함수값이 0일 경우 두 노드는 완벽하게 일치하는 것을 의미하며, 노드사이의 거리가 멀어질수록 *NF*에 가깝게 된다. 여기서 *NF*값은 normal factor로 0보다 크고 0.5보다 작거나 같은 값 중에 임의로 선택할 수 있다.

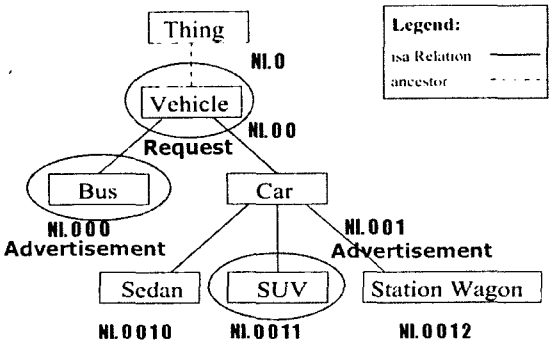

그림 3 노드 ID를 사용한 score 계산

그림3은 ID가 부여된 온톨로지에서 S_R ={Vehicle}이고, S_A ={Bus, SUV}일 경우, 유사도 측정함수의 계산과정을 보이기 위한 예이다.
기존의 유사도 측정규칙을 적용한 알고리즘에 따르면, M_{RA} ={subsume, subsume}이 된다. 그리고 만약 두 개의 DAML-S profile에서 provider정보가 동일하다면, 두 개 모두가 동일한 유사성을 가진다고 인정한다. 하지만 앞에서 정의한 유사도 측정함수를 적용하는 경우, 같은 subsume의 관계에도 score값을 이용한 랭킹이 가능하다 .
실제 적용 예를 들기 전에 설명의 편의를 위해, 아래 표3의 축약표현을 추가로 사용함을 명시하겠다.

S_{AI} = S_A 가 가지는 ID값들 의 집합
S_{RI} = S_R 이 가지는 ID값들의 집합
M_{SC} = 유사도측정에 대한 최종 score값들의 집합
$S_{AI}.n$ = S_{AI} 의 n번째 ID값
$S_{RI}.n$ = S_{RI} 의 n번째 ID값

표 3 추가된 축약표현

그림3의 예에서 S_R 과 S_A 가 매칭된 노드의 ID값들은 각각 S_{AI} ={" 0 0 0" , " 0 0 1 1" }과 S_{RI} ={" 0 0" }로 표현 할 수 있다. 따라서 수식1의 계층구조함수를 적용하는 경우 $S_{RI}.1$는 $S_{AI}.1$, $S_{AI}.2$에 대해 동일하게 상위노드의 관계를 가지고 있으며, 수식2의 계층계수 함수를 사용하여 노드사이의 거리에 대한 스코어계산을 하면, 각각의 ID의 길이차이가 1, 2이므로 M_{SC} ={ *NF* - *NF**0.33, *NF* - *NF**0.5}가 되고, $S_{AI}.1$의 ID값을 가지는 " Bus" 쪽이 더 유사하다고 판단하게 된다.

3.3 복수의 입출력에 대한 유사도 측정

가령 매치하고자 하는 대상이 복수로 존재할 경우, 즉 하나의 서비스가 복수의 입력과 출력을 가질 경우 각각의 입력과 출력에 대한 유사도가 구해지게 된다. 따라서 이러한 경우 서비스 전체의 유사도계산은 각각의 유사도값들의 평균값으로 계산하며, 이는 다양한 입력과 출력이 존재하는 경우 입출력모두의 내용을 가장 만족시키는 서비스를 선택함을 의미한다.

4. 결론 및 향후 연구 방향

본 논문에서는 기존의 유사도 측정규칙에 대한 단점을 보안하고자, 온톨로지노드사이의 거리관계에 기반한 유사도 측정 알고리즘을 제안하였다. 실제로 이러한 노드사이의 거리관계에 대한 고려는 서비스 발견의 측면에 있어 효과적으로 적용할 수 있다. 하지만 온톨로지 노드사이의 시블링관계 및 하나이상의 상위노드를 가지는 노드에 대한 해결방법등은 제시하고 있지 않다. 따라서 이러한 문제에 대한 향후연구가 필요할 것이다.

5. 참고문헌

[1] Massimo Paolucci, Takahiro Kawamura, Terry R. Payne, and Katia Sycara. Semantic Matching of Web Services Capabilities. The First International Semantic Web Conference (ISWC), 2002.
[2] Terry R. Payne, Massimo Paolucci, Rahul Singh, and Katia Sycara. Communicating Agents in Open Multi Agent Systems. First GSFC/JPL Workshop on Radical Agent Concepts (WRAC) , 2002.