시맨틱웹 데이터의 P2P 처리를 위한 유사도 측정*

김 병 곤** · 김 연 희***

Similarity measure for P2P processing of semantic data

Kim, Byung Gon · Kim, Youn Hee

〈Abstract〉

Ontology is important role in semantic web to construct and query semantic data. Because of dynamic characteristic of ontology, P2P environment is considered for ontology processing in web environment. For efficient processing of ontology in P2P environment, clustering of peers should be considered. When new peer is added to the network, cluster allocation problem of the new peer is important for system efficiency. For clustering of peers with similar chateristics, similarlity measure method of ontology in added peer with ontologies in other clusters is needed. In this paper, we propose similarity measure techniques of ontologies for clustering of peers. Similarity measure method in this paper considered ontology's strucural characteristics like schema, class, property. Results of experiments show that ontologies of similar topics, class, property can be allocated to the same cluster.

Key Words : Ontology, P2P, Similarity measure, Semantic web, Clustering

Ⅰ. 서론

갈수록 다양하고 복잡해지는 웹 환경에서 좀더 진보된 서비스를 제공하기 위하여, 여러 회사와 기관들은 시맨틱웹 개념을 이용한 차세대 웹 환경을 구축하기 위하여 각 호스트마다 온톨로지를 구축하여 다양한 서비스를 제공하려는 노력을 하고 있다[1-4]. 하지만, 각자 구축된 온톨로지들은 동음이의어와 같이 동일한 개념에 대한 용어상의 차이 혹은 기술언어의 차이 그리고, 데이터 모델의 차이를 기계가 자동으로 이해하고 식별할 수 없기 때문에 의미적으로는 동일한 온톨로지임에도 불구하고 서로 다른 온톨로지로 인식될 수밖에 없다. 또한, 서비스를 제공하는 호스트도 물리적으로 분산되어 있고 서비스를 제공하는 회사나 기관들이 논리적으로 분산되어 있는 것이 일반적이기 때문에 이러한 환경에서의 통합 처리능력을 구성하는 것이 중요한 이슈이다. 특히, 동적인 인터넷 환경을 가장 반영하는 P2P 환경은 앞으로의 인터넷 데이터 처리에 있어서 반드시 고려하여야 하는 환경이다.

P2P 환경에서 분산되어 있는 온톨로지들을 통합하여 질의를 처리하기 위해서는 온톨로지들을 의미적으로 통합하는 통합스키마를 구성하는 방식 즉 온톨로지의 스키마를 매핑, 정렬, 합병, 통합, 결합 등을 통하여 논리적으로 통합하여 유지하는 방식의 연구가 아닌 P2P 방식의

* 이 논문은 2009년도 부천대학 교비지원 연구비에 의하여 지원된 연구의 결과임.

** 부천대학 e-비즈니스과 부교수(교신저자)
*** 부천대학 e-비즈니스과 강의전담교수

질의 처리 방식으로의 전환이 필요하다고 할 수 있다.

P2P 환경에서의 온톨로지간 메시지 교환을 통한 통합 질의처리 환경은 P2P 네트워크상에서 분산된 데이터들의 효율적인 관리와 질의 처리를 목적으로 구성된다. 그러나, P2P 네트워크 시스템에서 단순한 토폴로지를 사용하게 되면 질의가 발생하였을 때 연결되어 있는 모든 피어에 질의가 전달되어야 하며 이는 각 피어의 전송 대역폭과 처리 능력의 낭비를 초래한다. 이를 극복하기 위해서는 피어 내에서의 질의 처리 전략뿐만 아니라, 질의를 필요한 피어로 전달하기 위한 효율적인 라우팅 기법이 필요하다. 이를 위하여 피어들간의 그룹핑을 통한 클러스터링을 이용하여 라우팅을 수행하면 효율적인 질의 처리가 가능하다. 이때, 피어들을 그룹핑하기 위해서는 새로운 온톨로지를 지니는 피어가 추가 될 때 여러개의 클러스터중에서 어떤 클러스터에 배정하여야 하는가의 문제는 시스템의 효율에 중요한 영향을 미친다.

본 연구는 위에서 설명한 바와 같이 P2P 환경에서 각 피어에 온톨로지가 구축되어 있는 경우에 피어간의 클러스터링을 위해서 온톨로지간의 유사도를 측정하는 기법을 소개하고 각 클러스터내에서의 인덱스 구성과 이를 이용한 질의처리 방안을 제시한다. 유사도의 측정은 특히, 온톨로지들간의 구조적인 특성을 반영하여 클러스터링할 수 있도록 하였다.

Ⅱ. 관련연구

2.1 시멘틱웹 P2P 기술 분석

P2P 네트워크는 Napster나 Gnutella와 같은 단순 시스템에서 CAN, CHORD와 같이 분산 인덱스를 사용하는 더욱 발전된 형태의 시스템으로 진화되어 왔다[5-6]. 그러나 이러한 시스템들은 아직도 더욱 복잡한 형태의 메타데이터나 질의를 처리하는데 많은 한계점을 지니고 있다.

인터넷이 더욱 일반화되고 시멘틱웹이 현재의 인터넷 환경의 대안으로 연구되어 지고 있는 상황에서 시멘틱웹 상에서의 P2P 네트워크에 대한 연구는 더욱 중요성을 더해가고 있다. 시멘틱웹의 중요한 측면은 RDF스키마, OWL과 같은 언어들을 사용하여 웹상의 자원들에 대한 온톨로지를 구성하고, 이를 이용하여 컴퓨터시스템간의 데이터 교환이 자율적으로 발생한다는 것이다.

현재의 인터넷 환경에서 메타데이터를 지니는 온톨로지는 단지 하나의 피어의 한 서버에만 존재하는 것이 아니다. 즉, 동일한 자원에 대한 메타데이터가 여러 피어에 분산되어 저장될 수 있다. 그러나, 현재 연구되어진 P2P 시스템들은 단순한 파일이름과 같은 제한된 형태의 메타데이터를 지원하므로 좀 더 복잡한 형태의 질의를 처리하지 못하는 것이 현실이며, 이를 극복하기 위하여 스키마를 지니는 온톨로지를 적용한 P2P 네트워크 시스템의 개발이 필요하다. 또한 네트워크간의 상이함을 고려하여 디자인하여 서로 다른 네트워크간의 결합이 용이하도록 디자인 하여야 한다.

스키마기반의 P2P 네트워크 시스템의 예로 Edutella 시스템이 있다. 이 시스템은 P2P 네트워크를 통한 분산 디지털 자원들에 대한 접근을 목표로 한다. 메타데이터를 표현하기 위하여 RDF와 RDF 스키마를 사용하며, 질의어로는 RDF-QEL을 사용한다[7].

Gloserv 시스템은 OWL DL을 사용하는 P2P 전역 서비스 검색 시스템으로서 광역 네트워크와 지역 네트워크에서 모두 동작하도록 설계되었다. 넓은 영역에 걸쳐 있는 서비스들은 OWL 온톨로지를 통하여 효율적으로 표현되고, CAN P2P 네트워크에 연결된 노드들을 걸쳐서 서비스를 검색하도록 동작한다. Gloserv 서버는 서비스분류(Service classification) 온톨로지, 시소러스(Thesaurus) 온톨로지, CAN 조사테이블(look up table)로 구성된다. 서비스분류 온톨로지는 전체 서비스 목록을 바탕으로 서비스들을 계층적으로 분류한 정보를 지닌다. 시소러스 온톨로지는 서비스분류 온톨로지에 나타난 단어들에 대한 유사어 사상 정보를 지니고, 서비스가 존재하는 올바른

서버를 찾는데 정확도를 높이는데 사용된다. CAN 조사 테이블은 CAN 환경의 P2P 네트워크에서 각 서버의 관련된 클래스들을 연결시켜주는 역할을 한다[8].

이러한 연구들은 P2P 네트워크상에서 분산된 데이터들의 효율적인 관리와 질의 처리를 목적으로 구성되었다. 그러나, P2P 네트워크 시스템에서 단순한 토폴로지를 사용하게 되면 질의가 발생하였을 때 연결되어 있는 모든 피어에 질의가 전달되어야 하며 이는 각 피어의 전송 대역폭과 처리 능력의 낭비를 초래한다. 이를 극복하기 위해서는 피어 내에서의 질의 처리 전략뿐만 아니라, 질의를 필요한 피어로 전달하기 위한 효율적인 라우팅 기법이 필요하다. 라우팅을 위해서는 각 피어에 라우팅 인덱스가 필요하다. 라우팅 인덱스는 자신과 연결되어 있는 피어들에 대한 메타데이터 정보를 지니게 된다. 라우팅 인덱스의 구조는 어느 한 피어의 지역 스키마에 의존하지 않고, 네트워크 전체에서 사용되는 일정한 형태의 구조를 지니도록 설계하여야 한다.

2.2 온톨로지 유사도 측정

앞 절에서 언급한바와 같이 P2P 네트워크의 클러스터링은 좀 더 유사한 정보를 지니는 피어들을 같은 클러스터에 결합하는 것을 기본으로 한다. 이러한 방법은 무작위로 피어를 클러스터에 배정하는 것보다 네트워크내에서의 메시지의 양을 줄여준다. 따라서, 어떤 종류의 유사성 측정을 통하여 유사한 온톨로지인지를 판단하고, 클러스터를 배정할지의 문제는 전체 네트워크의 성능에 영향을 미치는 중요한 요소이다.

온톨로지의 유사도 측정에 관한 연구는 P2P 환경에서의 온톨로지 운영 측면 이전에 온톨로지 간의 상호운영성 측면이나 재사용 측면에서 활발히 연구되어 왔다. 즉, 서로 다른 구조의 온톨로지들간의 유사도 측정을 통하여 온톨로지 매핑, 정렬, 합병, 통합 등에 사용하는 방식을 사용하였다. 이는, 기구축된 온톨로지를 대상으로 하여 새로운 통합된 온톨로지를 구축하는 연구 형태로서 이를 통하여 의미적인 통합이 가능하다. 통합의 방법으로는 온톨로지간 레벨에 따른 비교 연구, 온톨로지간 통합 방법론에 대한 연구와 통합된 온톨로지의 일관성을 유지하기 위한 연구 등으로 구분할 수 있다.

온톨로지의 유사도를 측정하기 위한 방법중 가장 많이 언급된 식은 쟈카드의 계수(Jaccard's coefficient)를 이용한 방법이다. 예를 들어, 실세계에 존재하는 두개의 개념 혹은 온톨로지는 다음 네가지 확률 분포를 따른다. 즉, P(A, B), P(A′, B), P(A, B′), P(A′, B′)의 네가지 경우로 나눌 수 있으며, 여기서 A′는 A의 여집합을 의미한다. 이 네가지 확률분포를 가지고 두 개의 개념에 대한 유사도는 다음과 같이 표현된다.

$$Jaccard\text{-}sim(A, B) = P(A \cap B) / P(A \cup B)$$
$$= P(A, B) / (P(A, B) + P(A, B') + P(A', B))$$

여기서, A와 B를 온톨로지라고 가정하면, 온톨로지 A와 B가 동일한 개념일 때 Jaccard-sim(A, B)의 값은 1이며, 반대로 전혀 다른 개념의 온톨로지라면 그 값은 0이 된다. 쟈카드의 계수 방법은 GLUE 시스템과 같은 많은 시스템에서 유사도 측정을 정의하는데 사용되었다[9]. GLUE 시스템은 두개의 온톨로지간의 의미적 매핑을 찾고, 유사도를 매핑해서 하나의 새로운 온톨로지로 만드는 역할에 사용하였고, [10]에서도 유사도에 근거한 온톨로지의 매핑에 쟈카드의 계수 방법을 사용하였다. [11]에서는 유사도를 측정할 때 쟈카드의 계수를 적용하는데 이때 퍼지 이론을 적용하여 유사도의 정도를 종합하여 7개의 유사도 그룹으로 분류하여 사용하는 방법을 사용하였다.

앞에서 언급한 바와 같이 이러한 유사도의 측정은 온톨로지 간의 상호운영성 측면이나 재사용 측면에서 연구되어 왔다. 더구나, 온톨로지의 스키마가 지니는 구조적인 관계에 관하여 고려하지 않고 유사도를 산출한다. P2P 환경에서 클러스터링을 위한 유사도 산출의 경우는 온톨로지간의 매핑의 개념보다 유사한 피어들을 동일한

클러스터에 묶어주는 것이 목적이며, 이를 위하여 온톨로지간의 구조적 상하 관계를 고려하여 유사도를 측정하도록 하여야 한다.

Ⅲ. 클러스터링에 의한 온톨로지 P2P 데이터 처리

3.1 구조적 관계를 고려한 유사도 측정

P2P 환경의 웹데이터 처리에서 질의가 발생하였을 때 모든 피어에 질의를 전송하게 되면 수많은 메시지 전송량과 결과 전송으로 인하여 전체적인 네트워크 효율에 영향을 미친다. 그러므로, 온톨로지가 연결되어 있는 P2P 네트워크에서의 효율적인 질의 전송을 위하여 클러스터를 구성하고, 각 클러스터에서는 인덱스를 사용하여 자기 클러스터에 소속되어 있는 피어들을 관리하고, 클러스터에서 얻을 수 없는 정보는 클러스터간의 정보 교환을 통하여 해결하는 방법을 사용하는 것이 일반적인 데이터 처리의 흐름이다.

클러스터를 구성할 때에는 질의의 특성이나 각각의 스키마간의 관계를 고려하는 방법이 사용될 수 있다. 질의의 특성을 고려하는 방법의 경우에 이종의 P2P 네트워크에서 질의의 특성이 미리 정의 되는 것은 어렵기 때문에 빈도 측정 알고리즘과 같은 방법을 사용할 수 있다. 이를 통하여 가장 많이 사용되는 스키마, 프로퍼티, 값 등을 분석하여 질의가 전송될 클러스터를 결정하거나 두 개의 피어가 같은 클러스터에 존재하게 할지 등의 문제를 해결하는데 사용될 수 있다. 그러나 통계적인 빈도 측정은 계속적인 데이터의 축적이 필요하며 유동적인 질의 패턴에는 적용하기가 어려운 단점이 있다.

반면에, 가장 단순한 클러스터 구성 방법으로는 새로운 피어가 네트워크에 발생하면 무작위에 의하여 클러스터에 배정하는 방법이 있지만, 이러한 방법 보다는 일반적으로 P2P 네트워크의 클러스터는 비슷한 특성을 지니는 피어들을 가까운 위치에 결합하는 방법이 효율적이다. 즉, 각 피어에 존재하는 온톨로지들을 비교하여 비슷한 데이터를 지니거나, 유사한 데이터에 대한 스키마를 지니는 피어들간에 동일한 클러스터에 배정하는 방법이다. 이러한 방법은 무작위로 클러스터에 피어를 배정하는 것보다 네트워크내에서의 메시지의 양을 줄여주며, 보다 정확한 질의 처리 결과를 기대할 수 있다. 이때, 가장 중요한 요소가 온톨로지간의 유사도 측정이다. 즉, 어떤 종류의 유사도 측정을 통하여 네트워크를 분할할지가 중요한 요소이다.

본 연구에서 제안하는 클러스터 구성 방법은 온톨로지의 특성인 구조적 관계에 기반하여 유사도를 측정하고 측정 결과에 따라 클러스터를 결정하는 기법이다. 다음은 클러스터에 새로운 피어가 추가되는 경우의 유사도를 측정하기 위한 요소들을 나열한 것이다.

<표 1> 유사도 관계 측정요소

분류	내용
스키마	새로운 피어가 클러스터의 피어들과 동일한 스키마를 사용하는 경우
클래스	새로운 피어에 클러스터의 피어내에 동일하거나 동치인 클래스가 존재하는 경우
	새로운 피어에 클러스터의 피어내에 서브클래스관계의 클래스가 존재하는 경우
프로퍼티	새로운 피어에 클러스터의 피어내에 동일하거나 동치인 프로퍼티가 존재하는 경우
	새로운 피어에 클러스터의 피어내에 서브프로퍼티관계의 프로퍼티가 존재하는 경우

<표1>에서 언급한 바와 같이 유사도 측정을 위하여 크게 스키마, 클래스, 프로퍼티의 3가지 요소와 세부적으로 5개의 요소를 가지고 유사도를 측정하도록 한다. 특징적인 것은 서브클래스관계와 서브프로퍼티관계에 대한 정보도 유사도 측정을 위하여 고려한다. 본 연구에서는 식(1)에서와 같이 두개의 온톨로지 간의 유사도를 측정하기 위한 함수를 제시한다.

$$S(O_a, O_b) = a_1 \cdot S_s(O_a, O_b) + a_2 \cdot S_c(O_a, O_b) + a_3 \cdot S_p(O_a, O_b)$$
$$단,\ a_1, a_2, a_3 >= 0 \quad (1)$$

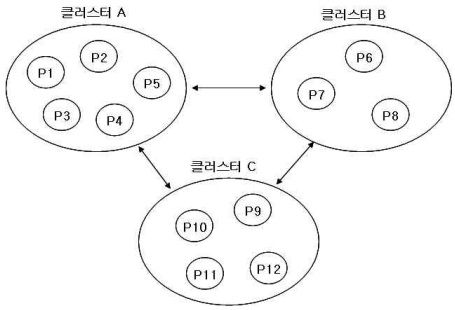
유사도 함수 S는 <표1>의 유사도 관계 측정에서 제시한 관계에 따라 두개의 온톨로지간의 유사도를 수치 값으로 표현하도록 하였다. S(Oa, Ob)는 온톨로지 Oa와 Ob간의 유사도를 나타낸다. Ss(Oa, Ob)는 두 온톨로지간의 스키마 유사도, Sc(Oa, Ob)는 두 온톨로지간의 클래스 유사도, Sp(Oa, Ob)는 두 온톨로지간의 프로퍼티 유사도이며, a1, a2, a3 는 각각의 가중치이다. Sc(Oa, Ob)와 Sp(Oa, Ob)는 관계측정요소 내용을 참조하여 다음과 같은 식에 따라 산출된다.

$$S_c(O_a, O_b) = w_1 \cdot S_c(O_a, O_b) + w_2 \cdot S_u(O_a, O_b)$$
$$S_p(O_a, O_b) = x_1 \cdot S_c(O_a, O_b) + x_2 \cdot S_u(O_a, O_b)$$
$$단,\ w_1, w_2, x_1, x_2 \geq 0 \quad (2)$$

각 가중치는 <표 1>에 나열된 요소들의 중요도에 따라 결정되며, 한 피어는 여러개의 클러스터 중에서 가장 높은 유사도가 있는 클러스터에 소속되도록 한다. 각각의 유사도는 2.2절에서 언급된 쟈카드의 계수 방법을 이용하여 측정된다.

전체 네트워크상의 피어들과의 유사도가 산출되면 클러스터의 배정 단계가 된다. 이때, 두가지의 선택적인 방법이 있다.

첫째 방법은 가장 유사한 온톨로지를 지니는 피어와 같은 클러스터에 배정하는 방법이고, 두 번째는 각 클러스터별로 산출된 유사도를 평균하여 가장 높은 평균값을 지니는 클러스터에 배정하는 방법이다. 이러한 사항은 실제 실험과 운영을 통하여 결정해야 한다. 이제, 클러스터가 결정되면, 새로이 등록하는 피어는 피어에서 사용하는 온톨로지의 스키마를 클러스터내의 피어들에게 알려 등록한다.



<그림 1> 클러스터 구성 개념

<그림 1>은 유사도 관계 측정에 의하여 구성된 클러스터의 개념을 보여준다. 모든 클러스터는 자신만의 규칙 즉, 어떤 피어들이 클러스터에 포함되어 있는지를 표현하는 정보들을 지닌다.

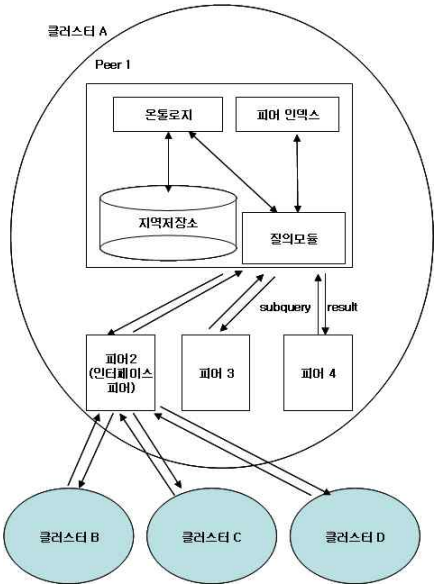
3.2 온톨로지 P2P 데이터 처리

P2P 질의 처리에서는 클러스터내에서 질의가 처리되는 경우에는 클러스터간 전송이 필요 없지만, 다른 클러스터로의 정보 전송이 요구되는 경우에 다른 클러스터로의 질의를 전송하는 것이 중요하다. 이를 위하여 인터페이스피어가 필요하다. 인터페이스 피어는 클러스터의 인터페이스 피어들을 지니는 클러스터인덱스를 지닌다.

인터페이스 피어는 클러스터에 있는 피어중에서 하나를 선정하며 일반적으로 가장 먼저 등록된 피어를 선정하는 방법이 가장 간단한 방법이 될 수 있다. 인터페이스 피어는 말 그대로 클러스터간의 인터페이스 즉 연락창구 역할을 하며, 다른 클러스터들에게 자신의 클러스터가 지닌 정보를 보여주고, 요구된 질의를 클러스터내의 적절한 피어로 보내주는 역할을 수행한다. 질의는 클러스터 인덱스를 바탕으로 이웃 클러스터의 인터페이스 피어로 전송되며 피어 인덱스에 의하여 연결된 소속 피어로 다시 전달된다. 클러스터 인덱스와 피어 인덱스의 수정은 연결된 피어들의 수정메시지에 의하여 수행된다. 이것은 하나의 피어는 임의의 하나의 클러스터에 연결되며

다음과 같은 등록 프로세스를 따른다. 클러스터에 새로운 피어가 등록되면 해당 피어의 스키마에 대한 정보를 멤버들에게 보낸다. 멤버들은 피어인덱스의 엔트리들과 내용을 비교하여 반영한다. 즉, 피어인덱스에 새로운 항목이 하나 추가되면 피어는 이와 연결된 네트워크에 이에 대한 사항을 전송하여 수정한다. 분산된 온톨로지 환경에서의 P2P 질의 처리 과정은 다음과 같다. 먼저, 한 피어의 온톨로지를 기반으로 한 지역질의가 발생하면, 지역 피어의 질의처리기는 질의를 트리플 단위로 파싱하여 클래스, 프로퍼티별로 분석하여 먼저 클러스터내의 다른 피어의 정보를 지니는 피어인덱스와 비교 검색할 준비를 한다. 파싱한 질의에 대해 피어인덱스를 사용하여 클러스터내의 클래스와 프로퍼티의 내용을 검색한다. 검색된 결과를 바탕으로 자신의 피어에서 처리가 가능한 경우에는 자체 질의 모듈로 질의를 처리하고 종료한다. 다른 피어로의 질의 전송이 요구되는 경우에는 각 피어에 전송될 별도의 질의를 산출한다. 소속된 클러스터내에서 질의의 처리가 수행되지 못한 경우에는 인터페이스 피어의 클러스터인덱스를 이용하여 외부 클러스터의 인터페이스피어로 질의가 전송된다. 각 지역 피어로 부분질의가 전송되면 지역 피어는 전송받은 질의를 실행하여 결과 데이터를 추출한다. 이때, 각 지역 피어는 질의 발생 피어에서 수행한 단계의 질의처리를 마찬가지로 수행한다. 최종적인 부분질의의 처리가 끝나면 결과 데이터는 다시 호출 피어로 전송되고, 호출피어는 전송받은 결과를 취합하여 결과를 반환한다.

<그림 2>는 질의처리개념을 구성도로 표현한 것이다. 각 피어별로 지역저장소가 존재하며 이는 XML, DB등의 각각의 구조로 데이터가 저장되어 있다. 데이터에 대한 데이터를 표현하는 온톨로지는 OWL로 구성되어 있다. 앞에서 설명한 단계의 질의처리를 위하여 질의모듈이 있으며, 질의모듈은 온톨로지와 피어인덱스를 참조하여 부분질의를 만들어 전송하고 취합하는 역할을 한다.



<그림 2> 질의처리 구성도

Ⅳ. 실험평가

본 논문에서 제안한 온톨로지 유사도 측정 함수를 평가하기 위해 인터넷 상의 여러 응용 도메인에서 사용되는 4개 그룹의 12개 OWL 문서를 대상으로 실험을 진행하였다. 실험에 사용된 OWL 문서들은 각각 가족, 책과 음악, 사람, 야구와 쇼핑에 대한 주제를 가지는 4개의 그룹으로 분류된다. 그룹별로 각 OWL 문서의 특성과 주로 사용하는 스키마는 <표 2>와 같다. 본 논문에서 제안한 유사도 측정 함수는 온톨로지를 대상으로 하기 때문에 OWL 문서에 기술되어 있는 개체(individual)에 대한 정의 부분은 실험에서 제외하였다.

본 논문에서 제안한 OWL 문서의 유사도 측정 알고리즘은 C 언어로 구현하였고 1GB RAM, 윈도우 XP 운영체제가 설치된 3.4GHZ Pentium 4 PC 환경에서 실험하였다.

본 논문에서 제안한 유사도 함수를 이용해서 OWL 문서들 간의 유사도를 측정하기 위해서는 7개 가중치의 값을 결정하는 것이 중요하기 때문에 <표 3>과 같이 각 가

<표 2> 실험 대상 OWL 문서의 특성

분류 \ 특성		주요 스키마	클래스 개수	프로퍼티 개수
그룹1	문서1	Family	14	14
	문서2	Family	9	7
	문서3	Family	12	11
그룹2	문서4	Book/Music	6	14
	문서5	Book	3	9
	문서6	Music	3	7
그룹3	문서7	People	24	10
	문서8	People	14	7
	문서9	People	17	13
그룹4	문서10	Baseball/Shopping	16	2
	문서11	Baseball	9	2
	문서12	Shopping	7	0

<표 3> 실험별 가중치 값

분류 \ 가중치	a_1	a_2	a_3	w_1	w_2	x_1	x_2
실험1	0.4	0.3	0.3	0.5	0.5	0.5	0.5
실험2	0.4	0.3	0.3	0.7	0.3	0.7	0.3
실험3	0.4	0.3	0.3	0.3	0.7	0.3	0.7
실험4	0.2	0.5	0.3	0.5	0.5	0.5	0.5
실험5	0.2	0.5	0.3	0.3	0.7	0.3	0.7
실험6	0.2	0.5	0.3	0.7	0.3	0.7	0.3
실험7	0.2	0.3	0.5	0.5	0.5	0.5	0.5
실험8	0.2	0.3	0.5	0.3	0.7	0.3	0.7
실험9	0.2	0.3	0.5	0.7	0.3	0.7	0.3

중치를 변화시켜 실험을 진행하였다.

<표 2>에서 제시한 4개의 그룹으로 분류되는 12개의 OWL 문서에 대해 <표 3>에서 제시한 가중치로 실험을 진행한 결과 모든 실험에서 동일 그룹에 속하는 문서들 간의 유사도가 높게 측정된 것을 확인할 수 있었다. 특히, 실험2와 실험6에서 다른 실험들보다 더 높은 유사도가 측정되었다.

<표 4>는 <표 3>에서 제시된 실험2의 환경에서 12개 문서들 간의 유사도를 측정한 결과를 보여준다. 같은 문서에 대한 유사도 측정 결과는 생략하였다. <표 4>에서 같은 그룹에 속해있는 문서들 간의 유사도가 상대적으로 높게 측정된 것을 확인할 수 있다. 예를 들어, 가족(Family)과 관련한 용어를 정의하고 용어들의 의미적 관계를 기술하는 온톨로지인 문서1은 6.23의 유사도가 측정된 문서 3과 가장 유사하다고 판단할 수 있다. 문서3은 문서1과 같은 그룹에 속한다. 그리고 두 번째로 높은 5.27의 유사도가 측정된 문서2도 문서1과 같은 그룹에 속한다. 문서1의 경우 다른 그룹에 속하는 문서들에 비해 그룹3에 속하는 문서들과는 비교적 높은 유사도가 측정된 것을 확인할 수 있다. 이것은 그룹3에 속하는 문서들이 사람(People)에 관련한 용어를 정의하고 용어들 간의 의미적 관계를 기술하는 온톨로지로 그룹1의 가족 온톨로지와 일치하는 용어들이 많이 존재하기 때문이다.

<표 4> 실험2의 유사도 측정 결과

(a) 문서1부터 문서 6까지의 유사도 측정 결과

	문서1	문서2	문서3	문서4	문서5	문서6
문서1	-	5.72	6.23	1.6	1.6	1.6
문서2	5.27	-	5.39	1.6	1.6	1.6
문서3	6.23	5.39	-	1.6	1.6	1.6
문서4	1.6	1.6	1.6	-	4.73	4.31
문서5	1.6	1.6	1.6	4.73	-	2.23
문서6	1.6	1.6	1.6	4.31	2.23	-
문서7	2.71	2.71	2.83	1.6	1.6	1.6
문서8	2.71	2.71	2.83	1.6	1.6	1.6
문서9	2.41	2.41	2.5	1.81	1.81	1.6
문서10	1.6	1.6	1.6	1.6	1.6	1.6
문서11	1.6	1.6	1.6	1.6	1.6	1.6
문서12	1.6	1.6	1.6	1.6	1.6	1.6

(b) 문서7부터 문서 12까지의 유사도 측정 결과

	문서7	문서8	문서9	문서10	문서11	문서12
문서1	2.71	2.71	2.41	1.6	1.6	1.6
문서2	2.71	2.71	2.41	1.6	1.6	1.6
문서3	2.83	2.83	2.5	1.6	1.6	1.6
문서4	1.6	1.6	1.81	1.6	1.6	1.6
문서5	1.6	1.6	1.81	1.6	1.6	1.6
문서6	1.6	1.6	1.6	1.6	1.6	1.6
문서7	-	6.81	7.56	1.6	1.6	1.6
문서8	6.81	-	6.54	1.6	1.6	1.6
문서9	7.56	6.54	-	1.6	1.6	1.6
문서10	1.6	1.6	1.6	-	4.71	3.87
문서11	1.6	1.6	1.6	4.71	-	2
문서12	1.6	1.6	1.6	3.87	2	-

<표 5> 실험6의 유사도 측정 결과

(a) 문서1부터 문서 6까지의 유사도 측정 결과

	문서1	문서2	문서3	문서4	문서5	문서6
문서1	-	5.98	6.49	0.8	0.8	0.8
문서2	5.98	-	5.71	0.8	0.8	0.8
문서3	6.49	5.71	-	0.8	0.8	0.8
문서4	0.8	0.8	0.8	-	4.15	3.73
문서5	0.8	0.8	0.8	4.15	-	1.43
문서6	0.8	0.8	0.8	3.73	1.43	-
문서7	2.65	2.65	2.85	0.8	0.8	0.8
문서8	2.65	2.65	2.85	0.8	0.8	0.8
문서9	2.15	2.15	2.3	1.15	1.15	0.8
문서10	0.8	0.8	0.8	0.8	0.8	0.8
문서11	0.8	0.8	0.8	0.8	0.8	0.8
문서12	0.8	0.8	0.8	0.8	0.8	0.8

(b) 문서7부터 문서 12까지의 유사도 측정 결과

	문서7	문서8	문서9	문서10	문서11	문서12
문서1	2.65	2.65	2.15	0.8	0.8	0.8
문서2	2.65	2.65	2.15	0.8	0.8	0.8
문서3	2.85	2.85	2.3	0.8	0.8	0.8
문서4	0.8	0.8	1.15	0.8	0.8	0.8
문서5	0.8	0.8	1.15	0.8	0.8	0.8
문서6	0.8	0.8	0.8	0.8	0.8	0.8
문서7	-	7.57	8.5	0.8	0.8	0.8
문서8	7.57	-	7.2	0.8	0.8	0.8
문서9	8.5	7.2	-	0.8	0.8	0.8
문서10	0.8	0.8	0.8	-	4.77	3.65
문서11	0.8	0.8	0.8	4.77	-	1
문서12	0.8	0.8	0.8	3.65	1	-

<표 5>는 <표 3>에서 제시된 실험6의 환경에서 12개 문서들 간의 유사도를 측정한 결과를 보여준다. <표 4>와 마찬가지로 <표 5>에서도 같은 그룹에 속해있는 문서들 간의 유사도가 상대적으로 높게 측정된 것을 확인할 수 있다. 한편, 같은 그룹에 속하는 문서들 간의 유사도는 <표 5>의 유사도가 <표 4>의 유사도보다 높게 측정되었다. 예를 들어, 같은 그룹에 속하는 문서1과 문서3의 유사도는 <표 4>에서는 6.23이지만 <표 5>에서는 6.49이다. 이것은 동일한 온톨로지를 기술하고 있는 같은 그룹에 속하는 문서들은 특히 일치하는 용어, 즉 클래스가 많이 존재하기 때문에 스키마나 프로퍼티에 대한 가중치보다 클래스에 대한 가중치를 0.5로 더 크게 설정한 실험 6에서 높은 유사도가 측정된 것이다. 그리고 같은 그룹에 속하는 온톨로지 문서들에는 서브 클래스 관계에 있는 클래스보다는 정확하게 일치하는 같은 클래스가 사용되는 경우가 많기 때문에 서브 클래스 관계에 대한 가중치보다 동일한 클래스에 대한 가중치를 0.7로 높게 설정한 실험6에서 실험4나 실험5보다 더 높은 유사도가 측정되었다. 또한, 의미적으로 일치하는 용어들이 많이 존재하지만 다른 그룹에 속하는 문서들 간의 유사도는 <표 4>에서 더 높게 측정된 것을 확인할 수 있다. 예를 들어, 그룹1에 속하는 문서1과 그룹3에 속하는 문서7의 유사도는 <표 5>에서는 2.65이지만 <표 4>에서는 2.71이다. 이것은 다른 그룹에 속하는 온톨로지 문서들은 일치하는 클래스가 존재하기는 하지만 같은 그룹의 문서들에 비해서는 그 수가 적기 때문에 클래스나 프로퍼티에 대한 가중치보다는 스키마에 대한 가중치를 0.4로 높게 설정하여 잠재적인 의미적 유사도를 측정했기 때문이다.

실험결과를 토대로 동일한 주제 혹은 유사한 클래스 등을 지니는 온톨로지 간에 동일한 클러스터의 배정이 가능한 것을 알 수 있다.

Ⅳ. 결론

P2P 환경에서 효율적인 질의처리를 위하여 유사한 온톨로지를 지니는 피어들을 클러스터 형식으로 구분하여 처리하는 방법을 사용한다.

본 연구는 위에서 설명한 바와 같이 P2P 환경에서 각 피어에 온톨로지가 구축되어 있는 경우에 피어간의 클러스터링을 위해서 온톨로지간의 유사도를 측정하는 기법을 소개하였다. 유사도 측정을 위하여 크게 스키마, 클래스, 프로퍼티의 3가지 요소와 세부적으로 5개의 요소를 가지고 유사도를 측정하는 유사도 함수를 제시하였다. 제시된 함수를 가지고 OWL 온톨로지 문서간의 유사도

를 측정하여 문서들간의 유사도를 분석을 통하여 새로운 온톨로지의 클러스터 배정에 사용할 수 있도록 하였다. 측정된 유사도 값은 가중치의 배정에 따라 조금씩 상이한 값을 나타내었으며, 동일한 주제의 온톨로지간의 동일한 클러스터 배정이 가능하였다.

본 연구에서 제시한 방법으로 클러스터를 배정하고 질의를 처리하는 경우에, 유사한 정보들이 동일한 클러스터로 묶이게 되며, 특히 빈번히 발생하는 유사 정보 질의를 처리하는데 주요 성능 향상 요인으로 작용할 수 있다.

참고문헌

[1] www.w3.org/TR/2004/REC-owl-guide-20040210/wine.rdf

[2] http://www.movieontology.org/

[3] http://www.schemaweb.info/

[4] 조정원, 차시호, 안병호, 조국현, “홈 헬스케어를 위한 온톨로지 기반 상황인지 플랫폼의 설계 및 구현,” 디지털산업정보학회 논문지, 제5권 제3호, 2009, pp77-86.

[5] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker., “A scalable content addressable network,” Proceedings of the 2001 Conference on applications, technologies, architectures, and protocols for computer communications. ACM Press New York, NY, USA, 2001, pp161-172

[6] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. “Chord: A scalable peer-to-peer lookup service for internet applications,” Proceedings of the 2001 Conference on applications, technologies, architectures, and protocols for computer communications. ACM Press New York, NY, USA, 2001.

[7] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmer, and T. Risch, “EDUTELLA: a P2P Networking Infrastructure based on RDF,” WWW 11 Conference Proceedings, Hawaii, USA, 2002.

[8] K. Arabshian, H. Schulzrinne, D. Trossen, and D. Pavel, “Gloserv: Global service discovery using the owl web ontology language,” Proceedings of the IEE International Workshop on Intelligent Environments, University of Essex, Colchester, UK, June 2005.

[9] Doan, J. Madhavan, P. Domingos and A. Halevy, “Learning to map between ontologies on the semantic web,” Proceedings of the 11th International WWW Conference, 2002.

[10] Laurel C. Y. Kong, C. L. Wang, and F. C. M. Lau, “Ontology Mapping in Pervasive Computing,” Proceedings of Environment International Conference on Embedded and Ubiquitous Computing., Aizu, Japan, August, 2004, pp. 1014-1023.

[11] Suphakit Niwattanakul, Philippe Martin, Michel Eboueya and Kanit Khaimook, “Ontology Mapping based on Similarity Measure and Fuzzy Logic,” Proceedings of E-learn 2007, Quebec City, Canada, October, 2007.

■ 저자소개 ■

김 병 곤
Kim, Byung Gon

2001년~현재 부천대학 e-비즈니스과 부교수
2001년 홍익대학교 전자계산학과 이학박사
1992년~1998년 국방과학연구소 연구원
1992년 홍익대학교 전자계산학과 이학석사
1990년 홍익대학교 전자계산학과 이학사

관심분야 : 다차원 인덱싱, 온톨로지, 시맨틱 웹 등
E-mail : bgkim@bc.ac.kr

김 연 희
Kim, Youn Hee

2007년 3월~현재 부천대학 e-비즈니스과 강의전담교수
2006년 8월 홍익대학교 컴퓨터공학과 (공학박사)
2002년 2월 홍익대학교 컴퓨터공학과 (공학석사)
2000년 2월 홍익대학교 컴퓨터공학과(공학사)

관심분야 : 시맨틱 웹, XML, 분산 데이터베이스, 모바일 데이터베이스
E-mail : yhkim@bc.ac.kr

논문접수일 : 2010년 9월 28일
수 정 일 : 2010년 10월 29일
게재확정일 : 2010년 11월 10일