의미 기반의 정보 검색을 위한 P2P 시스템

김성수 [1], John Strassner[2], 홍원기 [2]

[1] 포항공과대학교 컴퓨터공학과

[2] 포항공과대학교 정보전자융합공학부

{kiss, johns, jwkhong}@postech.ac.kr

요 약

P2P(Peer-to-Peer) 시스템은 이용자들의 효율적인 데이터 공유를 위한 대표적인 분산 시스템이다. 데이터가 중앙 서버에 집중되어있는 서버-클라이언트 모델과는 달리 P2P 시스템은 전체 시스템에 데이터가 분산되어있기 때문에 효율적인 쿼리의 라우팅과 검색이 중요한 이슈이다. 기존의 DHT(Distributed Hash Table) 방식의 P2P 시스템들은 단순 쿼리에 이용되는 단일 키워드 매칭을 통한 검색을 제공하지만 실제 사용자들은 범위 검색, 분류 검색 등의 확장된 검색기능을 필요로 한다. 본 논문에서는 단순한 키워드 매칭의 검색기법을 개선하여 쿼리와 노드의 컨텐츠 유사도를 이용한 DHT 기반의 검색 기법을 제안한다. 제안한 기법은 Kademlia 프로토콜을 기반으로 설계되었으며 시뮬레이션을 통해 검색 성능을 검증한다.

1 서론

Napster[1], eMule[2], Gnutella[3] 와 같은 응용프로그램의 성공으로 P2P (Peer-to-Peer) 시스템은 이용자들간의 데이터 공유를 위한 대표적인 분산 시스템으로 널리 알려졌다. P2P 시스템은 시스템의 노드들이 클라이언트와 서버의 역할을 동시에 수행할 수 있으므로 기존의 서버-클라이언트 모델과 달리 서버의 병목현상과 중앙서버의 부하를 획기적으로 줄일 수 있다.

대부분의 P2P 시스템은 단순한 키워드 매칭을 이용해 정보를 검색한다. 하지만 실제 사용자들은 범위 검색, 분류 검색 등의 복잡한 검색을 요구하며 키워드 매칭은 복잡한 쿼리나 단어 및 구문의 의미를 반영하는 검색을 지원하지 않는다. 비구조적 P2P 시스템은 브로드캐스팅을 이용하여 쿼리를 전체 노드에 전달하기 때문에 복잡한 쿼리를 수행할 수 있지만, 브로드캐스팅에 따른 과다한 트래픽 생성으로 네트워크에 과도한 부하를 유발한다. 구조적인 P2P 시스템인 CAN[4], Chord[5], Kademlia[6] 등은 DHT(Distributed Hash Table)를 이용하기 때문에 키워드 매칭에 있어서는 좋은 성능을 보이나 구조상의 한계로 복잡한 쿼리를 처리할 수 없다.

본 논문에서는 키워드 매칭과 더불어 컨텐츠의 내용에 기반한 검색을 지원하는 시멘틱 오버레이 네트워크를 제안한다. 시멘틱 오버레이 네트워크의 설계를 위해서는 각 노드 그리고 쿼리들의 시멘틱과 오버레이 네트워크 상에서의 효율적인 쿼리 라우팅이 중요한 문제이다. 제안하는 시멘틱 오버레이 네트워크는 노드의 컨텐츠와 쿼리의 시멘틱을 VSM(Vector Space Model)[8]을 이용해 나타내며 효율적인 라우팅을 위해 DHT 상위에 시멘틱 클러스터를(semantic cluster) 구성한다.

본 논문은 다음과 같이 구성된다. 2 장에서는 구조적 P2P 와 고급 검색과 관련된 연구에 대해 살펴볼 것이다. 3 장에서는 본 논문에서 제안하는 시멘틱 오버레이 네트워크의 구조와 동작, 쿼리 라우팅에 대해 알아본다. 4 장에서는 시뮬레이션과 그 결과에 대해 설명한다. 마지막으로 5 장에서는 결론 및 향후 연구에 대해 서술한다.

2 관련 연구

이 장에서는 구조적 P2P 네트워크와 시멘틱 오버레이 네트워크와 관련된 연구에 대해 설명한다.

2.1 구조적 P2P 시스템

DHT (Distributed Hash Table)는 네트워크 환경에 위치한 노드들에 해쉬 테이블을 분산하여 저장한다. DHT 를 사용함으로써 효율적인 검색 및 쿼리의 라우팅이 가능하다. DHT 에 키워드를 입력하면 P2P 시스템상에 존재하며 해당 키워드에 해당하는 파일이나 문서를 가지고 있는 노드의 위치를 반환한다. Chord[5]는 대표적인 DHT 기반의 P2P 프로토콜로 SHA-1 해쉬함수를 이용해 각 노드에 m-bit 식별자를 할당한다. 데이터 오브젝트의 키워드 해쉬값에 따라 해당 키워드를 포함하는 데이터 오브젝트가 어떤 노드에 위치할 것인지를 결정한다. 키워드에 따른 오버레이 네트워크 상의 위치를 지정할 때에는 식별자 값이 키워드 해쉬값과 같거나 더 큰 첫 번째 노드(successor)에 저장한다. Finger table 이라는 라우팅 테이블을 이용해 해쉬값을 삽입하거나 해당 키워드를 가진 노드를 찾기 위한 메시지를 전달하는데 이용한다. Finger table 는 contact 들로 구성되는데 네트워크 주소정보와 포트번호, ID 도 같이 저장

되므로 직접 통신이 가능다. 노드 ID 공간에서 노드의 위치를 기준으로 하여 시계 방향으로 거리를 지수적으로 증가시켜가면서 $p+2^n$ 에 해당하는 식별자 값을 가진 노드들에 대한 정보만 유지한다.

2.2 컨텐츠 기반의 검색을 지원하는 P2P 시스템

P2P 시스템에 키워드 매칭 이외의 확장된 검색 기능을 도입하는 연구는 논문 [7]에서부터 시작되었다. P2P 검색의 초기 연구로 노드와 데이터 오브젝트를 MUSIC, MOVIE, BOOK 등으로 분류하여 분류 검색을 가능하게 하였으나 적용할 수 있는 도메인의 범위가 한정되며 각 데이터 오브젝트에 수동으로 분류를 위한 메타데이터를 설정해야 하는 단점이 있다.

PSearch[9]와 SSW(Semantic Small World)[10]는 컨텐츠 기반의 검색을 지원하는 P2P 시스템이다. PSearch 의 경우 DHT 중 하나인 CAN[4] 위에 시멘틱 기반의 검색 엔진을 구현했다. 시멘틱 검색은 검색을 통해 쿼리와 내용이 가장 밀접한 문서를 반환한다. SSW 는 pSearch 의 검색과 유지에 드는 오버헤드를 줄이기 위해 컨텐츠의 내용이 유사한 노드들을 클러스터로 묶는 개념을 도입했다. SSW 는 pSearch 에 비해 우수한 성능을 보이나 여전히 검색에 필요한 경로 길이가 전체 노드 수가 1024 개일때 15hop 으로 비구조적 P2P 와 비교해 낮은 검색 성능을 보인다.

3 시멘틱 오버레이 네트워크

이 장에서는 제안하는 시멘틱 오버레이 네트워크[13]의 구조와 쿼리의 라우팅에 대해 설명한다. 또한 포함하고 있는 컨텐츠의 시멘틱이 비슷한 노드들간의 클러스터를 구성하는 법에 대해 설명한다.

대부분의 P2P 시스템은 키워드 매칭을 이용한 검색만을 지원한다. 즉, 키워드의 완전 일치 또는 부분 일치를 통해 정보를 검색한다. 하지만 실제 사용자들은 키워드 검색뿐만 아니라 범위 검색, 분류 검색 등 다양한 검색을 필요로 한다. 구조적 P2P 시스템은 키워드 매칭을 이용한 검색에 최적화된 시스템이지만 범위 검색, 분류 검색 등의 복잡한 검색의 수행이 불가능하다. 구조적 P2P 의 노드간 연결은 노드들이 소유한 컨텐츠들과는 무관하며 임의로 형성되기 때문이다. 따라서 쿼리의 도메인이 오버레이 네트워크 전체이며 쿼리가 전체 네트워크에 동시에 발생할 경우 네트워크 전체에 과도한 부하를 유발한다.

제안하는 오버레이 네크워크는 키워드 매칭 기반의 기본적인 검색기능과 더불어 쿼리와 노드의 컨텐츠 유사도를 검색하는 고급 검색기능을 제공한다. 노드가 포함하고 있는 컨텐츠들은 VSM 형식으로 표현하며 이를 시멘틱 벡터(Semantic Vector)라 한다. 각 노드는 $t \times d$ term-document 행렬로 포함하고 있는 컨텐츠를 나타낸다. d 는 문서의 수를, t 는 각 문서에 나타나는 용어의 개수를 의미한다. 두 문서간 혹은 노드간의 컨텐츠 유사도는 코사인 유사도(cosine similarity)를 이용해 측정된다. 코사인 유사도는 문서와 쿼리의 유사도를 측정하기 위해 주로 쓰이는 기법이다. 용어 벡터 X 는 $|X|=1$ 로 표준화되어 용어 벡터 $X=(x1, x2, ..,xl)$과 $Y=(y1, y2, …, yl)$의 유사도가 식 2 의 공식을 이용해 계산된다. $\cos(X, Y)$ 는 벡터 X 와 Y 가 이루는 각도의 코사인 값이다.

$$\cos(X,Y) = \frac{X \otimes Y}{|X| \bullet |Y|} = \sum_{i=1}^{l} x_i y_i \qquad (2)$$

그림 2 는 시멘틱 오버레이 네트워크를 나타내다. 기본 오버레이 네트워크는 Kademlia DHT 를 이용해 구성하고 상위레벨에서 시멘틱 클러스터를 생성한다. 시멘틱 클러스터는 명시적으로 정의되지 않고 각 노드별로 컨텐츠가 유사한 상위 20 개의 노드 주소를 가지는 방식으로 구성된다.

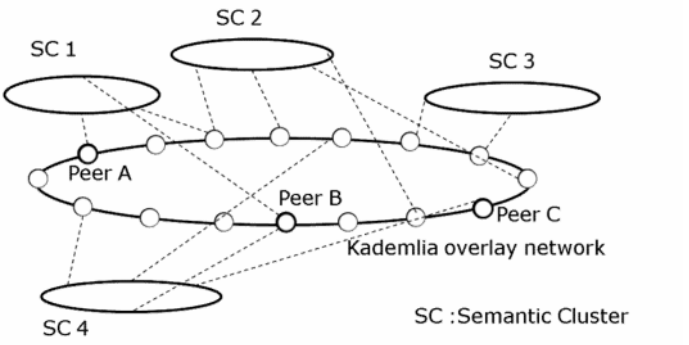

그림 1 시멘틱 오버레이 네트워크

예를 들어, 그림 1 의 노드 A 가 노드 C 가 가지고 있을 데이터를 찾기 원할 경우, 기존의 DHT 의 검색기능으로는 정확한 키워드를 모를 경우 찾는 것이 불가능하다. 본 논문에서 제안하는 시멘틱 오버레이 네트워크는 시멘틱 클러스터를 이용해 노드 A 가 클러스터내의 쿼리와 가장 유사도가 높은 노드 B 에게 쿼리를 전달하고, 노드 B 가 클러스터 내부의 노드들을 검색해 최종적으로 노드 C 를 찾아낸다.

시멘틱 오버레이 네트워크를 구성하는 모든 노드들은 Kademlia 프로토콜을 구현하고 추가로 노드가 포함하고 있는 파일 및 문서를 나타내는 시멘틱 벡터를 가진다. 컨텐츠 유사도에 기반한 검색을 위해 자신의 시멘틱 벡터뿐만 아니라 자신과 유사한 컨텐츠를 가지는 다른 노드들을 시멘틱 contact 라 하며 시멘틱 contact 의 주소와 시멘틱 벡터들의 목록을 저장함으로써 시멘틱 라우팅 테이블을 구성한다. 코사인 유사도를 바탕으로 각 노드들은 시멘틱 contact 들의 정보를 수집하며 유사도가 높은 contact 들의 정보를 시멘틱 라우팅 테이블에 저장한다. 그림 1 은 시멘틱 라우팅 테이블을 나타낸다. 노드와의 컨텐츠 유사도와 노드의 주소, UDP 포트 번호, 노드 식별자, 그리고 시멘틱 벡터를 포함한다. 시멘틱 contact 목록은 유사도가 큰 순에서 작은 순으로 정

렬되며 테이블 내의 노드보다 유사도가 높은 노드가 발견되면 가장 유사도가 낮은 노드를 테이블에서 삭제하고 새로운 노드를 테이블에 추가하여 정렬한다.

Index	Distance	Contact address
0	Semantic distance	(IP address, UDP port, Node ID, Semantic Vector)
1	Semantic distance	(IP address, UDP port, Node ID, Semantic Vector)
…	…	…
I	Semantic distance	(IP address, UDP port, Node ID, Semantic Vector)

그림 2 시멘틱 라우팅 테이블

알고리즘 1 Peer Joining

Peer *i* joining in semantic overlay network
1: Extract local semantic vector matrix V_i
2: Get Kademlia contact list from bootstrap peer
3: **for** peer ***j*** in set of Kademlia neighbors **do**
4: Get Kademlia neighbor ***j***'s semantic neighbors and semantic vectors
5: **for** peer ***m*** in the set of semantic neighbors of peer ***j***
6: **if** semantic neighbor is in semantic routing table of peer ***i***
7: **then** put semantic neighbor into semantic routing table of ***i***
8: **else**
9: drop the neighbor contact
10: **end if**
11: **end for**
12: **end for**

새로운 노드가 시멘틱 오버레이에 join 할 경우 알고리즘 1 의 과정을 통해 시멘틱 네이버들의 정보를 수집한다. 노드 *i* 가 오버레이 네트워크에 합류했을때, *i* 는 자신의 컨텐츠를 반영한 시멘틱 벡터 V_i 를 계산한다. 그리고 부트스트랩 노드로부터 Kademlia contact 의 목록을 가져온다. 노드 *i* 는 Kademlia contact 노드들로부터 각 contact 들의 시멘틱 contact 목록과 시멘틱 벡터를 가져와 자신의 시멘틱 벡터와 비교한 후 유사도가 높은 상위 20 개의 contact 들을 시멘틱 라우팅 테이블에 저장한다.

알고리즘 2 Semantic Searching (complex query)

Semantic searching of query *Q* (***K*** is the default value for choosing similar neighbors list, ***t*** is the threshold for document similarity with the query)
1: Complex query ***Q*** is received by peer ***i***
2: Peer ***i*** calculates the semantic similarity between query ***Q*** and its semantic neighbors *j*
3: Peer ***i*** chooses ***K*** semantic neighbors that have the shortest semantic distance to query ***Q***
4: **for** node *s* in the set of ***K*** semantic neighbors nodes
5: **do** Send query ***Q*** to node *s*
6: **if** node *s* found documents whose semantic similarity is exceeds threshold ***t***
7: **then** return document ID to peer ***i***
8: **end if**
9: **end for**

복잡한 쿼리를 시스템에 질의했을 때 DHT 기반의 P2P 시스템은 쿼리의 키워드를 전부, 혹은 부분적으로 만족할 때만 결과값을 반환한다. 이는 이용자가 검색하고자 하는 데이터 오브젝트의 이름과 키워드를 정확히 알아야 함을 의미한다. 또한 DHT 는 contact 들간의 관계가 임의로 형성되므로 범위검색, 분류검색 등의 복잡한 검색을 수행할 수 없다. 시멘틱 오버레이의 contact 들은 서로 유사한 컨텐츠들을 보유하고 있는 노드들이기 때문에 고급 쿼리들의 브로드캐스팅 범위를 한정할 수 있으며 쿼리와의 유사도를 기반으로 결과를 반환하기 때문에 내용적으로 유사한 검색 결과를 얻을 수 있다. 알고리즘 2 는 시멘틱 오버레이 네트워크내에서 쿼리가 라우팅 되는 과정을 설명한다. 노드 *i* 가 쿼리 *Q* 를 시스템에 질의했을 때, 시멘틱 contact 들과 쿼리의 코사인 유사도를 계산해 유사도가 가장 높은 *K* 개의 시멘틱 contact 로 쿼리를 전달한다. 쿼리를 전달받은 노드 *s* 는 쿼리와 노드 *s* 의 코사인 유사도를 검사하여 유사도가 주어진 임계값 *t* 보다 높으면 해당 노드의 컨텐츠를 결과값으로 반환하며 그렇지 않은 경우 가장 높은 유사도를 가진 노드 *K* 개의 쿼리를 전달한다.

4 성능 평가

이 장에서는 시멘틱 오버레이 네트워크를 실험을 통해 검증한다. 시멘틱 오버레이 네트워크의 시뮬레이션은 P2P 시뮬레이터 PeerSim 상에서 수행되었다. 임의로 생성된 토폴로지 상의 노드들에 160bit 의 ID 와 시멘틱 벡터 값을 임의로 할당한다. 표 1 은 시뮬레이션에 필요한 각 변수의 설명과 실험에 사용된 설정 값을 나타낸다. 시멘틱 벡터는 50 개의 차원으로 구성되며 노드의 수는 200 개에서 12800 개까지 각 단계별로 노드의 수를 두 배로 증가시켜가며 실험한다. 시멘틱 contact 의 수는 20 개로 제한했으며 매 시뮬레이션단계마다 새로운 노드가 오버레이 네트워크에 추가되거나 네트워크를 떠날 확률은 50%이다. 매 초 마다 임의로 네트워크 상의 노드를 선정하고 해당 노드에서 쿼리를 생성한다. 쿼리는 시멘틱 클러스터 내의 노드 중에 가장 컨텐츠 유사도가 높은 노드로 포워딩되고 임계값(threshold) 이상의 유사도를 보이면 쿼리의 전달을 중지하고 결과값을 반환한다. 임계값은 0.5-0.85 사이의 값으로 설정한다. 시멘틱 검색의 성능평가를 위해 결과를 찾는데 필요한 경로의 길이와 지연시간을 먼저 분석하였다.

표 1 시뮬레이션 설정 값

	설명	값
N	Number of nodes in the network	200-12800
S	Number of semantic contacts in node	20
p	Probability that one new node joins or an existing node leaves for every simulation step (1/100sec)	50%
t	Threshold for query	0.5-0.85
k	Dimension of VSM space	50
K	Query is sent to K semantically closest nodes	3-5

그림 3 은 임의로 생성된 쿼리 *Q* 와 노드들 사이의 컨텐츠 유사도를 나타낸다. 0.5 에서 0.6 사이의 유사도를 갖는 노드들이 전체의 34%를 차지했으며 0.8 이상의 유사도를 보이는 노드는 전체의 10%에 불과했다. 실험에서 0.8 이상의 임계값을 적용하여 쿼리와 0.8 이상의 유사도를 가지는 노드를 반환하도록 설정하였다. 시멘틱 검색의 성능평가를 위해 결과를 찾는데 필요한 경로 길이와 지연시간을 먼저 분석하였다.

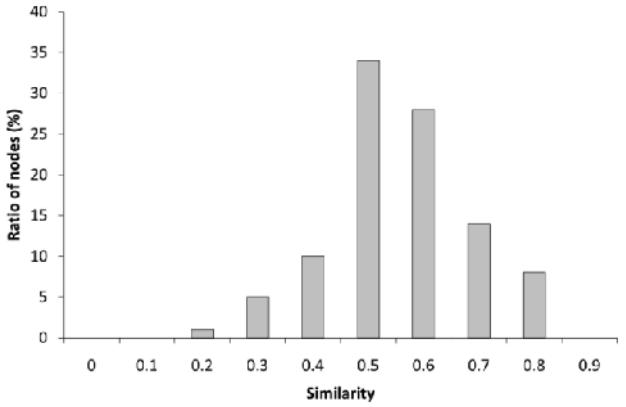

그림 3 노드 유사도 분포

그림 4 는 평균 임계값이 0.8 일 때 검색에 필요한 평균 경로 길이와 지연시간을 나타낸다. 그림 1 에서와 같이 쿼리와의 컨텐츠 유사도가 0.8 이상인 노드들은 전체의 10%인데 이 노드들을 검색하기 위한 경로 길이 평균값이 전체 노드 수가 12800 개일 경우에도 2.5 홉을 넘지 않는다. 지연시간은 800ms 이하의 값을 보이며 경로 길이와 비례한다. 지연시간 및 경로의 길이는 전체 노드수의 증가함에 따라 선형적인 증가를 보인다.

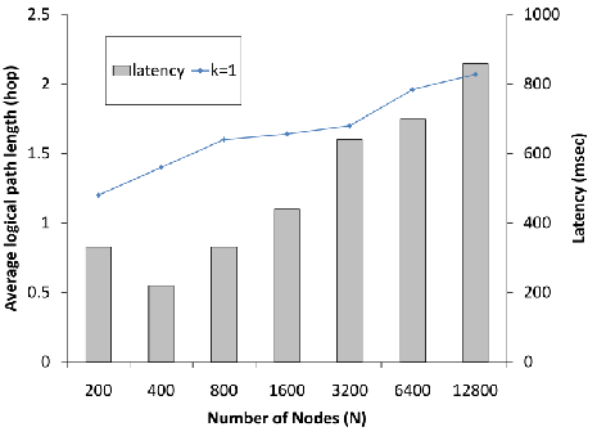

그림 4 검색에 필요한 평균 경로길이와 지연시간 (*t*=0.8)

그림 5 는 유사도가 0.85 이상인 노드들을 검색하는데 필요한 평균 경로 길이와 생성되는 메시지의 수를 나타낸다. 전체 노드중 쿼리와의 유사도가 0.85 이상인 노드는 약 0.01%밖에 존재하지 않는다. 즉 800 개의 노드가 오버레이 네트워크를 구성하고 있다면 약 8 개의 노드가 0.85 이상의 유사도를 보인다는 의미이다. K=1 일 때, 평균 경로길이와 생성되는 메시지의 수는 오버레이 네트워크의 전체 노드 수에 비례한다. K=2 일 때, 검색에 필요한 평균 경로 길이는 노드의 수에 비례하여 증가하나 생성되는 메시지의 수는 노드의 수가 증가함에 따라 기하급수적으로 증가한다. 그러나 그림에서 보이듯 0.85 이상의 유사도를 가지는 노드를 검색하는 경우에도 평균 경로 길이는 6 홉을 넘지 않는다. 특이한 점은 k 값이 클 때, 평균 경로 길이 역시 비례해서 증가한다는 점인데 이는 k 값이 클 경우 쿼리와 가장 유사한 노드뿐만 아니라 2 번째, 3 번째로 유사한 노드들에도 쿼리가 전달됨으로써 검색 경로가 길어지는 것으로 보인다.

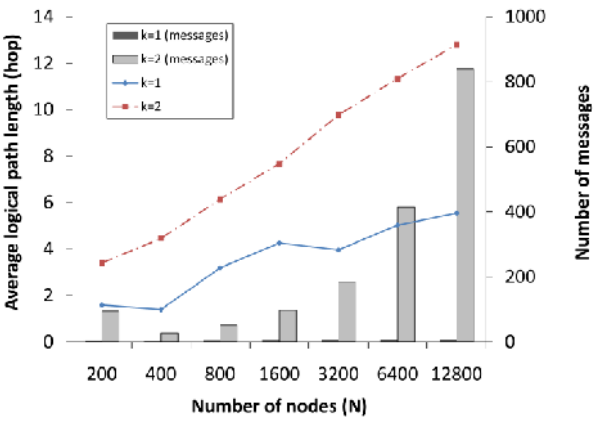

그림 5 검색에 필요한 평균 경로길이와 메시지의 수 (*t*=0.85)

5 결론

본 논문에서는 컨텐츠의 유사도 검색을 지원하는 시멘틱 오버레이 네트워크를 제안하였다. 각 노드가 가지고 있는 컨텐츠를 VSM 을 이용하여 표현하고 쿼리와 각 노드의 컨텐츠 유사도를 코사인 유사도를 이용해 비교하여 임계값 이상의 유사도를 가지는 노드의 검색을 가능하게 하였다. 실험을 통해 약 10000 개 이상의 노드를 가지는 오버레이 네트워크 상의검색에 서도 2.1 홉 안에 결과를 가져올 수 있다는 결론을 확인하였다.

향후 연구로는 코사인 유사도 이외의 다양한 문서 유사도 측정 기법의 적용과 검색을 통해 얻어진 문서와 쿼리와의 유사도 검증이 이루어질 것이다.

6 참고 문헌

[1] Napster, http://napster.com/.
[2] eMule. http://www.emule-project.net. 2006.
[3] Gnutella, http://gnutella.wego.com/.
[4] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, “A scalable content-addressable network”, In ACM SIGCOMM , Aug. 2001, pp. 161-172.
[5] I. Stoica, R. Morris, D. Karger,M. F. Kaashoek, and H. Balakrishnan, “Chord: A scalable peer-to-peer lookup service for Internet applications”, in Proceedings of ACM SIGCOMM, Aug. 2001, pp. 149–160.
[6] P. Maymounkov and D. Mazieres, “Kademlia: A Peer-to-Peer Information System Based on the XOR Metric”, 1st International Workshop on Peer-to-Peer Systems, Cambridge, MA, USA, March 7-8, 2002, pp. 53-62.
[7] A. Crespo and H. Garcia-Molina, “Semantic overlay networks for p2p systems”, International Workshop on Agents and Peer-to-Peer Computing (AP2PC’04), 2004, pp. 1–13.
[8] G. Salton, A. Wong, and C.S. Yang, “A vector space model for automatic indexing”, Communications of the ACM, Vol.18 n.11, Nov. 1975, pp.613-620.
[9] C. Tang, Z. Xu, and S. Dwarkadas, “Peer-to-peer information retrieval using self-organizing semantic overlay networks”, in Proceedings of ACM SIGCOMM, Aug. 2003, pp. 175–186.
[10] M. Li, W.-C. Lee, and A. Sivasubramaniam, “Semantic small world: An overlay network for peer-to-peer search”, in International Conference on Internet Protocols, 2004, pp. 228–238.
[11] PeerSim, http://peersim.sourceforge.net/.
[12] J. Strassner, S. Kim, and J. W. Hong, “Semantic Routing for Improved Network Management in the Future Internet”, Recent Trends in Wireless and Mobile Networks (WiMo), Vol. 84, 2010, pp. 163-176.
[13] Sungsu Kim, John Strassner, and James Won-Ki Hong, “Semantic Overlay Network for Peer-to-Peer Hybrid Information Search and Retrieval”, 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011), Dublin, Ireland, May 23-27, 2011. (Accepted to appear)