Content Popularity 를 이용한 P2P Semantic Overlay 기법

최승배, 황의영, 이춘화
한양대학교 전자컴퓨터통신공학과
e-mail : {choi, marieui}@mcc.hanyang.ac.kr, lee@hanyang.ac.kr

Content Popularity-Based Peer-to-Peer Semantic Overlay

Seungbae Choi, Euiyoung Hwang, and Choonhwa Lee
Dept. of Electronics Computer Engineering, Hanyang University

요 약

Peer-to-Peer(P2P) 시스템은 분산된 대용량 데이터를 효율적으로 공유하게 하여 사용자들에게 제공되는 killer application 으로 최근까지 여러 분야에서 연구가 되고 있다. 하지만 P2P 네트워크에서 피어가 소유한 데이터나 공통 관심사 또는 사회적인 관계를 고려하지 않고 무작위로 오버레이가 구성되기 때문에 검색 결과의 제약이 발생한다. 따라서 본 논문에서는 P2P 오버레이상의 효율적인 데이터 검색을 위해서 각 피어가 가지고 있는 데이터와 공통의 관심사를 기반으로 유사성을 측정하여 semantic overlay 를 구성하는 기법을 제안한다. 그리고 피어들 간의 semantic proximity 는 데이터 요약 기법을 사용하여 측정되며 측정 과정상에서 popular content 을 고려하여 semantic proximity 의 왜곡현상을 방지하여 semantic link quality 의 향상을 가져오는 방안을 도입한다.

1. 서론

P2P 오버레이 네트워크는 사용자들의 자원을 사용함으로써 저비용을 실현하고 확장성의 용이, 병목 현상 해결 등의 이유로 파일교환과 공유에 널리 활용되고 있다. P2P 의 소셜 클러스터링 기법은 각 피어 노드가 보유하고 있는 내용에 기반하여 오버레이 링크 혹은 라우팅 테이블 정보를 설정하는 방식인 content-based 기법과 공동의 관심(프로파일에 명시된 관심사나 검색 키워드 이력으로부터 관심사 추측)이나 비슷한 행동 패턴(유사한 태깅, 북마킹, 코멘팅 등)을 보이는 피어들 간의 유사성을 추출하여 오버레이가 구성하는 spiritual overlay 기법이 있는데 기법들을 통칭하여 본 논문에서는 semantic overlay 라고 칭한다.

semantic overlay 에서는 버디캐스트를 이용하여 사회적 관계 정보(버디 리스트)를 통해 각 피어는 자신과 취향이 비슷한 노드들과 우선적으로 연결이 되어 데이터를 검색하는 경우 유사한 취향을 가진 버디에게서 우선적으로 검색을 수행하게 됨으로 더욱 효율 높은 검색을 지원 할 수 있게 된다. 이러한 특징을 이용하여 본 연구진은 피어들 사이에 대용량 데이터를 고려하여 데이터 요약 기법을 통해 데이터를 구성함으로써 피어들 간의 압축된 컨텐츠 정보를 교환하여 취향이 비슷한 노드를 중심으로 오버레이를 구성, 유지하기 위해 데이터 요약 기법인 블룸 필터를 이용하여 유사도를 측정하여 semantic overlay 를 구성하고자 한다. 피어들 간의 semantic proximity 는 데이터 요약 기법을 사용하여 측정되며 측정 과정상에서 popular content 을 고려하여 semantic proximity 의 왜곡현상을 방지한다. 이를 통해 본 논문에서는 semantic link quality 의 향상을 가져오는 방안을 제안한다.

※ 이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0059045)

2. 관련연구

소셜 클러스터링 기법중 content-based 방식에서는 structured 와 unstructured 가 조금 다르게 동작한다. structured p2p 의 경우에는 파일의 키워드나 태그와 같은 메타 데이터를 해시 함수의 입력으로 하여 DHT 기반 프로토콜을 이용할 수 있다. Unstructured 의 경우에는 파일 내 특정 키워드나 주제의 출현 빈도를 피어 간에 교환하여 라우팅 자료로 사용하거나[1], 데이터 마이닝기법을 이용한 소셜 클러스터링도 가능하다[2]. spiritual overlay 는 공동의 관심(프로파일에 명시된 관심사나 검색 키워드 이력으로부터 관심사 추측)이나 비슷한 행동 패턴(즉, 유사한 태깅, 북마킹, 코멘팅 등)을 보이는 피어들 간의 유사성을 추출하여 오버레이가 구성된다[3].

3. Content Similarity and Popularity

블룸 필터는 주어진 정보를 비트 스트링으로 변환하여 압축함으로써 저장 용량과 계산량(computation)간의 절충을 꾀하는 손실 정보 요약(lossy aggregation) 기법이다. 블룸 필터는 0 으로 초기화되는 m 비트로 이루어진 벡터비트집합이다. 압축하고자 하는 정보가 주어지면 이를 k 개의 독립된 해시 함수($h_1, h_2, ..., h_k$)에 입력으로 사용하고, 이 해시 함수의 결과값이 지정하는 블룸 필터의 위치 비트가 1 로 세팅된다. 즉, 아래 수식과 같이 입력 워드 w 에 대해 해시 함수가 지정하는 위치는 1 로 바뀌고 나머지는 0 으로 남게 된다.

$BF[h_i(w)] = 1$, where $i = 1,2, ..,k$. <식 1>

컨텐츠의 유사성에 따라 semantic overlay 를 구성할 때 content popularity 로 인해 semantic proximity 의 왜곡이 발생할 수 있다. 그래서 popular content 를 배제하고 시맨틱 유사성을 측정함으로써 semantic link quality 를 최적으로 유지할 수 있도록 블룸 필터의 content aggregation 을 활용한다. 이 기법은 popular content 을 공유하는 semantic overlay 보다는 어떤 기준의 데이터를 공유하는 의미있고 유익한 오버레이 형성을 지원한다. 먼저 노드는 주변 이웃 노드 i 로부터 취합한 BF_i 로부터 popular content 으로 인한 $BF_{popular}$ 를 계산한다.

$BF_{world} := \cup_{i \in S} BF_i$, where S is the set of neighbors

$BF_{popular} = BF_{popular}[j] = 1$, if $|\{j \mid BF_{world}[j] = 1\}| >$ threshold

$BF_{popular}[j] = 0$, otherwise <식 2>

블룸 필터가 주어졌을 때 popular content 를 제외한 부분에 대한 유사도를 측정함으로써 유용한 semantic neighbor 를 감식해내는 것이 가능해질 것이다. 블룸 필터를 알고 있는 모든 노드에 대해 content popularity 를 적용한 유사성을 측정하면 semantic neighbor 리스트를 얻을 수 있다. 노드 i 와 j 의 content similarity 는 다음과 같이 측정할 수 있다.

$$\text{Content Similarity}(i, j) = \frac{|BF_{pi} \cap BF_{pj}|}{\sqrt{|BF_{pi}| \times |BF_{pj}|}} \times 100\%$$

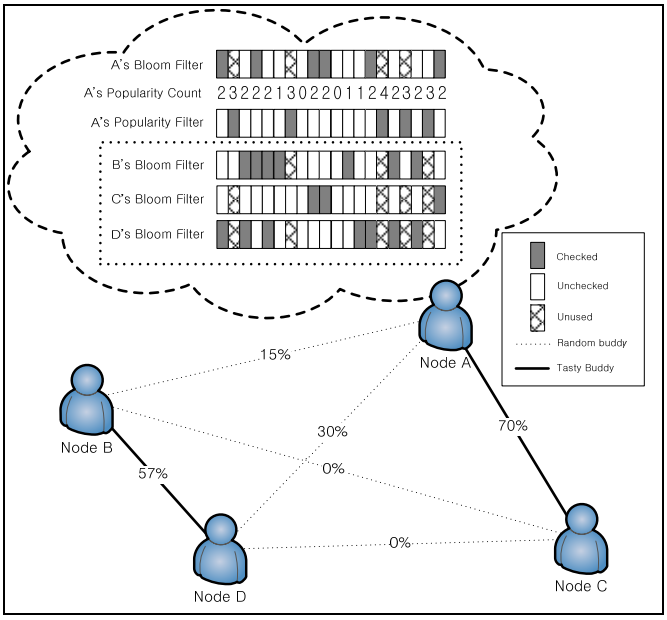
$$BF_{pi} = BF_i \cap BF_{popluar}$$
$$BF_{pj} = BF_j \cap BF_{popluar}$$

<식 3>

4. Peer-to-Peer Semantic Overlay 기법

4.1 Tasty buddy 의 형성

취향이 비슷한 노드와의 연결은 노드들이 가지고 있는 컨텐츠를 데이터 요약 기법을 통해 구성함으로써 피어들 간의 압축된 컨텐츠 정보를 교환하여 취향이 비슷한 노드를 중심으로 오버레이를 구성한다.



(그림 1) popularity 가 적용된 semantic overlay

블룸 필터를 사용하여 노드간의 semantic proximity 를 측정하여 일정 기준이상의 컨텐츠를 동일하게 소유하고 있는 노드는 tasty buddy 로 선정된다. 그림 1 은 popularity 를 적용한 semantic overlay 의 모습이다. 체크되어 있는 박스가 popular content 로 판명되어 유사성측정에 사용되지 않은 비트집합이다. 노드들은 popular content 로 인해 tasty buddy 가 되었는데 popular content 는 오버레이상에서 쉽게 발견되기 때문에 semantic link novelty 향상은 semantic proximity 를 측정시에 popular content 를 배제함으로 얻을 수 있다.

4.2 Buddycast protocol

Buddycast protocol 은 컨텐츠 자체가 아닌 컨텐츠를 소유한 사람을 찾아서 그와 관계를 맺는 것을 목적으로 한다. 즉,버디캐스트는 buddy list(tasty buddy)를 소셜 네트워크 오버레이 형성과 라우팅 목적에 사용하는 P2P 프로토콜이다. Epidemic protocol 을 기본으로 하는 버디캐스트에서는 3 가지 리스트를 가지고 있다. 아직 정보를 교환하지 않은 리스트인 C_C(candidate list), semantic proximity 를 측정후 Tasty buddy 로 선정된 C_T(tasty buddy list) 와 선정되지 못한 C_R(random list)가 있다. 버디캐스트를 기반으로 tasty buddy 를 선정하는 과정은 다음과 같다.

Algorithm 1 : Buddycast procedure

```
begin
    loop nodeout()
      if  C_C is empty  then
        C_C ← receiveRandomList(C_R)
      end if
    remove any peer from C_C if node is blocked
    Q = selectnode(C_C)
    estimateBF(Q)
    remove Q from C_C
    blocknode(Q, 4hours)
  End loop
End
```

5. 결론

Popularity 를 이용한 content-based 기법은 semantic overlay 상의 semantic proximity 의 왜곡현상을 방지하고 semantic link novelty 를 향상 시킬 수 있다. 이렇게 형성된 semantic overlay 는 기존의 semantic overlay 보다 효율적인 P2P 검색과 라우팅을 지원한다.

참고문헌

[1] A. Crespo and H. Garcia-Molina, "Routing Indices for Peer-to-Peer Systems," In Proc. of ICDCS, July 2002.

[2] A. Fast, D. Jensen, B. Levine, "Creating Social Networks to Improve Peer-to-Peer Networking", In Proc. of ACM SIGKDD, August 2005.

[3] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems," In Proc. of INFOCOM 2003, March-April 2003.