BitTorrent Swarm의 시간적 가용성 분석

조은상, 권태경, 최양희
서울대학교 컴퓨터공학부
escho@mmlab.snu.ac.kr, {tkkwon, yhchoi}@snu.ac.kr

Analysis on Temporal Availability of BitTorrent Swarms

Eunsang Cho, Taekyoung "Ted" Kwon, Yanghee Choi
School of Computer Science and Engineering, Seoul National University
escho@mmlab.snu.ac.kr, {tkkwon, yhchoi}@snu.ac.kr

요 약

BitTorrent는 데이터 공유를 목적으로 최근 가장 널리 쓰이는 Peer-to-peer 응용의 하나이다. 데이터 공유 및 배포를 목적으로 한 전통적인 수단으로는 CDN과 같은 서버 기반의 방법이 존재한다. 그러나 BitTorrent는 CDN과 다르게 사용자의 참여와 이에 따른 비용적인 이득으로 인하여 최근 널리 사용되는 추세이다. BitTorrent는 하나의 torrent 파일을 바탕으로 peer들 사이에 데이터를 공유하는데 이러한 데이터 공유의 단위를 하나의 swarm이라고 한다. 이러한 swarm은 참여한 peer들 사이에서 데이터의 완전한 사본이 존재하는 경우에만 온전히 데이터 공유가 가능하며, 이 때 완전한 사본을 가지고 있는 peer를 Seed라고 부른다. BitTorrent의 주된 목적인 데이터 공유를 달성하기 위해서는 swarm이 생성되고 나서 지속적으로 Seed가 존재하여야 하고 그렇지 않을 때에는 데이터 공유를 달성할 수 없다. 이와 다르게 CDN에서는 데이터가 서버에 항상 이용 가능한 형태로 보관되어 있으므로 계약된 기간 동안 항상 가용한 상태에 있다. 따라서 본 논문은 BitTorrent swarm의 가용성을 분석하기 위하여 swarm에 참여한 Seed의 수를 중심으로 하여 시간에 따른 데이터의 완전한 사본의 존재 여부를 확인하고, 특성을 분석한다. 약 64만 건 이상의 BitTorrent swarm에 대한 상태 정보를 분석한 결과 BitTorrent swarm이 생성 후 수년이 지난 뒤에도 유지되는 경우가 다수 존재하지만, 장기간 존재하는 swarm의 경우 가용성이 50% 미만으로 저하되는 경우가 존재함을 확인하였다.

1. 서론

BitTorrent[1]는 데이터 공유를 목적으로 최근 가장 널리 쓰이는 Peer-to-peer(이하 P2P) 응용의 하나이다. [2]에 따르면, 2007년 독일에서 P2P는 전체 인터넷 트래픽의 73.79%를 차지하고 있으며 그 가운데 BitTorrent는 66.7%에 해당한다. 즉, BitTorrent가 독일 인터넷 트래픽의 49.2%를 차지한다는 것이다. 이듬해 발표된 [3]에 따르면, 독일 이외에도 전세계적으로 P2P의 트래픽 점유율이 40% 이상으로 나타나 P2P가 전세계적으로 많이 쓰이는 데이터 공유의 수단이며 특히 BitTorrent가 널리 이용되고 있는 것을 알 수 있다.

BitTorrent는 하나의 torrent 파일을 바탕으로 peer들 사이에 데이터를 공유하는데 이러한 데이터 공유의 단위를 하나의 swarm이라고 한다. Torrent 파일은 공유하고자 하는 데이터에 대한 메타데이터(meta-data)를 담고 있으며, 이를 통해 peer들은 서로 관심 있는 데이터를 공유할 수 있다. 이렇게 같은 데이터를 공유하고자 하는 peer들의 집합을 swarm이라고 하며, swarm의 개수는 torrent 파일의 개수만큼 존재하게 된다. 이러한 swarm은 참여한 peer들 사이에서 데이터의 완전한 사본이 존재하는 경우에만 온전히 데이터 공유가 가능하며, 이 때 완전한 사본을 가지고

있는 peer를 Seed라고 부른다. 또한 Seed가 아닌 peer를 가리켜 Leecher라고 부른다.[1]

데이터 공유 및 배포를 목적으로 하는 전통적인 수단으로는 Content Delivery Network(이하 CDN)과 같은 서버 기반의 방법이 존재한다.[4] CDN은 서버에 데이터가 항상 이용 가능한 형태로 보관되어 있으므로 계약된 기간 동안 항상 가용한 상태에 있는 것이 장점이다. 그러나 CDN 서비스를 이용하기 위하여 데이터를 배포하고자 할 때 서버 이용에 대한 비용을 지불해야 하며, 사용자가 급격히 증가할 경우에 대처가 어려운 단점이 있다.

BitTorrent는 CDN과는 달리 데이터 공유 및 배포를 위하여 비용을 지불하지 않으며, 사용자가 급격히 증가하는 경우에도 큰 문제없이 동작한다. 이는 BitTorrent가 사용자의 참여에 기반하기 때문으로, BitTorrent의 사용자는 다운로드 뿐 아니라 업로드도 해야 시스템에 참여할 수 있다. 반면, 사용자 참여는 BitTorrent의 가용성에 영향을 미치는 요소가 되는데 완전한 데이터를 가진 사용자가 공유에 참여하지 않으면 그 이후에는 데이터를 이용 가능하지 않게 된다.

본 논문에서는 BitTorrent swarm의 상태 측정을 통해 가용성을 분석한다. Swarm에 참여한 Seed의 수를 중심으로 하여 시간에 따른 데이터의 완전한 사본의 존재 여부를 확인하고, 특성을 분석한다.

2. BitTorrent Swarm 상태 측정 방법

가. Torrent 메타 검색

BitTorrent Swarm의 상태를 측정하기 위하여 torrent 메타 검색 사이트인 Torrentz (http://www.torrentz.com/)의 검증된 파일 목록을 이용하였다. Torrent 파일들은 주로 Mininova (http://www.mininova.org/)와 같은 전문 웹사이트를 통해 유통된다. 이런 공개 웹사이트의 경우 검증되지 않은 torrent 파일들이 있기도 하고, 다른 웹사이트에 등록된 torrent 파일에 대해서는 찾을 수 없는 단점이 있다. 이와 같은 단점을 해결하기 위해 등장한 것이 torrent 메타 검색 사이트이다. 이를 통해 여러 웹사이트에 있는 torrent 파일들을 한 곳에서 확인할 수 있고, 사용자들의 참여를 바탕으로 실제 이용할 수 있는 파일인지를 미리 알 수 있다.

또한 torrent 메타 사이트는 각 torrent 파일에 대해 swarm에 존재하는 Seed와 Leecher의 수를 제공하고 있다. BitTorrent 시스템에서 각 swarm을 관리하는 역할을 맡고 있는 것이 tracker인데, tracker는 자신이 관리하는 swarm에 대해 Seed와 Leecher의 수를 파악하고 있다. Tracker에서 제공하는 정보를 바탕으로 하여 torrent 메타 사이트는 swarm에 대한 정보를 제공할 수 있는 것이다.

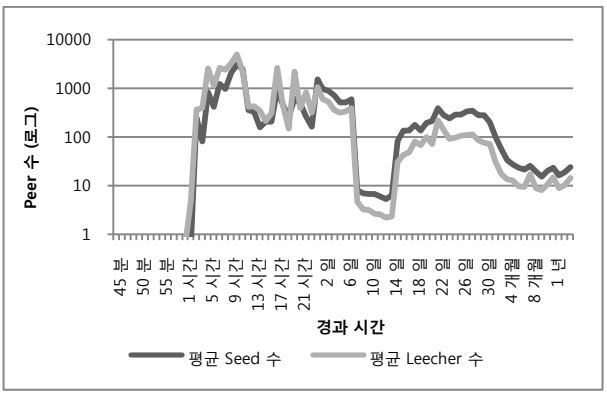

그림 1 BitTorrent swarm의 경과 시간별 평균 peer 수 (세로축: 로그)

나. 웹 크롤링을 통한 정보 수집

BitTorrent swarm에 대한 상태 정보를 수집하기 위하여 일주일간 매일 정해진 시각에 torrent 메타 검색 사이트에 대해 웹 크롤링(crawling)을 수행하였다. Torrent 메타 검색 사이트인 Torrentz에서는 검증된 torrent 파일들에 대한 리스트를 제공하는데, 최대 10만 건까지 조회가 가능하다. 2009년 11월 4일부터 11월 10일까지 일주일 간 매일 오후 10시경에 http://www.torrentz.com/verified/ 웹 페이지를 페이지 숫자를 변경해가며 크롤링하여 최대 일 10만 건에 해당하는 상태 정보를 수집하였다. 수집한 정보는 총 649,352건이며 수집한 정보의 종류는 torrent 파일의 제목, 생성일, 파일 크기, Seed와 Leecher의 수 등이다.

3. 측정 결과 및 분석

가. BitTorrent swarm의 경과 시간별 peer 수 분석

<그림 1>은 BitTorrent swarm의 생성일을 기준으로 하여 경과 시간별 평균 Seed와 Leecher의 수를 나타낸 것이다. 생성된 초기에는 swarm의 활동이 활발하여 peer의 수가 매우 많다. 이를 잘 나타내기 위하여 생성 초기에는 분 단위 및 시간 단위로 나누고 시간이 경과함에 따라 일, 월, 년으로 단위를 구분하여 그래프로 나타내었다. 이 때 가로

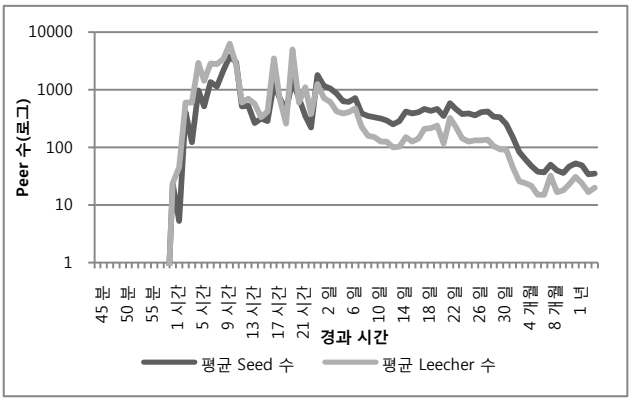

그림 2 BitTorrent swarm의 경과 시간별 평균 peer 수 (Seed 수가 0 초과인 swarm, 세로축: 로그)

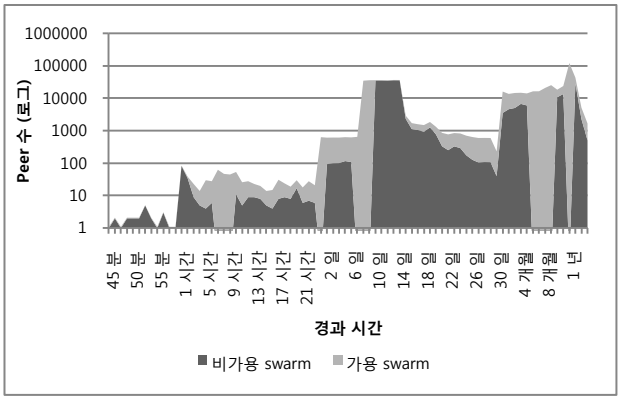

그림 3 경과 시간별 수집된 비가용 swarm의 수와 가용 swarm의 수(누적영역형 그래프, 세로축: 로그)

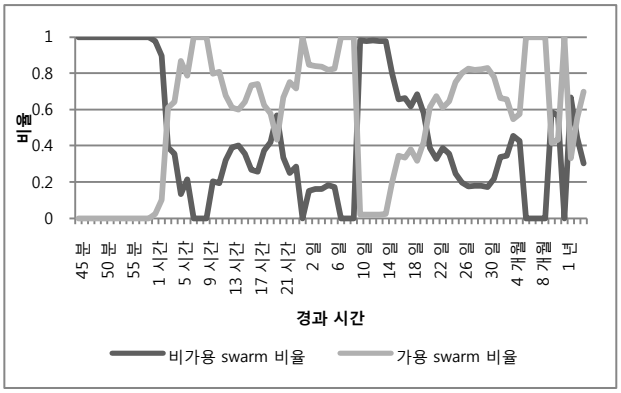

그림 4 경과 시간별 수집된 전체 BitTorrent swarm에 대한 비가용 swarm의 비율과 가용 swarm의 비율

축의 성격이 로그와 유사하므로 세로축도 로그로 나타내었다.

생성 초기에는 Leecher의 수가 Seed의 수보다 많다가 1일 경과를 기준으로 하여 Seed의 수가 더 많아지는 현상을 보인다. 또한 전체 peer의 수는 생성 초기 2~3일까지는 Seed와 Leecher 모두 1000이 넘어 매우 활발한 swarm이 형성됨을 확인할 수 있다. 수집된 자료에서는 4년여까지의 경과 시간이 지난 torrent 파일도 존재했는데, 이렇듯 수개월, 수년이 지난 오래된 torrent의 경우에도 가용한 swarm이 존재한다는 결과가 나타났다.

한편 경과 시간 8일부터 14일 사이에는 급격한 평균 peer 수 감소가 보이는데, 이는 약 일별 3만 건을 상회하는 peer 수 0인 torrent가 수집되었기 때문이다. 이 torrent들은 수집 기간인 2009년 11월 4일부터 11월 10일까지 7일간 나타났으며 모두 10월 27일에 발행되었다. 이 데이터들은 다른 날짜에 발행된 torrent에 비해 그 숫자가 지나치게 많아 Torrentz 사이트의 일시적인 오류로 인한 노이즈로 판단된다.

가용한 swarm의 특성을 확인하기 위하여 Seed 수가 0인 swarm(이하 비가용 swarm)을 제외한 결과를 바탕으로 새로운 그래프를 <그림 2>와 같이 그렸다. <그림 1>과 비교하여 경과 시간 10일부터 14일 사이의 구간에서 급격한 평균 peer 수의 감소가 나타나지 않음을 확인할 수 있다. 또한 1일 경과 시점을 기준으로 하여 Seed 수와 Leecher 수가 역전되는 현상이 그대로 유지되는 것으로 나타났다. 비가용 swarm이 제외되어 <그림 1>보다 시간 경과에 따른 peer 수 감소가 완만하게 나타났으며, 생성 후 장기간이 경과한 swarm의 경우에는 <그림 1>과의 차이가 그리 크지 않은 것이 특이점이다.

나. BitTorrent swarm의 경과 시간별 가용성 분석

BitTorrent swarm의 가용성을 분석하기 위하여 전체 수집된 swarm 가운데 Seed 수가 0 초과인 swarm(이하 가용 swarm), 즉 비가용 swarm을 제외한 swarm들에 대한 비율을 구하여 분석하였다. 비가용 swarm은 Leecher들 사이에 데이터의 완전한 사본이 존재하는 드문 경우를 제외하면 완전한 사본을 구할 수 없다. 이러한 swarm에서는 데이터가 이용 가능하지 않기 때문에 이를 제외한 나머지 swarm이 가용한 것으로 한다.

<그림 3>에서는 경과 시간별로 수집된 비가용 swarm의 수와 가용 swarm의 수를 누적영역형 그래프로 나타내었다. <그림 4>에서는 경과 시간별로 수집된 전체 BitTorrent swarm에 대한 비가용 swarm의 비율과 가용 swarm의 비율을 나타내었다. 앞서 설명했던 특이성을 보이는 경과 시간 10일에서 14일 사이의 데이터를 제외하고 보면, 시간이

흐름에 따라 비가용 swarm의 수와 비율이 모두 점차 증가하고 있다. 이는 시간이 지남에 따라 점차적으로 swarm의 가용성이 저하되는 것을 의미한다. 특히 1개월 이후 시점에서 가용성이 50% 이하로 하락하는 경우가 많은 것을 감안해 볼 때 가용성이 단기간에 저하되는 것이 BitTorrent swarm의 단점이다.

4. 결론

BitTorrent는 최근 널리 쓰이는 데이터 공유 시스템으로, CDN과 비교하여 저비용으로 데이터 공유를 달성할 수 있어 주목받고 있다. 본 논문에서는 이러한 BitTorrent의 가용성을 알아보기 위하여 일주일 간 약 64만 건 이상의 BitTorrent swarm의 상태 정보를 수집하여 분석하였다.

분석 결과 BitTorrent swarm의 생성 초기에는 Leecher의 수가 Seed의 수를 상회하나, 약 하루가 지난 뒤에는 반대의 상황이 발생한다. 또한 BitTorrent swarm은 생성된 지 수 년 뒤에도 이용 가능한 경우가 있어 장기간 데이터를 공유할 수 있었다. 그러나 가용성의 측면에서 생성 이후 약 1개월이 지나게 되면 이용 가능한 swarm의 비율이 절반 이하로 하락하는 경우가 다수 발생하여 장기간 공유할 때 가용성이 저하되는 경향을 보였다. 이러한 점은 CDN과 비교하여 단점이 될 수 있다.

이러한 단점을 보완하기 위하여 HTTP Seeding(BEP17)이나 HTTP/FTP Seeding(BEP19)과 같은 기법이 제안되어 일부 BitTorrent 클라이언트에서 활용되고 있다.[5] 그러나 이러한 방법은 CDN의 단점인 서버의 비용 부담 및 이용자 증가에 따른 확장성(scalability) 문제 등을 그대로 가지고 있어 과도기적 해결책에 가깝다. 따라서 BitTorrent가 가진 장점을 살리면서 이 문제를 해결하기 위한 노력이 필요하다.

향후 BitTorrent swarm의 속성을 더욱 잘 알기 위해 공유되는 데이터의 속성을 다양하게 수집하여 각 속성에 따라 가용성에 미치는 영향을 분석하는 작업이 이루어질 필요가 있다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT산업 원천기술개발사업의 일환으로 수행하였음. [2007-F-038-03, 미래 인터넷 핵심기술 연구]

이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사 드립니다.

참고문헌

[1] B. Cohen. “Incentives Build Robustness in BitTorrent,” Workshop on Economics of P2P systems. Jun. 2003.

[2] H. Schulze and K. Mochalski, “ipoque :: Internet Study 2007 Data about P2P, VoIP, Skype, File Hosters like RapidShare and Streaming Services like YouTube.” http://www.ipoque.com/resources/internet-studies/internet-study-2007/.

[3] H. Schulze and K. Mochalski, “ipoque :: Interet Study 2008/2009.” http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009/.

[4] A.-M.K. Pathan and R. Buyya, “A taxonomy and survey of content delivery networks,” Technical Report, GRIDS-TR-2007-4, Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia. http://www.gridbus.org/reports/CDN-Taxonomy.pdf 12 Feb. 2007.

[5] BitTorrent.org, “Index of BitTorrent Enhancement Proposals.” http://www.bittorrent.org/beps/bep_0000.html.