트래픽 분류를 위한 Hash 함수의 성능 분석

박준상, 최지혁, 김명섭
고려대학교 컴퓨터정보학과
{junsang_park, jihyeok_choi, tmskim}@korea.ac.kr

Performance analysis of hash function in traffic classification

Jun-Sang Park, Ji-Hyeok Choi, Myung-Sup Kim
Dept. of Computer and Information Science, Korea Univ.

요 약

응용 레벨 트래픽 분류를 위한 페이로드 시그니쳐 기반 분류 방법의 패턴 매칭 과정의 처리 속도 향상을 위해서 Hash 테이블 기반 방법론이 제시되었다. 하지만 Hash 테이블 기반 패턴 매칭 방법의 처리 속도는 Hash 함수의 계산 복잡도, Hash 키 값의 분포, 키 충돌 처리 방법에 따라 영향을 받는다. 따라서 본 논문에서는 Hash 테이블의 탐색 속도 개선을 위해 기존 연구에서 사용되고 있는 Bob Jenkins, ELF Hash 함수를 실험적으로 평가하여 그 성능 분석하고, 응용 레벨 트래픽 분석에 적합한 Hash 함수를 제시한다. 학내망의 실제 트래픽에 적용한 결과, Bob Jenkins 함수가 ELF 함수보다 키 값의 충돌이 적고, 계산 복잡도 측면에서도 빠른 처리 속도를 보였다.

Ⅰ. 서 론

응용 레벨 트래픽 분류 방법에 있어 페이로드 시그니쳐 기반 분석 방법은 높은 분류 정확성과 분석률을 보인다[1]. 하지만 분류 시스템의 처리 속도에 있어 현재의 고속 네트워크 상에서 발생하는 대용량 트래픽을 실시간으로 처리하기에 부적합한 방법이다.

페이로드 시그니쳐 기반 분석 시스템의 처리 속도 향상을 위해서 Hash 테이블을 이용한 패턴 매칭 방법이 제안되고 있다[2, 3]. 이러한 방법은 Hash 키 값을 우선 매칭하여 패턴 매칭을 위한 시그니쳐 탐색 공간을 최소화할 수 있기 때문에 분석 시스템의 처리 속도를 향상 시킬 수 있다. S. Kawano et al. [2]은 악성 URL 탐지를 위해서 HTTP 트래픽의 URI 를 Hash 기반으로 매칭하는 방법론을 제안하였고, G. He et al. [3]은 인터넷 응용 트래픽의 분석을 위해서 Hash 기반 패턴 매칭 방법을 제안하였다. 하지만 이러한 방법은 Hash 함수의 성능에 따라 분류 시스템의 성능이 결정되기 때문에 Hash 함수에 대한 성능 평가에 대한 연구가 선행되어야 한다. 따라서 본 논문에서는 대용량 트래픽을 실시간으로 처리하기 위해서 사용되는 Bob Jenkins[4], ELF[5] Hash 함수를 실험적으로 평가하고, 응용 레벨 트래픽 분석에 적합한 Hash 함수를 제시한다.

본 장의 서론에 이어, 2 장에서는 Hash 기반 매칭 방법을 기술하고, 3 장에서는 Hash 함수를 실험적으로 평가한다. 마지막으로 4 장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. Hash 기반 트래픽 분석 방법

Hash 기반 트래픽 분석 방법은 패턴 매칭 모듈의 시그니쳐 탐색 공간을 최소화하여 분석 시스템의 처리 속도를 향상 시킬 수 있는 방법이다. 단, 패킷의 페이로드로부터 시그니쳐의 일부 또는 전체를 자동으로 추출할 수 있는 경우에 적용 가능한 방법이다.

그림 1 은 Hash 기반 트래픽 매칭 모듈을 도식화한 것이다. 분석에 사용되는 시그니쳐를 고정된 크기의 키 값으로 변환하여 Hash 테이블을 생성한다. 플로우의 페이로드에서 시그니쳐로 사용된 필드를 추출하여 Hash 키 값으로 변환하고, 시그니쳐 Hash 테이블에 매핑한다. 해당 키에 대응되는 시그니쳐는 패턴 매칭 모듈에 전달되어 해당 플로우의 추출 스트링과 패턴 매칭을 수행하여 최종적으로 플로우가 식별된다.

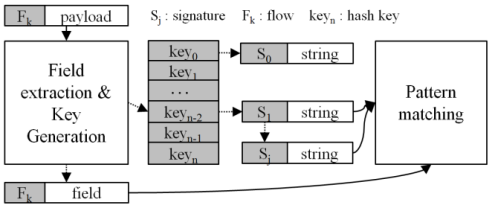

Figure 1. Hash 기반 매칭 방법

Hash 기반 분석 시스템의 성능을 최적화하기 위해서는 Hash 테이블의 크기, Hash 함수의 계산 복잡도, Hash 키의 출동 처리 방안을 고려한 Hash 테이블 적용이 요구된다.

* 본 연구는 BK21 플러스 사업 및 2012 년 정부(교육과학기술부)의 재원으로 한국연구재단(2012R1A1A2007483)의 지원을 받아 수행된 결과임.

Ⅲ. Hash 함수의 성능 분석

본 장에서는 기존 연구에서 대용량 트래픽을 실시간으로 처리하기 위해서 사용되는 Bob Jenkins, ELF Hash 함수를 실험적으로 평가하고, 효과적인 트래픽 분석을 위한 Hash 함수를 제시한다. 두 Hash 함수는 비트 연산을 기반으로 키를 생성하여 속도가 빠른 장점을 갖는다. Hash 함수에 대한 성능 평가를 위해 Hash 크기에 따른 키 값의 충돌 빈도, Hash 함수의 계산 복잡도를 비교하였다. 실험을 위해서 학내 망 전체에서 발생하는 1 일 동안의 HTTP 트래픽을 페이로드를 포함한 플로우 단위로 수집하였고, HTTP 트래픽을 서비스 별로 분석하기 위한 675 개의 페이로드 시그니쳐를 추출하였다. Hash 테이블 기반 분석 시스템은 flow/packet/byte 단위로 66.2/74.7/78.6%의 분석률을 나타냈다.

Hash 함수에 의해서 계산된 키 값의 충돌이 빈번하게 발생하면 Hash 테이블의 탐색 속도가 저하된다. 키 값의 충돌 빈도를 비교하기 위해서 두 함수의 CR(Conflict Ratio)를 수식 1 에 의해서 측정하였다.

$$\text{Conflict Ratio} = \#\ of\ sig./\#\ of\ hash\ key \qquad (1)$$

CR 값이 1.0 이면 시그니쳐와 키값이 1:1 로 매핑되어 O(1)의 탐색 속도를 보장할 수 있다. CR 값이 증가하면 충돌 빈도가 많아져서 충돌에 대한 처리가 필요하다. 따라서 키 충돌 빈도가 최소가 되는 Hash 함수와 Hash 크기를 결정해야 한다.

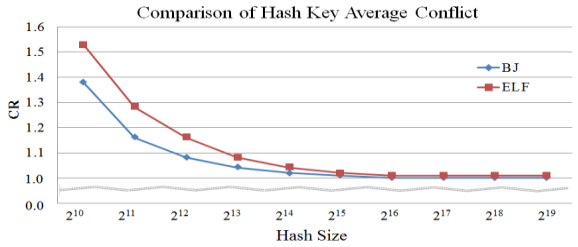

Figure 2. Hash 크기에 따른 CR 값

그림 2 는각각의 Hash 함수의 Hash 테이블의 크기에 따른 CR 갑의 변화를 보여주고 있다. BJ 는 2^{16} 의 Hash 크기일 때 CR 값이 1 되었고, ELF 는 Hash 크기를 2^{19} 까지 증가 시켜도 출동이 발생하는 것을 알 수 있다.

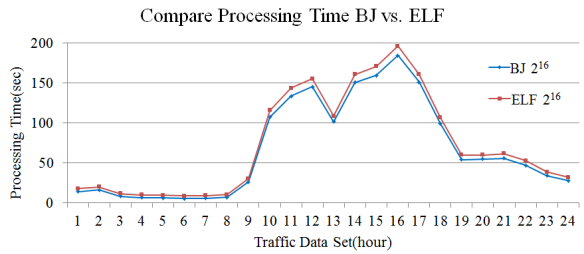

Figure 3. Hash 함수에 따른 분석 시간 비교

그림 3 은 BJ 와 ELF 의 계산 복잡도를 비교한 결과이다. 그림 2 의 결과에 따라 2^{16} 의 Hash 테이블에 출동이 발생하는 시그니쳐를 제외하고 동일한 트래픽을 분석하는 시간을 측정한 결과이다. BJ 함수가 ELF 를 사용했을 때 보다 분석 속도가 더 빠른 것을 알 수 있다.

Hash 테이블의 성능은 키 값의 충돌이 발생했을 때 처리하는 방법에 따라 영향을 받는다. Hash 키 충돌 시 처리 방법은 체이닝과 이중 해싱 등이 있다. 체이닝은 키 충돌 발생 시 해당 데이터를 연결 리스트나 트리로 구성하여 탐색는 방법이며, 이중 Hash 는 2 개의 Hash 함수를 이용하여 1 차 Hash 의 충돌이 발생하면 2 차 Hash 함수로 키 값을 생성하고, 그 값을 1 차 Hash 키 값에 더하여 새로운 키 값을 구하는 방법이다. 그림 2 와 같이 Hash 크기가 충분히 크다면 키 값의 충돌이 거의 발생하지 않기 때문에 2 차 Hash 키 값을 계산하는 부하를 줄일 수 있는 체이닝 방법이 적합하다.

Ⅳ. 결론 및 향후 과제

본 논문에서는 페이로드 시그니쳐 기반 트래픽 분류 시스템의 처리 속도 향상을 위해서 제안된 Hash 테이블 기반 분석 방법의 성능 개선을 위해서 Hash 테이블 매칭에 범용적으로 사용되는 2 가지 함수를 평가하였다. 키 값의 충돌 빈도, 계산 복잡도를 고려했을 때 Bob Jenkins Hash 함수가 Hash 테이블의 탐색 속도가 빠른 함수 이며, 키 값의 충돌이 빈번하지 않기 때문에 체이닝 기법으로 키 충돌 처리하는 것이 적합하다.

고속 링크의 대용량 트래픽을 실시간으로 분석하기 위해서는 패턴 매칭을 위한 Hash 함수뿐만아니라, 트래픽을 저장, 탐색하기 위한 Hash 함수가 요구된다. 트래픽 관리를 위한 Hash 함수에 대한 성능 평가에 대한 연구를 진행할 계획이다.

참 고 문 헌

[1] J. S. Park, S. H. Yoon, M. S. Kim, “Software Architecture for a Lightweight Payload Signature-based Traffic Classification System”, Traffic Monitoring and Analysis(TMA) Workshop 2011, Vienna, Austria, Apr. 27 2011, pp. 136-149.

[2] S. Kawano, T. Okugawa, T. Yamamoto, T. Motono, and Y. Takagi, “High-speed DPI method using multi-stage packet flow analyses” Information and Telecommunication Technologies, Santiago and Valparaiso, Chile, Dev. 5-9 2012, pp. 1-6.

[3] G. He, B. Sun, Y. Liu, X. Wu, “I-Hash Multiple Various Posotion Pattern Matching Algorithm in Internet Application Identification”, Network Infrastructure and Digital Contents(IC-NIDC), Beijing, China, Sept. 21-23, 2012, pp. 280-283.

[4] Z. Istvan, G. Alonso, M. Blott, K. Vissers, “A flexible hash table design for 10Gbps key-value stores on FPGAS”, Field Programmable Logic and Application(FPL), Porto, Portugal, Sept. 2-4, 2013, pp. 1-8.

[5] L. Tonglai, J. Hua, W. Zhang, “Research on Network Content Audit Based on Information Fingerprint”, Genetic and Evolutionary Computing, Guilin, China, Oct. 14-17, 2009, pp. 185-187.