페이로드 시그니쳐 품질 평가를 통한 고효율 응용 시그니쳐 탐색

Finding the Highly Efficient Application Signature through Payload Signature Quality Evaluation

이성호, 구영훈, Baraka D. Sija, 김명섭

고려대학교 컴퓨터정보학과

{gaek5, gyh0808, sijabarakajia25, tmskim}@korea.ac.kr

요 약

인터넷 속도의 증가와 다양한 응용의 개발로 인해 인터넷 사용자와 이들이 발생시키는 인터넷 트래픽의 양이 급격히 증가하고 있다. 트래픽 분석에 있어서 트래픽 응용 식별 방법은 페이로드 시그니쳐에 의존적이기 때문에 시그니쳐의 구성이나 개수에 따라 높은 부하와 처리 속도가 느린 단점을 갖는다. 따라서 본 논문에서는 응용 식별을 위한 페이로드 시그니쳐의 중요도를 평가하는 방법과 이를 바탕으로 높은 효율의 시그니쳐를 탐색하는 방법을 제안한다. 각 시그니쳐 별로 3 가지 기준을 바탕으로 가중치를 계산하고 계산된 가중치와 시그니쳐 맵을 통해 고효율의 시그니쳐 세트를 탐색한다. 제안하는 방법을 실제 트래픽에 적용했을 때 기존 대비 약 4 배의 응용 식별 능력을 가진 높은 효율의 시그니쳐들을 정의할 수 있었다.

Keyword : Application Traffic Identification, Application Signature, Signature Quality Evaluation, S-MAP

1. 서론

네트워크의 고속화와 더불어 다양한 서비스와 응용프로그램이 개발됨에 따라 개인 또는 기업은 인터넷으로 대표되는 네트워크에 대한 의존이 상당히 커져가고 있다. 이와 같은 현실 속에서 네트워크의 효율적 운용과 관리를 위한 응용 레벨의 트래픽의 모니터링과 분석은 네트워크 사용현황 파악과 확장 계획 수립 등의 다양한 분야에서 필요성이 증가하였다. 따라서 다양한 종류의 응용 레벨 트래픽을 정확하게 분류할 수 있는 방법과 고속 링크에서 발생하는 대용량의 트래픽을 실시간으로 처리하는 방법이 요구된다.

응용 레벨 트래픽 분류 방법에 있어 페이로드 시그니쳐 기반 분석 방법은 다른 분석 방법들에 비해 상대적으로 높은 분류 정확성과 식별률을 보였지만 낮은 처리 속도는 여전히 문제점으로 남아있었다.[1,2] 응용의 사용이 증가하고 있는 추세를 고려했을 때 페이로드 기반 분석 방법의 처리 속도 문제는 반드시 해결되어야 하는 과제이다.

본 논문에서는 트래픽 응용 식별을 위한 시그니쳐 생성 시스템에서 생성된 페이로드 시그니쳐의 중요도를 3 가지 기준으로 평가한다. 그리고 평가는 내용을 수치화해 응용 식별 측면에서 상대적으로 높은 효율을 갖는 시그니쳐를 탐색하는 방법을 제안한다.

제안하는 방법의 검증을 위하여 토렌트 응용 트래픽을 통해 실험을 진행하였다. 실험을 위해 응용 시그니쳐 자동 생성 시스템을 통해 시그니쳐들을 추출했다. 추출된 시그니쳐에 제안하는 시그니쳐 품질 평가 방법을 적용시켰고 이를 통해 탐색된 고효율의 시그니쳐들은 이전의 시그니쳐들 보다 평균적으로 4 배의 응용 식별 효율을 보였다. 품질 평가 방법은 시그니쳐의 활용성, 고유성 그리고 매칭 속도 3 가지 측면에서 이루어진다. 3 가지 기준으로 Quality Weight 값을 계산하고 시그니쳐 맵인 S-MAP 을 구성한다. 최종적으로 S-MAP 을 통해 본 논문에서 제안하는 고효율의 응용 시그니쳐를 탐색할 수 있다. 응용 식별률은 기존의 모든 시그니쳐들과 비교했을 때 약 3% 미만의 차이를 보였기 때문에 제안하는 시그니쳐 품질 평가 방법은 불필요한 시그니쳐는 제외하고 가치가 높은 시그니쳐만을 탐색할 수 있는 효율적인 방법이라고 판단할 수 있다.

본 논문의 구성은 다음과 같다. 본 장의 서론에 이어, 2 장에서는 관련 연구에 대해 기술하고, 3 장에서는 해결하고자 하는 문제에 대해 정의한다. 4 장에서는 제안하는 방법의 핵심이 되는 시그니쳐 품질 평가 방법과 탐색 방법에 대해 설명한다. 5 장에서는 제안하는 방법을 통해 정의된 고효율 시그니쳐를 트래픽 응용 식별 시스템에 적용해 실험해보고 그 결과를 통해 제안하는 방법의 타당성을 증명한다. 마지막으로 6 장에서는 결론 및 향후 연구에 대해 기술한다.

이 논문은 2015 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2015R1D1A3A01018057)

2. 관련 연구

응용 프로그램 서비스 제공자는 방화벽을 우회하여 사용자에게 원활한 서비스를 제공하기 위해 복잡한 구조의 응용 레벨 프로토콜 구성하기 때문에 시그니쳐 또한 복잡하고 다양한 형태로 나타난다. 또한 인터넷에 기반한 응용의 증가로 인해 시그니쳐의 개수가 증가하고 그 가치 또한 높아지고 있다. 시그니쳐의 복잡도가 커지고, 개수가 증가하면서, 페이로드 시그니쳐 기반 응용 식별 시스템의 처리 속도는 전체적인 트래픽 분석 시스템의 성능을 결정하는 중요한 요소로 작용하게 되었다. 따라서 본 논문에서 제안하는 시그니쳐의 중요도 평가 방법과 이를 통한 고효율 시그니쳐 탐색 방법을 통해 앞에서 정의한 시그니쳐의 비효율을 개선하고 트래픽 응용 식별 시스템의 근본적인 성능 향상 효과를 기대할 수 있다. 이러한 방법과 관련해 기존에도 많은 연구들이 진행되어 왔다.

응용 식별 시스템의 속도 향상을 위한 다양한 패턴 매칭 알고리즘들이 제안되었다. 하지만 패턴 매칭 알고리즘의 성능은 입력 데이터의 구성에 의존적이며, 제한적인 성능 향상을 나타낸다[5]. 또한 매칭 단계에서 시그니쳐별로 갖는 오프셋이나 패턴에 대한 고려를 하지 않기 때문에 시그니쳐의 구성이나 구조에 종속적이라는 한계점이 있었다.

따라서 기존의 매칭 알고리즘 성능 개선의 한계적 문제점을 해결하기 위해 응용 레벨 트래픽의 발생 패턴을 분석 시스템에 반영하여 시그니쳐의 탐색 공간을 최소화하는 방법이 제안되었다[6]. 트래픽의 발생 패턴이나 특징을 반영해 응용 식별 효율을 높인다는 점에서 본 논문에서 제안하는 방법과 유사하다. 그러나 기존의 연구에서는 각 시그니쳐의 HC(Hit Count)만을 고려했고 중요도를 수치나 값을 통해 평가 하진 못하였다. 결과적으로 탐색 공간은 최소화 했지만 방법을 통해 효율적인 시그니쳐를 탐색하고 분류하지 못했다. 결과적으로 시그니쳐의 비효율성을 근본적으로 해결하지 못했고 방법 역시 제한적이었다.

시그니쳐 패턴 매칭 알고리즘이나 시그니쳐의 구성에 의존하지 않고 트래픽 간의 지역적인 연관성을 이용한 분석 방법도 제시되었다[3,4,7]. 하지만 CDN(Content Delivery Network)트래픽과 같은 상호 연관성은 존재하지만, 서로 다른 응용을 나타내는 트래픽을 식별하는 데에는 그 한계점이 존재 하고 또한 시그니쳐를 사용하지 않고 트래픽 간 연관성을 활용한 응용 식별 방법이기 때문에 그 정확도가 낮다. 따라서 전체적인 트래픽 분석 시스템의 구성에 있어서 한계점이 존재한다.

기존의 연구는 페이로드 시그니쳐 기반 트래픽 분류 시스템의 처리 속도 향상을 위해서 패턴 매칭 기법을 소프트웨어 또는 하드웨어적으로 개선하려는 노력과 트래픽의 특징을 정의하고 그룹화해서 시그니쳐 탐색 범위를 최소화 하는 방법이 주를 이루었다. 하지만 이러한 방법은 모두 앞에서 정의한 시그니쳐의 비효율성을 판단하고 수치화할 수 있는 방법이 없었다. 때문에 비효율적인 시그니쳐들은 계속 활용되어왔고 그 결과 응용 트래픽 식별 방법이나 전체 트래픽 분석 시스템의 성능 향상에 있어 많은 개선을 이루어낼 수 없었다.

3. 문제 정의

현재의 응용 트래픽 시그니쳐 생성 시스템에서 생성된 페이로드 시그니쳐는 단순히 분석을 목표로 하는 응용의 트래픽 파일에서 공통적으로 나오는 문자열을 찾아 나열하는 단계에 지나지 않았다. 따라서 생성된 시그니쳐들 중에는 응용 식별 측면에서 의미를 갖는 높은 수준의 시그니쳐도 있지만 단순히 의미 없는 공통된 문자들이 나열된 낮은 수준의 시그니쳐들도 존재하게 된다.

생성된 시그니쳐의 정확성과 그 의미를 평가할 명확한 지표와 기준이 없었다. 그 결과 불필요한 시그니쳐들 또한 분석 과정에 포함될 수 밖에 없었고 응용 식별률 측면에서는 일정 수준 이상을 만족했지만, 각 시그니쳐의 식별 효율과 그 중요성을 알 수 없었기 때문에 어떤 시그니쳐가 중요한 시그니쳐인지 판단할 수 없었다.

본 논문에서는 이러한 현상을 시그니쳐의 비효율로 정의하고 각각의 시그니쳐 별로 식별 효율과 시그니쳐로서의 중요도를 평가하는 방법과 이를 구체적으로 수치화해 고효율의 시그니쳐를 탐색하는 방법을 제안한다.

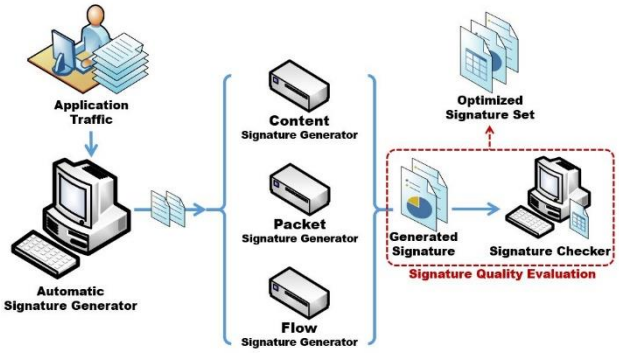

그림 1. 제안하는 방법의 개념도

그림 1 은 현재의 응용 트래픽 페이로드 시그니쳐 생성 시스템을 나타낸다. 시그니쳐 생성을 목표로 하는 응용에 대한 트래픽 파일들을 Automatic Signature Generator 에 넣게 되면 각 트래픽 파일들에서 공통적으로 나오는 문자열을 바탕으로 해당 응용에 대한 Flow, Packet, Content 시그니쳐가 생성되게 된다. 본 논문에서 제안하는 방법의 핵심은 그림 1 의 Signature Quality Evaluation 부분이다. Quality Evaluation 방법은 앞서 Signature Generator 에서 생성된 시그니쳐들을 받아와 각 시그니쳐들을 3 가지 기준을 바탕으로 평가하고 수치화해 가장 최적의 시그니쳐을 탐색한다. 이를 통해

높은 응용 식별 효율과 시그니쳐로서 특징을 갖는 Optimized Signature Set 을 탐색할 수 있다.

4. 제안하는 시그니쳐 품질 평가 방법

Quality Evaluation 은 Signature Generator 에서 생성된 시그니쳐의 응용 식별 효율, 중요도 그리고 의미를 평가하는 방법으로 그림 2 와 같이 크게 3 가지 기준으로 나뉜다.

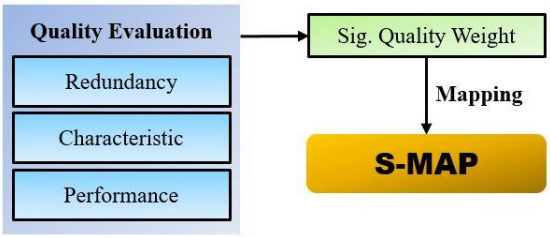

그림 2. 시그니쳐 중요도 평가 방법

첫번째 기준은 시그니쳐의 활용성(Redundancy)이다. 만약 어떤 응용 트래픽의 Flow, Packet 등이 다수의 시그니쳐에 의해 매칭 되었다면 해당 Flow, Packet 들을 분석하는 시그니쳐들은 같은 Flow 를 중복적으로 여러 번 매칭하게 된다. 따라서 시그니쳐로서의 가치가 상대적으로 낮다고 볼 수 있다. 이와 반대 개념으로 특정 시그니쳐가 여러 개의 Flow 나 Packet 에 매칭된 다면 해당 시그니쳐는 활용성과 효율이 높다고 판단할 수 있다. Quality Evaluation 에서는 이러한 시그니쳐의 활용 정도를 '시그니쳐 활용성' 이라는 기준으로 삼아 생성된 모든 시그니쳐들을 수식(1)과 같이 계산 한다.

$$\text{Redundancy Value} = |\{\text{Flows} \mid \text{Identified by } Sig_x\}| \quad (1)$$

두번째 기준은 시그니쳐의 고유성(Characteristic)으로 생성된 시그니쳐 전체 문장에서 의미를 갖는 글자와 숫자의 비율을 고려한다. 실제로 생성된 시그니쳐들 중에는 응용 트래픽 식별의 측면에서 시그니쳐로써 의미나 고유성을 갖지않는 Padding bits 나 Random String 또한 포함되어 있다. 따라서 보다 시그니쳐로써의 특징과 고유성을 갖는 시그니쳐를 찾기 위한 방법이 필요하다. Quality Evaluation 에서는 시그니쳐 문자열에서 표현가능한 Character 와 Numeric 의 비율 및 응용 이름의 유무 등을 평가한다. 그리고 이러한 기준을 '시그니쳐의 고유성' 이라 정의하고 수식 2 와 같이 계산한다.

$$\text{Characteristic Value} = \frac{|character| + |numberic|}{Sig_x TotalLength} \quad (2)$$

(if application name is in the Sig, CV is fixed to 1)

세번째 기준은 시그니쳐의 매칭 속도(Performance)이다. 시그니쳐의 매칭 속도를 판단하기 위해서는 매칭 오프셋을 고려해야 한다. 시그니쳐의 매칭 오프셋이란 시그니쳐가 나타나는 응용 트래픽 패킷의 페이로드 위치를 의미한다. 시그니쳐의 매칭 속도를 기준으로 설정한 이유는 응용 트래픽 식별이나 전체 트래픽 분석 시스템의 성능을 고려할 때, 시그니쳐가 존재하는 오프셋이 앞쪽에 고정되어 있다면 보다 빠른 응용 식별이 가능하고 이는 트래픽 분석 시스템의 성능 향상과 밀접한 연관이 있다. 따라서 시그니쳐의 매칭 속도가 빠를수록 좋은 시그니쳐이다. 따라서 '시그니쳐의 매칭 속도' 를 세번째 기준으로 정의하고 수식 3 과 같이 계산한다.

$$\text{Performance Value} = \frac{L - Sig_x MatchedOffset}{L} \quad (3)$$

(L=Max Payload Length)

결과적으로 본 논문에서 제안하는 Quality Evaluation 은 정의한 시그니쳐의 활용성(Redundancy), 시그니쳐의 고유성(Characteristic), 시그니쳐 매칭 속도(Performance)를 바탕으로 수행한다.

$$\text{Quality Weight} = PV \times CV \times \log(RV) \quad (4)$$

(PV=Performance Value, CV=Characteristic Value, RV=Redundacy Value)

각 시그니쳐 별로 3 가지 정의에 대한 내용들을 수치화 하고 계산해 최종적으로 Quality Evaluation 을 위한 최종 값인 Quality Weight 을 계산한다. Quality Weight 의 계산 방법은 수식 4 과 같다. 수식 4 와 같이 구성한 이유는 각 시그니쳐 간의 Quality Weight 값의 편차를 낮추고 정규화 하기 위해서이다.

Sig ID	QV Weight	Identified Flow Sequence
1	0.9	3,5,7,10,11,12
2	2.5	2,5,9,11
3	2.9	3,5
4	3.0	10,12
5	2.0	1,3,5,7,9,10
6	3.4	11
7	3.8	6
8	1.5	1,2,5,8,9,12
9	4.0	5
10	1.7	2,4,5,6,9
11	4.5	4
12	4.7	1

표 1. S-MAP Table(Example)

그림 2 와 같이 Quality Weight 값은 이후 고효율 시그니쳐 탐색 방법인 S-MAP 을 구성하기 위해 사용된다. 그림 2 에서 정의한 S-MAP 은 고효율의 시그니쳐 세트를 탐색하기 위한 방법으로 앞서 정의한 Quliaty Weight 값을 바탕으로 구성된다.

표 1 과 그림 3 은 S-MAP 을 구성한 예이다. 각

시그니쳐가 분석하는 Flow 들과 Quality Weight 를 사용해 표 1 과 같은 S-MAP Table 을 구성하고 2 차원 맵을 구성해 각 Flow 를 분석하는 시그니쳐를 매핑한다. 그리고 각 포인트마다 Quality Weight 를 참조해 모든 X 축 Flow 포인트를 연결하는 선을 그린다.

그림 3 과 같이 X 축에 매핑 된 모든 Flow 들을 연결하는 선은 각 포인트의 최대 Quality Weight 값들을 연결을 통해 구성된다.

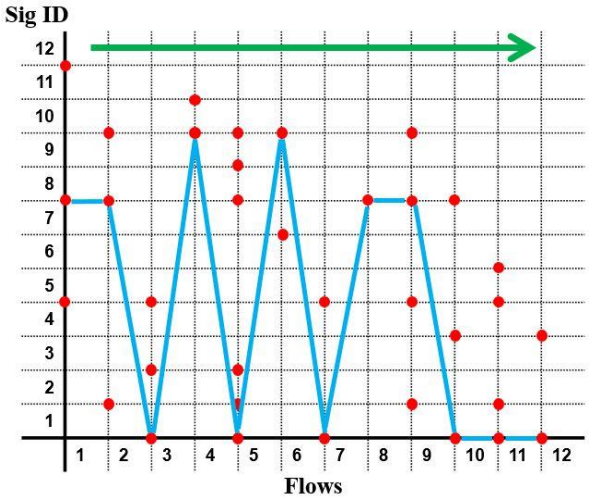

그림 3. S-MAP(Example)

그림 3 의 S-MAP 예를 참고했을 때, 각 포인트의 Quality Weight 값을 참조한 선이 그려지고 이 때 사용된 시그니쳐 세트(Y 축)은 {1, 8, 10}번 시그니쳐이다. 따라서 고효율의 시그니쳐는 1, 8 10 번 시그니쳐이고 이때 해당 시그니쳐의 효율은 식 5 의 Signature Average Efficiency 를 통해 구한다.

SAE 의 의미는 단위 시그니쳐가 분석한 Flow 의 개수이다. 이전보다 시그니쳐의 개수가 1/4 로 압축

되었으므로 SAE 는 4 배 커지게 된다. 따라서 S-MAP 을 통해 탐색된 시그니쳐 세트는 이전의 시그니쳐들 보다 평균 4 배 높은 시그니쳐 효율을 갖는다고 판단할 수 있다.

$$SAE = \frac{\sum_{x=1}^{|SigSet|} |\{Flows \mid IdentifiedbySig_x\}|}{|SigSet|} \quad (5)$$

(SAE=Signature Avergae Efficiency)

5. 실험 및 결과

본 장에서는 3 장에서 제안한 시그니쳐 Quality Evaluation 방법을 실제 응용 트래픽에 적용시켜보고 그 결과를 분석한다.

성능 평가를 위해 사용한 트래픽은 표 2 와 같다. 6 개의 트레이스 모두 토렌트 응용의 트래픽으로 각각 동영상 파일을 다운로드하며 수집했다.

Trace ID	Size(MB)	Flow	Pkt
1	113	2500	114,705
2	110	405	104,785
3	46	1452	48,867
4	34	1503	52,069
5	55	501	52,302
6	54	646	52,048

표 2. Traffic Trace

표 1 의 트레이스들을 바탕으로 그림 1 의 시스템을 적용해 Quality Evaluation 평가를 진행했고, 그 결과 그림 4,5 와 같은 결과를 얻을 수 있었다. 평가를 위해 적용한 시그니쳐들은 그림 1 의 Content Siganture 로 가장 낮은 단위의 시그니쳐이다. Content Signature 를 사용한 이유는 가장 낮은 단위의 시그니쳐에서 Quality Evaluation 방법의 효율성이 검증되면 이후 Packet 이나 Flow 시그니쳐에도 쉽게 적용 가능하기 때문이다.

그림 4 는 3 장에서 정의한 방법을 통해 구현된 S-MAP 이다. 6 개의 트레이스에서 총 139 개의 응용 Content Signature 와 약 14,000 개의 패킷 시퀀스가 24 개의 시그니쳐로 압축되었다.

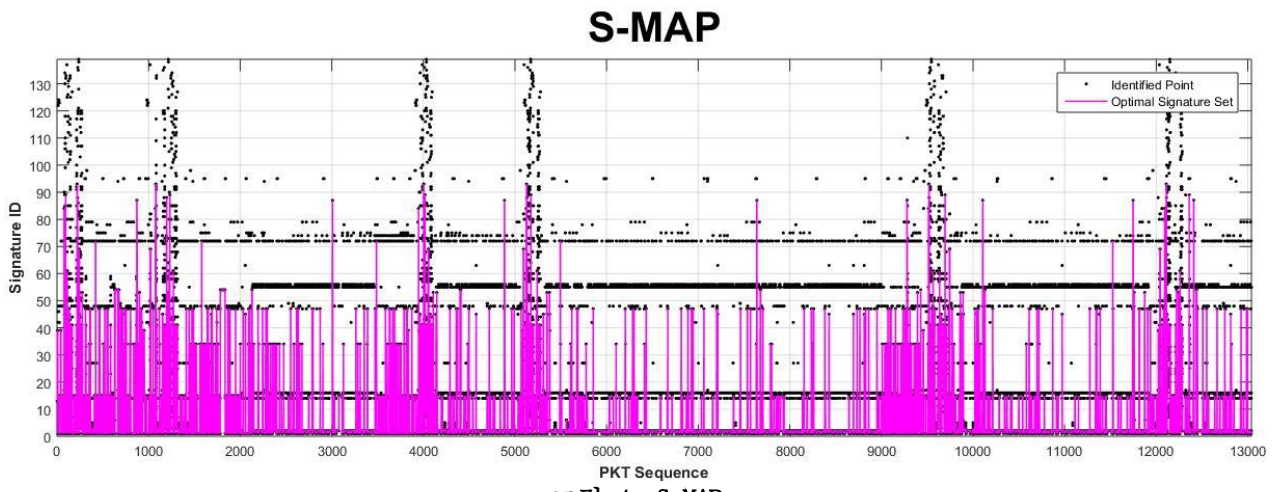

그림 4. S-MAP

그 결과 앞서 정의한 Signature Average Efficiency 의 측면에서 Quality Evaluation 을 적용하지 않은 시그니쳐 보다 평균 5 배 정도 높은 효율을 갖는 시그니쳐들을 탐색할 수 있었다. 그림 4 의 S-MAP 에서 각 점은 시그니쳐로 식별된 패킷 시퀀스를 의미한다. S-MAP 을 통해 확인하였을 때에 Y 축의 139 개의 시그니쳐 중 실제 연결된 점들의 수준이 24 개로 일정한 것을 확인 할 수 있다.

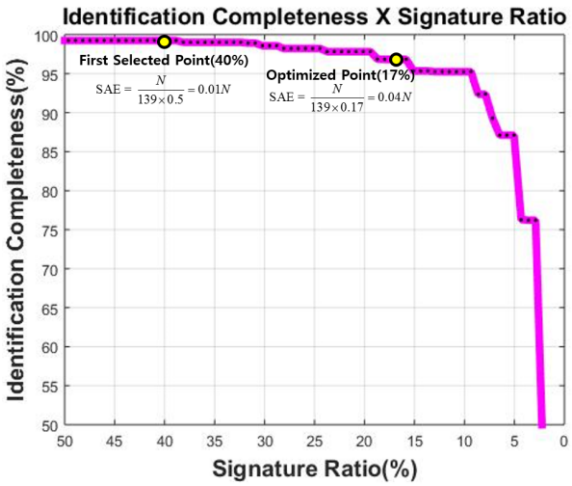

그림 5. 응용 식별률에 따른 시그니쳐 비율

그림 5 는 시그니쳐 비율에 따른 식별률을 나타낸다. 생성된 139 개의 시그니쳐를 100%로 정의했을 때 각 시그니쳐의 Quality Weight 가 낮은 시그니쳐부터 순차적으로 제외시키면서 시그니쳐 세트를 압축한다.

그림 5 를 참고했을 때 시그니쳐 세트의 압축률이 약 30%가 된 시점부터 식별률이 점진적으로 감소하는 것을 확인할 수 있다. 139 개의 시그니쳐를 같은 식별률을 유지하며 전체 시그니쳐의 35%인 40 개의 시그니쳐로 압축할 수 있었다. 하지만 본 논문에서 제안하는 Quality Evaluation 방법을 사용했을 때 압축된 24 개의 시그니쳐는 전체 시그니쳐의 약 17%로 1 차적으로 추려진 40 개의 시그니쳐 보다 18% 낮은 시그니쳐 비율을 갖는 동시에 응용 식별률 측면에서는 3% 미만의 차이를 보이고 있다.

시그니쳐 비율이 17% 보다 낮아지면서 본 논문에서 제안한 Quality Evaluation 을 통해 정의된 높은 Quality Weight 값을 갖는 시그니쳐들이 시그니쳐 세트에서 제외되고있다. 그리고 그 결과 시그니쳐 비율이 약 10%가 되는 시점부터 응용 식별률이 급격하게 낮아지는 것을 확인할 수 있다. 앞서 정의한 Signature Average Efficiency 를 바탕으로 계산했을 때 Identified Flow 의 개수를 N 으로 가정하고 그림 5 의 First Selected Point 에서는 0.01N 의 SAE 값을 갖는다. 반면 Optimized Point 에서는 0.04N 의 SAE 값을 갖는다. 따라서 본 논문에서 제안하는 QE 를 통해 이전에 비해 약 4~5 배의 응용 식별 효율을 갖는 시그니쳐를 탐색할 수 있었다.

6. 결론 및 향후 연구

결과적으로 본 논문에서 제안한 Quality Evaluation 방법을 통해 탐색된 시그니쳐 세트는 기존의 시그니쳐들 보다 높은 효율성을 갖는 동시에 응용 식별률 측면에서 큰 차이 없는 것을 확인할 수 있다. 또한 그림 5 의 그래프를 참고하였을 때, 정의된 높은 Quality Weight 의 시그니쳐들은 매우 고효율의 시그니쳐인 것을 알 수 있다.

본 논문에서 제안한 응용 트래픽 식별을 위한 페이로드 시그니쳐의 Quality Evaluation 방법은 결과적으로 그 의미와 효율성 측면에서 매우 긍정적인 것을 확인할 수 있다. 따라서 제안하는 방법을 이후 연구할 실시간 응용 시그니쳐 자동 생성 시스템에 적용해 실시간으로 생성된 시그니쳐의 중요도와 가치를 평가해볼 계획이다.

참고 문헌

[1] J. S. Park, J. W. Park, S. H. Yoon, Y. S. Oh, M. S. Kim, "Development of signature Generation system and Verification Network for Application Level Traffic Classification", in Proc. KIPS conf. Apr. 23-24, 2009, pp. 1288-1291, PuSan, Korea.

[2] S. H. Yoon, H. G. Roh, M. S. Kim, "Internet Application Traffic Classification using Traffic Measurement Agent ", in Proc. KICS Jul. 2-4, 2008, pp.618. Jeju Island, Korea.

[3] Fnag Yu, Zhifeng Chen, Yanlei Dino, T. V. Lakshman, Randy H. Katz, “Fast and memory Efficient Regular Expression Matching for Deep Packet Inspection” ANCS 2006, December , 2006, San jose, California USA.

[4] Christopher L. Hayes , Yan Luo, “DPICO: a high speed deep packet inspection engine using compact finite automata”, ACM/IEEE Symposium on Architecture for networking and communications systems, December 03-04, 2007, Orlando, Florida, USA

[5] C. L. Hayes and Y. Luo, “DPICO: A high speed deep packet inspection engine using compact finite automata,” in *Proc. ACM/IEEE Symp. Architecture Netw. Commun Syst. (ANCS ‘07)*, pp. 195-203, Orlando, USA, Dec. 2007.

[7] J. S. Park and M. S. Kim, “Performance improvement of application-level traffic classification system using application traffic pattern,” in *Proc. KICS Summer Conf.*, pp. 3-7, Jeju, Korea, Jun. 2011.

[8] J.-S. Park, S.-H. Yoon, and M.-S. Kim, “Performance improvement of the payload signature based traffic classification system using application traffic locality,” *J. KICS*, vol. 38B, no. 7, pp. 519-525, Jul. 2013.