Journal of The Korea Society of Computer and Information
Vol. 19, No. 3, March 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.3.045

개선된 휴리스틱 규칙 및 의사 결정 트리 분석을 이용한 P2P 트래픽 분류 기법

예 우 지 엔*, 조 경 산**

P2P Traffic Classification using Advanced Heuristic Rules and Analysis of Decision Tree Algorithms

Wujian Ye*, Kyungsan Cho **

요 약

본 논문에서는 기존 기법들의 제한점을 개선하기 위해 휴리스틱 규칙 및 기계학습 분석 결과를 이용한 두 단계의 P2P 트래픽 분류 기법을 제안한다. 첫 번째 단계는 패킷 레벨의 시그니처 기반 분류기이고, 두 번째 단계는 플로우 레벨에서 수행되는 패턴 휴리스틱 규칙 및 통계 기반 분류기이다. 제안된 패턴 휴리스틱 규칙은 분류의 정확도를 높이고 통계 기반 분류기가 처리할 트래픽의 양을 줄일 수 있다. 다양한 의사 결정 트리 알고리즘의 분석을 기반으로 통계 기반 분류기는 가장 효율적인 REPTree로 구현하고, 앙상블 알고리즘을 통해 통계 기반 분류기의 성능을 개선한다. 실제 환경의 데이터 집합을 이용한 검증 분석을 통해, 본 제안 기법이 기존 기법에 비해 높은 정확도와 낮은 과부하를 제공함을 제시한다.

▸ Keywords : P2P 트래픽 분류, 하이브리드 기법, 시그니처 기반, 휴리스틱 규칙, 의사 결정 트리

Abstract

In this paper, an improved two-step P2P traffic classification scheme is proposed to overcome the limitations of the existing methods. The first step is a signature-based classifier at the packet-level. The second step consists of pattern heuristic rules and a statistics-based classifier at the flow-level. With pattern heuristic rules, the accuracy can be improved and the amount of traffic to be classified by statistics-based classifier can be reduced. Based on the analysis of different decision tree algorithms, the statistics-based classifier is implemented with REPTree. In addition, the ensemble algorithm is used to improve the performance of statistics-based classifier.

•제1저자 : 예우지엔 •교신저자 : 조경산
•투고일 : 2013. 12. 17, 심사일 : 2014. 1. 21, 게재확정일 : 2014. 2. 18
* 단국대학교 대학원 컴퓨터학과(Dept. of Computer, Dankook University)
** 단국대학교 소프트웨어학과 교수(Dept. of Software Science, Dankook University)

Through the verification with the real datasets, it is shown that our hybrid scheme provides higher accuracy and lower overhead compared to other existing schemes.

▸ Keywords : P2P traffic classification, hybrid scheme, signature-based, heuristic rules, decision tree

I. 서 론

Peer-to-peer(P2P)는 중앙 서버의 간섭을 받지 않고 사용자 간의 직접적인 통신을 가능하게 하는 기술이며, P2P 파일 공유 시스템은 P2P 기술의 대표적인 응용이다[1]. 사용자는 Gnutella, Kazaa, eDonkey 또는 BitTorrent와 같은 P2P 응용 프로그램을 통해 인터넷에 접속하여 원하는 파일을 쉽게 검색하고 획득하여 유포할 수 있다. 최근 몇 년 동안 P2P 사용자의 수가 급증함에 따라 P2P 트래픽은 인터넷 트래픽의 상당부분을 차지하게 되었다[2]. 하지만, P2P 트래픽이 증가할수록 대역폭, 보안 및 관리 등의 문제도 증가한다[3]. 이러한 문제를 해결하는 동시에 인터넷 서비스 공급자(ISP) 및 non-P2P 사용자의 이익을 보호하기 위해, P2P 트래픽의 양을 조절해야 한다. 따라서, P2P 트래픽의 정확한 분류가 필요해졌다[4].

P2P 트래픽의 분류를 위해 포트 기반, 시그니처 기반, 패턴 기반 및 통계 기반 등의 기법이 제안되었다. 그러나 기존 P2P 식별 기술을 회피하기 위해 포트 변조(port disguise)에서 페이로드 암호화 및 패킷 크기 제어 기술까지 등장하였다. 다양한 회피 기술을 극복하기 위해 시그니처 기반 및 통계 기반의 두 가지 기법을 결합한 하이브리드 분류 기법도 제시되었지만 여러 제한점을 가진다.

본 논문에서는 통계 기반 분류 기법의 정확도를 높이고 계산량을 줄일 수 있는 개선된 휴리스틱 규칙 및 기계학습의 분석 결과를 적용한 두 단계의 P2P 트래픽 분류 기법을 제안한다. 첫 번째 단계는 패킷 레벨에서 동작하는 시그니처 기반 분류기이며, 두 번째 단계는 플로우 레벨에서 동작하는 패턴 휴리스틱 규칙을 이용한 통계 기반 분류기이다. 패턴 휴리스틱 규칙을 통해 분류 정확도를 높이고 통계 기반 분류기가 처리할 트래픽의 양을 줄일 수 있다. 다양한 유형의 의사 결정 트리 알고리즘의 분석을 기반으로 REPTree를 활용하는 통계 기반 분류기와 앙상블 알고리즘을 이용하여 분류의 정확도를 높일 수 있다. 검증을 위해, 실제 환경에서 형성된 데이터 집합(dataset)을 이용하여 제안 기법의 성능을 분석한다.

본 논문의 구성은 다음과 같다. 2장은 관련연구로 기존 P2P분류 기법의 장단점을 제시한다. 3장에서는 2장의 분석을 기반으로 휴리스틱 규칙 및 의사 결정 트리를 이용한 두 단계 기법을 제안한다. 4장에서는 제안 기법의 성능을 검증하고, 5장의 결론으로 본 논문을 마무리 짓는다.

II. 관련 연구

P2P 트래픽의 정확한 분류는 효율적인 네트워크 관리와 네트워크 자원의 합리적인 활용에 중요한 역할을 한다. P2P 트래픽을 분류하기 위해 여러 가지 기법들이 제안되었다.

대표적인 초기의 기법은 포트 기반 기법이다. 초기 P2P 응용 프로그램은 지정된 포트 번호를 사용하므로 이 기법은 패킷 헤더에 있는 포트 번호만으로 트래픽의 유형을 쉽게 판단할 수 있다. 이 기법은 간단하고 빠르지만, 많은 P2P 응용 프로그램은 동적으로 임의의 포트 번호를 사용하기 때문에 모든 P2P 트래픽을 식별할 수 없다. 예를 들어, Kazaa P2P 프로토콜의 전체 트래픽에서 기본 포트 번호는 단지 30%만을 차지하고 있다[5]. 또한 포트 기반 분석은 인터넷 트래픽의 30%-70%만을 식별할 수 있다고 보고되었다[6].

포트 기반 기법의 단점을 보완하기 위해, 시그니처 기반 기법이 등장하였다. 이는 패킷의 페이로드에 있는 특정 문자열을 인식하여 P2P 트래픽을 식별한다. 이 기법은 높은 정확도와 효율성을 보이지만 패킷의 페이로드를 분석해야 하므로 많은 작업이 필요하다. 또한 암호화된 페이로드와 새로운 유형의 P2P 트래픽에는 적용할 수 없다. 시그니처 기반의 분류 기법은 P2P의 오탐(false positives)과 미탐(false negatives)을 5%까지 감소한 것으로 분석되었다[5]. 시그니처 기반 기법이 96%의 높은 정확도를 보였지만, 암호화된

eMule 트래픽에 대한 정확도는 30%에서 70%까지인 것으로 분석되었다[7].

위 두 가지 기법의 문제점을 해결하기 위해 패턴 기반 기법 및 통계 기반 기법이 제안되었다.

패턴 기반 기법은 알려진 활동이나 응용이 나타내는 패턴들과 특정 호스트에 의하여 생성된 통신 패턴을 비교하는 기법이다[8]. Karagiannis 등은 한 호스트가 자신의 목적 IP 주소 및 수신 포트 번호에서 연결을 기다릴 때, 이 호스트에 연결된 서로 상이한 IP 주소의 수와 상이한 포트 번호의 수가 동일하다는 P2P 트래픽의 특성에 따라 P2P 트래픽을 탐지했다[9]. 그러나 이 기법은 단일의 플로우는 분류할 수 없으며 P2P 트래픽을 특정 응용 프로그램으로 매핑할 수 없다.

통계 기반 기법은 패킷 크기, 패킷의 도착 시간 간격과 플로우의 지속 시간 등의 통계 특징들(features)에 기초하여 인터넷 트래픽을 분류한다[10]. 이 기법은 각 응용 프로그램에 의해 생성된 트래픽 플로우는 고유한 특성을 나타낸다고 가정한다[11].

플로우의 특징 유형이 다양해짐에 따라 수동적으로 특징들과 플로우의 클래스(class)를 매핑하는 것은 더 어려워졌다. 따라서, 기계학습(machine learning) 알고리즘을 통해 플로우들의 특징값에 따라 플로우의 클래스를 자동으로 표시하는 통계 기반 분류기가 제시되었다. 현재 많이 사용되는 기계학습 알고리즘은 k-최근접 이웃(k-nearest neighbors), 인공 신경망(artificial neural network), 지지 벡터 머신기계(support vector machine), 의사 결정 트리(decision tree), 규칙 학습자(rule learner) 및 나이브 베이즈(naïve bayes) 등이다. 본연구의 선행연구로 위에서 제시한 여섯 가지 기계학습 알고리즘을 분석하여 의사 결정 트리 알고리즘이 P2P 트래픽 분류에 가장 적합함을 보였다[12].

의사 결정 트리 알고리즘은 트리 성장(tree-growing) 및 가지치기(tree-pruning)의 두 단계를 통해 트래픽 분류를 위한 트리를 형성한다. 트리 성장 단계는 분리 기준(splitting criteria)에 의해 트리를 생성한다. 가지치기 단계는 의사 결정 트리 분류기의 복잡성을 감소시키고 예측 정확도를 높이도록 트리의 일부분을 제거하여 트리의 크기를 줄인다[13].

의사 결정 트리에서 과적합화(overfitting) 문제를 극복하기 위해 가지치기 알고리즘을 사용한다. 비용-복잡도 가지치기(cost-complexity pruning, CCP) 및 에러-감소 가지치기(reduced-error pruning, REP) 등의 기술은 과다-가지치기(over-pruning)하는 경향이 있고 정확도는 좀 낮다. 반면에 에러 기반 가지치기(error-based pruning, EBP) 같은 기술은 과소-가지치기(under-pruning)하는 경향이 있다[14].

표 1. 의사 결정 트리들의 비교
Table 1. Comparison of Decision Trees

항목	C4.5	CART	REPTree
분리기준	Gain Ratio	Gini Index	Information Gain
가지치기 알고리즘	EBP	CCP	REP
가지치기 효과	Under	Over	Over
적합화 효과	Over	Under	Under

CART, C4.5, REPTree 등의 의사 결정 트리 알고리즘이 트리를 생성하기 위해 사용된다. 표 1에 이 세 가지 알고리즘의 특성을 비교한다.

기존 연구에 의하면 대부분의 경우에 의사 결정 트리의 생성에 사용되는 분리 기준의 선택은 분류기의 성능 개선에 큰 영향이 없는 것으로 분석되었다[15]. 또한, 이런 분리 기준은 본질적으로 불안정하므로 트리 분류의 정확도는 학습 데이터의 변화에 민감하다[14].

의사 결정 트리가 갖는 성장 단계의 불안정을 극복하기 위해 bagging, boosting 및 random subspace(RS) 등의 앙상블(ensemble) 알고리즘이 제안되었다. 이 알고리즘들은 원래의 훈련 데이터 집합(training dataset)을 수정하여 생성된 여러 개의 데이터 집합들로부터 기본 기계학습 알고리즘을 통해 여러 개의 기본 분류기를 생성한다. 마지막 단계에서 다수 투표(majority voting)와 같은 기법을 통해 기초 분류기들의 분류 결과를 결합하여 최종 분류한다[16].

앙상블 알고리즘의 성공은 여러 기본 분류기들이 서로 충분히 상이한 가의 여부에 달려 있다. 즉, 한 기본 분류기에서 오류가 발생한 경우, 다른 기본 분류기에서는 동일한 오류를 발생할 가능성이 낮아야 한다[17]. 표 2는 bagging, boosting 및 RS의 앙상블 알고리즘을 비교분석한다.

많은 경우에서 boosting이 좋은 결과를 보이지만, RS는 bagging보다 성능이 낮고 어떤 경우에는 boosting보다 더 우수하다고 제시되었다[18]. Boosting은 노이즈(noise)가 없는 데이터를 처리할 때 bagging과 RS보다 성능이 더 우수하다. 그러나 노이즈가 있는 경우에는 bagging과 RS가 boosting보다 훨씬 더 우수한 것으로 분석되었다[17].

특징 집합 중에서 랜덤(random)으로 부분집합(subset)을 선택하는 RS는 많은 특징과 샘플(sample)을 가진 데이터 집합을 잘 다룰 수 있다[18]. 또한, 선택한 특징들만을 저장

하기 때문에 RS는 훨씬 적은 메모리가 필요하다[19]. Boosting은 순차적으로 훈련하는 반면에, bagging과 RS는 병렬 훈련을 수행하므로 시간을 절약할 수 있다[20]. 위의 분석을 기반으로 RS는 boosting과 bagging보다 우수한 것으로 제시된다.

표 2. 앙상블 알고리즘들의 비교
Table 2. Comparison of Ensemble Algorithms

항목	Bagging	Boosting	Random Subspace
데이터 집합 수정 기술	Random sampling	Weighted	Feature vector subspace
훈련 방식	Parallel	Serially	Parallel
결정 규칙	Simple majority voting	Weighted majority voting	Simple majority voting
노이즈 극복	Good	Bad	Good
많은 특징들의 샘플처리 능력	Bad	Bad	Good
메모리 절약	Medium	Low	High
훈련 속도	Medium	Low	Fast

앞에서 소개된 통계 기반 기법은 P2P 트래픽을 잘 분류할 수 있지만, 트래픽을 특정 P2P 응용 프로그램으로 정확하게 매핑할 수 없으며 온라인 상황에서는 잘 작동하지 않고 계산량이 많다는 제한점이 있다.

기존 기법들의 분석에서 보듯이, 한 가지의 기법을 적용하여 P2P 트래픽을 분류하기는 어려우므로, 위의 방법들을 결합한 분류 기법들이 제안되었다.

Zhen-Xiang Chen 등은 두 단계로 구성된 복합 분류기를 제안했다. 첫 번째 단계에서는 static features 기반의 하드웨어 분류기를, 두 번째 단계에서는 flexible neural tree 기반의 소프트웨어 분류기를 사용하며 95.67%의 정확도를 보였다[21]. Jun Li 등은 coarse-grain 분류 단계 및 fine-grain 분류 단계로 구성되는 복합 분류기를 제안하였고, 정확도는 96.03%였다[22]. Ram Keralapura 등은 time correlation metric (TCM) 및 시그니처 기반 기법을 사용하는 복합 기법을 제안하고, 그 정확도는 95%였다[23]. 본 연구의 선행연구에서는 패킷 레벨에서 시그니처 기법 및 connection heuristics를 사용하고 플로우 레벨에서 C4.5를 적용한 두 단계 기법을 제안하였고, 또한 pattern heuristics를 추가한 기법도 소개하였다[24, 12].

그러나 기존의 하이브리드 기법은 두 번째 단계에서 통계 기반 기법의 특성 때문에 계산량이 많다는 취약점을 가지며, 정확도를 개선하고 계산량을 줄일 필요가 있다.

III. 제안 시스템

본 장에서는 기존 하이브리드 기법에 적용되는 통계 기반 기법의 정확도를 높이고 계산량을 줄이기 위해 그림 1과 같은 P2P 트래픽 분류 시스템을 제안한다.

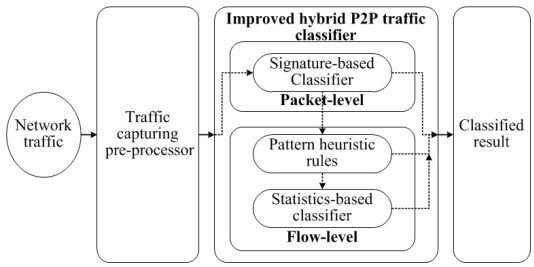


그림 1. 제안 시스템 구조
Fig. 1. Proposed System Architecture

제안 시스템은 트래픽 캡처 처리기(traffic capturing pre-processor) 및 개선된 하이브리드 P2P 트래픽 분류기(improved hybrid P2P traffic classifier)로 구성된다. 검증을 위해 본 연구에서는 제안 시스템을 직접 프로그래밍한다. 휴리스틱 규칙과 분류의 전체 과정은 Java언어로 구현하였는데, Jpcap은 네트워크 패킷을 캡쳐하기 위한 java용 라이브러리이고[25], Weka는 데이터 전처리, 분류, 회귀 등 데이터 마이닝 툴을 포함하는 오픈소스 소프트웨어이다[26]. 패킷 캡쳐기는 이와 관련된 API를 제공하는 JPcap를 이용하여 구현하고, 통계 기반 기법의 알고리즘(REPTREE, RS 등)은 Weka를 이용하여 구현한다.

트래픽 캡처 처리기는 네트워크에서 패킷을 캡처하고, 패킷의 유효성을 검사한 후에 불필요한 패킷을 필터링한다.

하이브리드 P2P 트래픽 분류기는 패킷 레벨 분류기 및 플로우 레벨 분류기로 구성된 두 단계를 통해 트래픽을 P2P 및 non-P2P로 분류하며, 이들의 특성은 다음 소절들에서 제시된다.

3.1 첫 번째 단계 분류기

분류기의 첫 번째 단계는 패킷 레벨의 시그니처 기반 분류기이다. 표 3은 일부 P2P 응용 트래픽의 시그니처이다. 첫 번째 단계에서 분류되지 않는 트래픽은 두 번째 단계에서 분류한다.

3.2 두 번째 단계 분류기

두 번째 단계는 패턴 휴리스틱 규칙 및 통계 기반 분류기로 플로우 레벨에서 동작한다. 분류 과정은 그림 2과 같다. 본 제안에서는 플로우 레벨 분류기에 패턴 휴리스틱 규칙을 추가하여 분류의 정확도를 개선하고 통계 기반 분류기의 계산량을 줄인다.

표 3. P2P 트래픽의 시그니처
Table 3. Signatures in P2P Traffic

응용 유형	시그니처 문자열
BitTorrent	"0x13BitTorrent protocol" "Get announce? info_hash="
eMule/ eDonkey2000	"0xe3", "0xc5", "0xd4"
KazaA	"X-kazaa"
Gnutella	"GNUTELLA" ,"GNUT", "GIV"
MP2P	"GO!!", "MD5", 'SZZ0x20"
DirectConnect	"$MyN", "$Dir", "$SR"
Fasktrack	'GIVE/ . Hash'
Ares	"GET hash:"

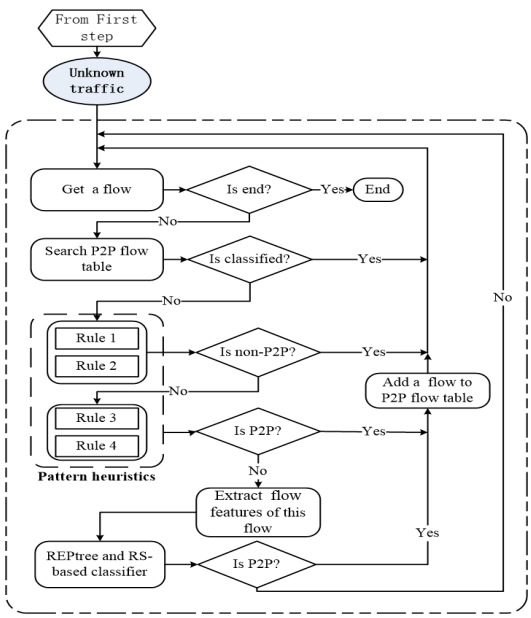


그림 2. 두 번째 단계의 분류 과정
Fig. 2. Process of Second Step Classification

3.2.1 패턴 휴리스틱 규칙

P2P 플로우는 각 프로그램에 따라 다음과 같은 특수한 특성 패턴을 갖는다.

FTP 플로우는 P2P 플로우와 특성이 비슷하기 때문에, 때때로 P2P로 분류된다. 하지만 페이로드를 가진 첫 번째 패킷의 방향과 같은 방향의 페이로드를 가진 패킷들의 ACK 플래그를 이용하여 잘못된 오탐을 수정할 수 있다.

암호화된 BitTorrent의 플로우는 110 바이트와 122 바이트의 크기의 두 패킷이 그 플로우의 처음 10개의 패킷에 포함될 수 있다. 또한, 암호화된 emule의 플로우는 65 바이트와 109 바이트의 크기의 두 패킷이 그 플로우의 처음 10개의 패킷에 포함될 수 있다.

위에서 제시된 특수한 패킷 패턴을 기반으로, 다음과 같은 휴리스틱 규칙을 정의할 수 있다.

규칙 1: 페이로드를 가진 첫 번째 패킷의 방향이 역방향인 경우에는, 이 플로우를 non-P2P로 분류한다.

규칙 2: 같은 방향의 페이로드를 가진 모든 패킷들의 ACK 플래그의 값이 1에 해당하는 경우에는, 이 플로우를 non-P2P로 분류한다.

규칙 3: 페이로드를 가진 처음 10 개의 패킷 중에 110 바이트와 122 바이트의 크기의 두 패킷이 함께 존재한다면, 이 플로우를 P2P로 분류한다.

규칙 4: 페이로드를 가진 처음 10 개의 패킷 중에 65 바이트와 109 바이트의 크기의 두 패킷이 함께 존재한다면, 이 플로우를 P2P로 분류한다.

3.2.2 플로우 특징 선정

플로우의 특징은 일반적으로 플로우에 속하는 여러 패킷들로부터 얻은 수치적인 특징이다[27]. 관련 연구의 분석과 실습을 기반으로, 본 연구에서는 플로우의 페이로드 크기에 대한 통계(최소, 평균, 최대 및 표준 편차)와 페이로드를 가진 첫 번째 패킷의 크기를 플로우의 특징으로 선정하였다[12].

3.2.3 의사 결정 트리

의사 결정 트리 생성 알고리즘의 선정을 위해 앞장에서 설명된 C4.5, CART, 및 REPTree의 세 가지 대표적인 의사 결정 트리 알고리즘을 비교하여 분석하였는데, 그 결과는 그림 3와 같다.

C4.5의 가지치기 알고리즘은 과소-가지치기의 경향이 있기 때문에, C4.5에서 과적합화 문제가 REPTree나 CART보다 더 심각하다. 따라서, C4.5에 의해 생성된 의사 결정 트리는 성능이 낮은 편이다. CART의 CCP와 REPTree의 REP는 과다-가지치기의 경향이 있는데, CCP는 REP보다 과다-가지치기가 더 심한 것으로 분석된다. 따라서, REPTree는

CART보다 높은 성능을 제공한다.

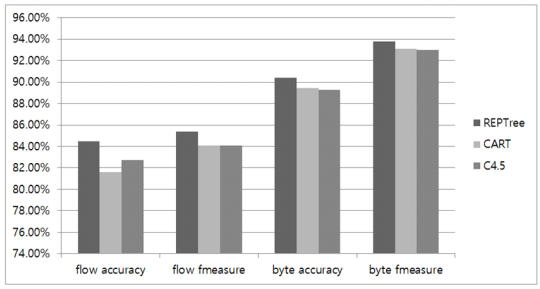


그림 3. 의사 결정 트리의 비교
Fig. 3. Comparison of Decision Trees

3.2.4 앙상블 알고리즘

Bagging, boosting 및 RS의 세 가지 앙상블 알고리즘을 C4.5, CART, 및 REPTree의 의사 결정 트리 알고리즘에 적용한 결과는 그림 4, 그림 5, 그림 6와 같다.

Bagging 및 boosting은 CART, REPTree의 성능에는 영향이 적지만, C4.5의 성능을 개선한다. RS를 CART, REPTree, 및 C4.5에 적용하면 원래보다 정확도가 더 높게 된다. 결과적으로, bagging 및 boosting은 과소-가지치기의 의사 결정 트리의 성능 개선에 도움이 된다. RS는 과소-가지치기의 경향이 있는 트리나 특히 과다-가지치기의 경향이 있는 트리에 대해 더 좋은 결과를 얻을 수 있으며 안정한 성능을 제공한다.

따라서, 본 연구에서는 통계 기반 분류기를 REPTree와 RS로 구현한다.

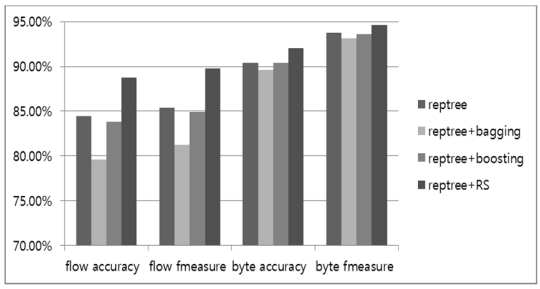


그림 4. REPTree 및 앙상블 알고리즘
Fig. 4. REPTree with Ensemble Algorithms

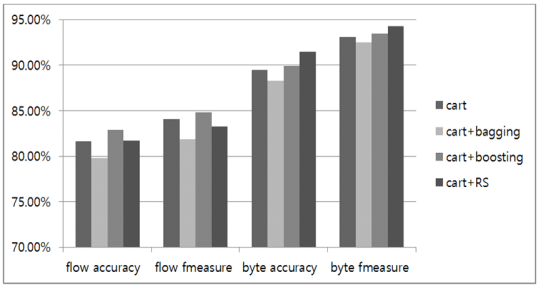


그림 5. CART 및 앙상블 알고리즘
Fig. 5. CART with Ensemble Algorithms

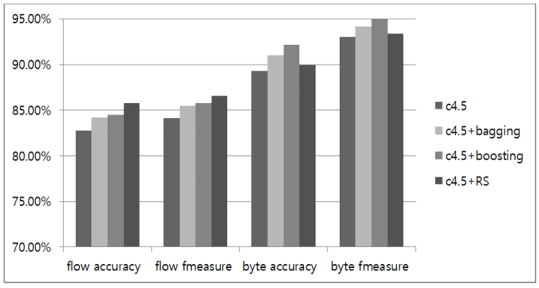


그림 6. C4.5 및 앙상블 알고리즘
Fig. 6. C4.5 with Ensemble Algorithms

IV. 제안 분류 기법의 검증 분석

4.1 평가 기준

제안 분류기에 의한 분류는 표 4와 같은 네 가지의 가능한 결과가 있다[28].

표 4. 평가 기준
Table 4. Evaluation Metrics

		예측 클래스(Predicted class)	
		P2P	Non-P2P
실제 클래스 (Actual class)	P2P	true positive(tp)	false negative(fn)
	Non-P2P	false negative(fp)	true negative(tn)

본 연구에서는 다음과 같은 두 가지 기준을 평가에 적용한다.

$$\text{정확도(accuracy)} = \frac{tp+tn}{tp+fn+fp+tn}$$

$$\text{f-측도(f-measure)} = \frac{2\times recall\times precision}{recall+precision}$$

(recall = tp/(tp+fn), precision = tp/(tp+fp))

4.2 데이터 집합

표 5에 보인 네 개의 데이터 집합(UNIBS, DKU1, DKU2 및 DKU3)을 제안 기법의 평가에 사용한다. UNIBS는 Brescia 대학에서 제공하는 트래픽 데이터 집합이며, Ground Truth 시스템으로 수집되었다[29]. DKU1, DKU2 및 DKU3은 페이로드를 가진 데이터 집합이며 단국대학교 내의 제어된 환경에서 수집되었고, 각 트래픽은 실제 응용 프로그램의 유형으로 표시되었다.

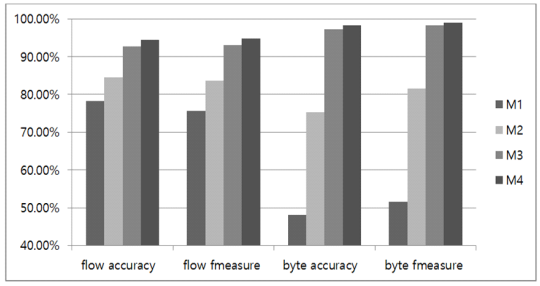
표 5. 데이터 집합
Table 5. Dataset

	UNIBS	DKU1	DKU2	DKU3
Non-P2P	53279	11247	13190	11860
P2P	21716	15734	14464	23988
Total	74995	26981	27654	35848

4.3 실험 결과 및 분석

제안 기법의 검증을 위해, 기본 시그니처 기반 분류기(M1)를 기본 모델로 하고, 이에 기능을 추가한 M2, M3 및 M4에 대해 성능을 비교하여 분석하였는데 그림 7에 각 모델에 대한 결과를 보인다. M4가 본 연구의 제안 기법이다.

암호화된 페이로드를 가진 패킷이 많이 있기 때문에 M1은 정확도와 f-측도가 가장 낮다. M2는 M1에 패턴 휴리스틱 규칙을 추가하여 FTP 플로우와 일부 암호화된 BitTorrent 및 eMule의 플로우를 분류한다.



M1: 시그니처 기반 기법
M2: M1에 패턴 휴리스틱 규칙 적용한 기법
M3: M2와 REPTree를 이용한 통계 기반 기법
M4: M3 및 RS 앙상블 알고리즘을 이용한 기법

그림 7. M1,M2,M3 및 M4 기법의 비교
Fig. 7. Comparison of M1,M2,M3 and M4 schemes

M3은 M2에 통계 기반 기법을 추가하여 정확도와 f-측도를 개선한다. 페이로드 크기 특징은 모든 P2P 트래픽에 동일하지 않기 때문에, M3은 모든 P2P 트래픽을 탐지할 수 없다. 제안 기법인 M4는 M3에 앙상블 기법을 적용하여 다른 세 가지 기법보다 성능이 더 높다.

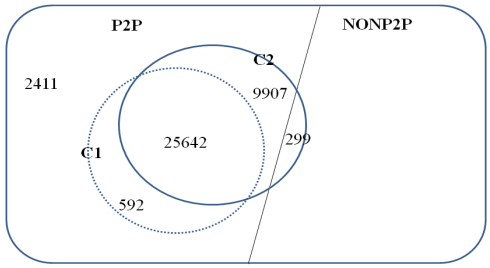
플로우 레벨에서 분류의 계산량을 평가하기 위해 각 기법에서의 처리한 패킷의 수와 플로우의 수를 분석하였는데, 그 결과는 표 6과 같다. 순수한 통계 기반 기법인 S1, 시그니처 기반 기법와 S1을 직접 결합한 하이브리드 기법인 H1, H1에 패턴 휴리스틱 규칙을 추가한 H2의 세 기법의 계산량을 비교하였다. 패턴 휴리스틱 규칙이 추가된 H2는 H1에 비해 통계 기반 분류기의 특징 수집과정에서 처리한 패킷의 수를 67.95%까지 감수시키고, 분류한 플로우의 수도 62.90%까지 크게 감소시켰다.

표 6. 계산량의 비교
Table 6. Computation Comparison

	플로우 레벨	
	특징 수집 과정에서 처리된 패킷의 수	분류한 플로우의 수
S1	11567317	63502
H1	8363862	37368
H2	3707640	23557
H1에 의한 감소 (%)	27.69	41.15
H2에 의한 감소 (%)	67.95	62.90

S1: 통계 기반 기법
H1: S1과 시그니처 기반 기법을 직접 결합한 하이브리드 기법
H2: H1에 패턴 휴리스틱 규칙을 적용한 기법

그림 8은 제안 기법의 두 단계 분류기가 서로 보완적임을 보여준다. 시그니처 기반 분류기는 알려지지 않은 패킷 또는 새 유형의 패킷을 감지할 수 없으며, 통계 기반 분류기는 훈련받지 않은 트래픽을 분류할 수 없다. 따라서 그림 7과 같이, 두 분류기는 각각 다른 분류기가 분류하지 못한 트래픽을 감지하여 두 단계 분류기에서는 높은 정확도를 달성할 수 있다.



C1: 첫 단계 분류기가 분류한 P2P 플로우들
C2: 두 단계 분류기가 분류한 P2P 플로우들
C1UC2: 제안 분류기가 분류한 P2P 플로우들

그림 8. 두 분류기의 상호보완성
Fig. 8. Complementary Operation of Two Classifiers

IV. 결 론

P2P 응용 프로그램은 편리함을 제공하지만 트래픽을 증가시키므로 네트워크상의 대역폭 문제 및 보안 문제를 유발한다. 정확한 P2P 트래픽 분류는 이런 문제를 해결하여, 인터넷 서비스 공급자 및 non-P2P 사용자에게 도움이 된다.

본 논문에서는 기존의 P2P 트래픽 분류 기법의 분석을 바탕으로, 개선된 휴리스틱 규칙 및 기계학습의 분석 결과를 적용한 두 단계 하이브리드 기법을 제안한다. 첫 번째 단계는 패킷 레벨에서 동작하는 시그니처 기반의 분류기이며, 두 번째 단계는 플로우 레벨에서 동작하는 패턴 휴리스틱 규칙 및 통계 기반의 분류기이다. 본 제안은 패턴 휴리스틱 규칙을 통해 분류 정확도를 높이고 통계 기반 분류기가 처리할 트래픽의 양을 줄인다. 또한 여러 유행의 의사 결정 트리 알고리즘의 분석을 기반으로, REPTree를 활용하는 통계 기반 분류기를 제안하고, 앙상블 알고리즘을 이용하여 통계 기반 분류기의 성능을 높인다.

제안 기법의 특성은 다음과 같이 분석되었다.

1) 시그니처 기반 기법과 통계 기반 기법은 상호 보완적인 탐지를 통해 높은 정확도를 달성할 수 있다.
2) 패턴 휴리스틱 규칙을 통해 통계 기반 기법의 처리할 트래픽의 양을 줄일 수 있다.
3) 앙상블 알고리즘을 이용하여 통계 기반 분류기의 성능을 높일 수 있다.

본 제안에서 첫 번째 단계에서는 특정 P2P 트래픽을 식별할 수 있지만, 두 번째 단계에서는 트래픽을 P2P 및 non-P2P로만 분류한다. 따라서, 향후 연구로 특정 P2P 트래픽을 분류할 수 있는 통계 기반 분류기를 제안한다. 더 많은 패턴 휴리스틱 규칙이 개발하고 P2P의 UDP 트래픽을 포함하도록 확장할 것이다.

참고문헌

〔1〕 Myung-Yoon Lee, Jang-Su Park and Im-Yeong Lee, "SPNS realization for secure P2P Service," Korea Multimedia Society, pp. 67-70, Nov. 2006.

〔2〕 Jaehak Yu, Hansung Lee, Yuonghee Im, Myung-sup Kim and Daihee Park, "Hierarchical Internet Application Traffic Classification using a Multi-class SVM," Korean Institute of Intelligent Systems, Vol. 20, No. 1, pp. 7-14, Oct. 2010.

〔3〕 Nam-Kyoung Um, Sung-Hee Woo and Sang-Ho Lee, "Flow-based P2P traffic identification using SVM," Vol. 13, No. 3, pp. 123-130, May 2008.

〔4〕 Yu-Shui Geng, Tao Han and Xue-Song Jiang, "The Research of P2P Traffic Identification Technology," Proc. of International Conference on E-Business and Information System Security, Wuhan, pp. 1-4, May 2009.

〔5〕 Subhabrata Sen, Oliver Spatscheck and Dong-Mei Wang, "Accurate, scalable in network identification of P2P traffic using application signature," Proc. the 13th international conference on World Wide Web, New York, pp. 512-521, May 2004.

〔6〕 Alok Madhukar and Carey Williamson, "A longitudinal study of P2P traffic classification," Proc. 4th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, pp. 179-188, Sept. 2006.

〔7〕 Xin-Bin Liu, Jian-Hua Yang, Gao-Gang Xie and Yao Hu, "Automated mining of packet signatures for traffic identification application layer with apriori algorithm," Journal on Communications, Vol. 29, No. 12, pp. 51-59, March 2008.

〔8〕 G´eza Szab´o, D´aniel Orincsay, Szabolcs Malomsoky, and Istv´an Szab´o, "On the Validation of Traffic Classification algorithms," Passive and Active Network Measurement Lecture Notes in Computer Science, pp. 72-81, April 2008.

〔9〕 Thomas Karagiannis, Andre Broido, Michalis Faloutsos and Kc Claffy, "Transport layer identification of P2P traffic," Proc. the 4th ACM SIGCOMM Conference on Internet Measurement, NewYork, pp. 121-134, Oct. 2004.

〔10〕 Yaou Zhao, Xiao Xie and Mingyan Jiang,

"Hierarchical real-time network traffic classification based on ECOC," TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol. 12, No. 2, pp. 1551-1560, Feb. 2014.

〔11〕 Aiqing Zhu, "A P2P Network Traffic Classification Method Based on C4.5 Decision Tree Algorithm," Proc. of the 9th International Symposium on Linear Drives for Industry Applications, Vol. 4, pp.373-379, Jan. 2014.

〔12〕 Wujian Ye and Kyungsan Cho, "Hybrid P2P traffic classification with heuristic rules and machine learning," Soft Computing Journal (to be published)

〔13〕 Pruning, http://en.wikipedia.org/wiki/Pruning_(decisi on_trees)

〔14〕 S. B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, Vol. 39, No. 4, pp. 261-283, April 2013.

〔15〕 Oded Maimon and Lior Rokach, "Data Mining and Knowledge Discovery Handbook," Second Edition, Springer, 2010.

〔16〕 Marina Skurichina and Robert P. W. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers," Pattern Analysis and Applications, Vol. 5, No. 2, pp. 121-135, June 2002.

〔17〕 S. Kotsiantis, "Combining bagging, boosting, rotation forest and random subspace methods," Artificial Intelligence Review, Vol. 35, No. 3, pp. 223-240, March 2011.

〔18〕 Tin Kam Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 8, pp. 832-844, Aug. 1998.

〔19〕 Robert E. Banfield, Lawrence O. Hall, Kevin W, Bowyer and W. P. KegeImeyer, "A comparison of decision tree ensemble creation techniques," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 1, pp. 173-180, Jan. 2007.

〔20〕 Tianyan Jiang, Jian Li, Yuanbing Zheng and Caixin Sun, "Improved Bagging Algorithm for Pattern Recognition in UHF Signals of Partial Discharges," Energies, Vol. 4, No. 7, pp. 1087-1101, April 2011.

〔21〕 Zhen-Xiang Chen, Bo Yang, Yue-Hui Chen, Ajith Abraham, Crina Grosan and Li-Zhi Peng, "Online Hybrid traffic classifier for Peer-to Peer Systems based on Network Processors," Applied Soft Computing, Vol. 9, No. 2, pp. 685-694, Mar. 2009.

〔22〕 Jun Li, Shui-Yi Zhang, Yan-Qing Lu and Jun-Rong Yan, "Hybrid Internet Traffic Classification Technique," Journal of Electronics (China), Vol. 26, No. 1, pp. 101-112, Jan. 2009.

〔23〕 Ram Keralapura, Antonio Nucci and Chen-Nee Chuah, "A novel self-learning architecture for p2p traffic classification in high speed networks," Computer Networks, Vol. 54, No. 7, pp. 1055-1068, May 2010.

〔24〕 Wujian Ye and Kyungsan Cho, "Two-Step P2P Traffic Classification with Connection Heuristics," Proc. of IMIS2013-Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp.135-141, July 2013.

〔25〕 JPcap, http://www.eden.rutgers.edu/~muscarim/jpcap/index.html

〔26〕 Weka, http://www.cs.waikato.ac.nz/ml/weka/

〔27〕 Thuy T. T. Nguyen and Grenville J. Armitage, "A survey of techniques for internet traffic classification using machine learning," Proc. of IEEE Communications Surveys and Tutorials, Vol. 10, No. 4, pp. 56-76, Fourth Quarter 2008.

〔28〕 Precision and recall, http://en.wikipedia.org/wiki/Recall_and_precision

〔29〕 F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso and K.C. Claffy, "GT: picking up the truth from the ground for Internet traffic,"ACM SIGCOMM Computer Communication Review, Vol. 39, No. 5, pp. 12-18, Oct. 2009.

저 자 소 개

예 우 지 엔

2010: 중국광동공업대학교 컴퓨터과학 및 기술학과 (공학사)
2012: 단국대학교 컴퓨터학과 (공학석사)
2012~현 재: 단국대학교 컴퓨터학과 (박사과정)
관심분야: 네트워크시스템
Email : yewujian@dankook.ac.kr

조 경 산(교신저자)

1979: 서울대학교 전자공학과 (학사)
1981: 한국과학원 전기전자공학과 (공학석사)
1988: 텍사스 대학교 (오스틴) 전기전산공학과 (Ph.D.)
1988~1990: 삼성전자 컴퓨터부문 책임연구원, 실장
1990~현 재: 단국대학교 소프트웨어학과 교수
관심분야: 네트워크시스템 및 이동통신 보안, 컴퓨터시스템
Email : kscho@dankook.ac.kr