Fit-FA Finder를 이용한 SMBC 플랫폼 설계

A design of the SMBC Platform using the Fit FA-Finder

박노경*, 한성호*, 서상진*, 진현준*
Nho-Kyung Park*, Sung-Ho Han*, Sang-Jin Seo*, Hyun-Joon Jin*

Abstract

Recently, e-mail has become an important way of communications in IT societies, but it creates various social problems due to increase of spam mails. Even though many organizations and cooperation have been trying researches to develop spam mail blocking technologies, a lot of cost and system complexities are required because of varieties of spam blocking technologies. In this paper, we designed of the SMBC(Spam Mail Blocking Center) using the Fit FA(Filtering Algorithm) Finder. Fit-FA Finder that search and applises spam mail filtering algorithm of the most suitable confrontation according to type of spam mail. The system of spam mail filtering is decided performance of the system by procedure that spam filter is used. Go through designed Fit-FA Finder and reduced unnecessary filtering process and processing time and load than appointment order filter application way of existent spam mail interception system.

요 약

최근 전자 우편은 IT 사회의 중요한 의사소통의 수단이 되고 있다. 그러나 스팸 메일의 증가로 인해 다양한 사회 문제가 발생되고 증가하는 추세이다. 스팸 메일을 차단하기 위해 정부와 민간 단체에서 많은 연구와 개발을 하고 있으나 다양한 스팸 메일의 증가로 인해 많은 비용과 시스템의 복잡성이 요구되어 지고 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 Fit-FA Finder를 이용한 SMBC(Spam Mail Blocking Center)를 설계 하였다. Fit-FA Finder는 스팸 메일의 유형에 따라 필터링 데이터베이스에서 적절한 알고리즘을 적용 시키는 시스템으로서 필터의 적용 순서에 따라 스팸 메일 오인율(False-Positive Error)이 달라져 시스템 처리 신뢰도에 큰 영향을 준다. 본 논문에서 제안한 Fit-FA Finder를 이용한 SMBC 플랫폼은 불필요한 필터링 처리 과정을 줄임으로써 시스템의 부하를 줄 일수 있다.

Keywords : spam mail, Fit-FA Finder, SMBC, 신뢰도, 오인율

Ⅰ. 서론

인터넷은 현대 생활에서 없어서는 안 될 정도로 큰 부분을 크게 차지하고 있다. 보편화된 인터넷을 통해 우리는 여러 정보를 쉽게 모을 수 있지만, 우리가 원하지 않는 정보도 참조하게 되는 단점도 있다. 원하지 않는 정보를 보내는 사람은 핸드폰이나, 전자우편 같은 방법을 활용하여 무차별적으로 정보를 전송하여,

* 호서대 정보통신공학과
(Dept. of Information Technology and Communication Engineering, Hoseo University)
接受日:2006年 2月 9日, 修正完了日: 2006年 6月 28日

※ 본 과제(결과물)는 교육인적자원부와 산업자원부의 출연금 및 보조금으로 수행한 산학협력 중심대학 육성사업의 연구결과입니다.

사회 문제로 대두 되고 있다[1].

특히, 스팸 메일은 현재 심각한 사회적 문제로 대두되면서 스팸 메일 차단 시스템 및 다양한 솔루션 등 새로운 기술에 대한 연구가 계속 되고 있지만, 새로운 유형의 스팸 메일이 등장하면서 그 폐해가 당분간 계속 될 것으로 보인다[2][3].

스팸 메일 차단 필터링 기술은 다양한 유형에 따라 개별적인 대응을 각각 달리 수행해야 하므로 시스템의 복잡도와 많은 처리 비용 및 노력을 요구한다. 필터의 적용 순서에 따라 스팸 메일 오인율(False-Positive Error)이 달라져 시스템 처리 신뢰도에 큰 영향을 끼친다.

본 논문에서는 스팸 메일의 유형에 따라 대응 가능한 최적의 스팸 메일 차단 알고리즘을 검색 및 적용하기 위한 Fit-FA(Filtering Algorithm) Finder를 이용한 SMBC(Spam Mail Blocking Center)를 설계 하였다. 본론에서는 1절 기존의 스팸 메일 차단 시스템 유형 및 스팸 메일 대응 기술한다. 2절에서는 Fit-FA Finder를 적용한 SMBC 플랫폼을 설계하고, 3장에서는 결론 및 향후 연구 과제에 대해 기술한다.

II. 본론

1. 스팸 메일 유형과 처리 알고리즘

본 장에서는 현재 스팸 메일 차단을 위한 시스템 유형을 살펴보고, 본 논문에서 설계한 SMBC에 적용될 Fit FA-Finder 기법을 분석 하였다.

1.1 스팸 메일 차단 시스템 유형

스팸 메일 차단 시스템은 적용 분야 구현 기술을 기준으로 구분된다. 적용 분야에 따른 차단 시스템 유형은 서버형과 클라이언트형으로 구분된다. 표 1은 Server 기반의 스팸 차단 시스템의 특징을 나타낸 것이다.

클라이언트 기반의 스팸 메일 차단 시스템은 사용자의 PC에 설치되어 메일 서버에 일정 시간마다 접근하여 메일 박스로부터 수신된 스팸 메일을 삭제하는 기술이다. 구현 기술에 따라 프록시 서버형과 에이전트(Agent)형으로 구분된다. 표 2는 클라이언트 기반의 스팸 차단 시스템의 특징을 나타내고 있다.

프록시 서버형은 클라이언트 PC에 설치된 프로그램이 프록시 서버의 역할을 수행한다. 일정 시간마다 사용자의 메일 서버에 접근하여 메일을 수집한 후 스팸 메일을 필터링한다. 이러한 기술은 사용자 입장에서 스팸 메일이 직접 사용자에게 도달하는 가능성을 원천적으로 차단되는 장점이 있다. 그러나 소프트웨어 자체의 용량이 크며, 웹 메일에 대해서 필터링 작업을 수행할 수 없다는 단점이 있다.

표 1. 서버 기반의 스팸 차단 시스템
Table 1. Spam filtering system of based on server

Server 기반의 스팸 차단 시스템	장 점	단 점	비 고
Proxy Server형 스팸 차단 시스템	- 필요시 스팸 메일로 분류한 메일 복구가능 - 스팸 차단기 상에서 메일 트래픽은 양호 특성	- DNS 참조 없이 직접 접근하는 스팸 메일에 대한 무력 - 소극적인 방화벽이나 서버 공격은 Bounce를 통한 메일 서버의 Outbound 메일 직접 유입	- 현재 가장 많이 사용
Gateway형 스팸 차단 시스템	- 모든 메일 트래픽 감시 가능	- System Down 시 전체 시스템 마비 - 필터링된 메일 복구 불가능	- 대부 네트워크의 입부 네트워크를 구분 - 스팸 차단 시스템과 트래픽과 메일 서버 트래픽의 상호 상존적

표 2. 클라이언트 기반의 스팸 차단 시스템
Table 2. Spam filtering system of based on client

Client 기반의 스팸 차단 시스템	장 점	단 점	비 고
Proxy Server형 스팸 차단 시스템	- 사용자 입장에서 스팸메일이 직접 도달하는 가능성을 원천적으로 차단	- 스팸 차단 솔루션의 막대한 자원 소모 - 웹 메일 필터링 불가	- 무위시간 간격으로 사용자 PC의 메일 서버로부터 메일수신 - 메일 클라이언트가 접근하는 메일서버의 정보를 자신의 정보로 교체하여 메일 송신
Agent형 스팸 차단 시스템	- 웹 메일 필터링 가능	- 실시간 필터링 적용 제외 수신된 스팸에 대해서 처리 불가	- 웹 메일 필터링을 위한 스크립 스크래핑 사용

1.2 스팸 메일 유형에 따른 차단 알고리즘

수신된 메일은 스팸 메일 차단을 위한 키워드 변형이나 내용 참조를 막기 위한 형식 변형을 제거한 후, 메일 내용을 분석하여 스팸 메일을 판정한다. 표 3은 Fit-FA Finder가 처리 가능한 스팸 메일 차단 알고리즘과 특징을 보여주고 있다.

보다 효율적인 스팸 메일 차단을 위해서는 휴리스틱 룰 기반의 스팸메일 차단 매커니즘을 동적 처리 기반으로 확장을 하여야 하고, 필터 스크립팅 기술은 노력 및 시간 손실을 최소화 하기 위해 제로-어드민 처리 구조를 도입해야 한다.

표 3. 기존 스팸 메일 차단 알고리즘
Table 3. Spam mail filtering algorithm

구 분	장 점	단 점
Heuristic Rule 기반 스캐닝 기법	대량 메일 발송 차단 RFC 규약 미 준수 차단	정적 처리기반의 기술적 특성으로 한계성 대포
Filter Scripting 기법	높은 정확도	지능화된 DB가 없는 경우 관리자의 노력과 시간 낭비
발신자 주소 유효성 검사 기반의 필터링 기법	메일 주소의 정확성 판단	효율성 떨어짐
Bayesian 필터링 기법	학습통해 피존 개인별 스팸 분류	기존 메일 시스템과의 호환성, 사용성면에서 자원낭비 큼

또한 발신자 주소 유효성 검사기술은 관계형 스팸 메일 발신자 리스트 검색 기법을 이용한 검사 기술로 보완하고, 베이시안 필터링 기술은 확률계산 기법 간소화 및 처리 과정 최적화를 통한 처리 부하를 최소화 시켜야 한다[4][6][7].

2. Fit-FA Finder를 이용한 SMBC 플랫폼 설계

본 장에서는 스팸 메일 차단 기법 중에 본 논문에서 제안한 Fit-FA Finder 이용하여 SMBC 플랫폼을 설계한다.

2.1 Fit FA Finder 설계

Fit-FA Finder는 불필요한 필터링 과정을 줄이기 위해 수신된 메일의 유형을 분석하는 Mail 유형 분석기와 분석된 유형에 따라 스팸 메일 차단 알고리즘을 결정하는 알고리즘 적용기로 구분되며, 그림 1에서 구성 모듈을 나타내고 있다.

메일 유형분석기는 수신된 메일을 OCR(Otical Character Reader)을 통해 순수 텍스트를 추출하고 인코딩 타입을 분석한다. HTML 타입의 경우 HTML tag(s), HTML- clocking를 검사하여 HTML stuffing, Space stuffing를 제거한다. Syntax Noise(s)는 Regular expression을 적용하고, Foreign Language는 한국어로 구성되었는지 확인한다.

Environment Analysis Manager에서는 스팸 메일 필터 DB 분석 요청과 결과 반환이 이루어지고 Mail Analysis Manager에서는 수취 메일 헤더 및 바디 분석이 이루어진다. Filter DB Analysis Agent에서는 스팸 메일 필터 정보 검색 요청과 결과 반환이 이루어지고 스팸 메일 필터 알고리즘 DB는 적용 알고리즘에 관한 정보가 들어 있다.

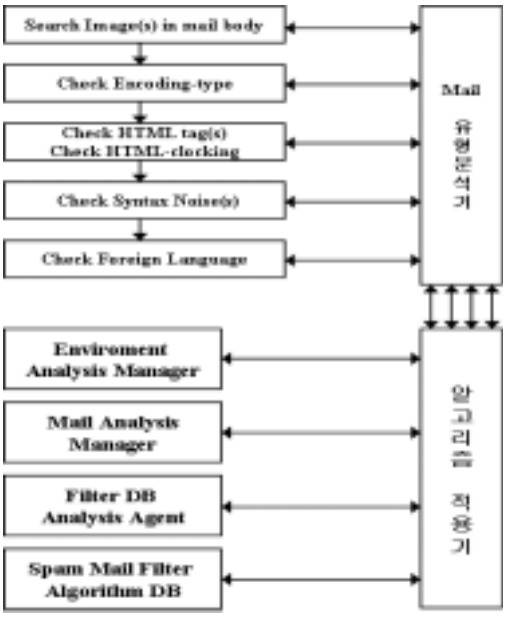

그림 1. Fit-FA Finder의 블록도
Fig 1. The block of Fit-FA Finder

TCP MDA(Mail Delivery Agent)로부터 메일을 수신 받은 후 메일수신 관리자를 통해 Fit FA-Finder로 보내진다. Fit-FA Finder에서는 우선 수신된 메일이 이미지 기반 형태의 메일인지 텍스트 기반 형태의 메일인지 분석 한 후 순수 텍스트 문을 추출한다. 추출된 순수 텍스트문의 분석을 통해 적용할 스팸 메일 차단 알고리즘의 유형과 순서를 결정하고 처리 적용한다.

Fit-FA Finder에 적용할 스팸 메일 필터링 알고리즘은 크게 포맷 검사(Validation) 알고리즘과 사용자의 의도에 따라 차단시킬 Filtering Scripting, 그리고 사용자 취향에 따라 차단 학습을 위한 Bayesian Filtering으로 분류된다. 적용 순서는 포맷 검사를 선행하여 사용자의 Filtering Scripting 적용이 유효하게 처리될 수 있도록 보증한다. 그리고 사용자가 의도한 차단 Filtering Scripting 적용은 False-Positive 메일을 발생시킬 수 있으므로, Bayesian Filtering을 적용하여 오인률을 최소화 할 수 있도록 적용 순서를 지정하였다.

표 4는 Fit-FA Finder가 참조하고 적용 순서를 지정하는 Filtering Algorithm을 간략히 나타내고 있다.

표 4. Fit-FA Finder의 스팸 차단 알고리즘
Table 4. Spam mail filtering algorithm in Fit-FA Finder

algorithm	summary (apply-seq.)
발신자 주소 유효성	Black-White List, DB ref(①)
Heuristic Rule	RFC 규약 준수(②)
Filtering Scripting	사용자 정의 필터링(③)
Bayesian Filtering	이용 성향에 따른 필터링(④)

그림 2에서 Fit-FA Finder의 처리 과정을 알고리즘으로 나타내고 있다.

```
Input : 수신 메일함 내에 i 번째 메일 (mail)i
Output : FFL(Filtering Function List)
Reference : Filtering Algorithm in Table 4
initialized Fit-FA Finder
Fetch a (mail)i From TCP MDA
IF((mail)i ∈ Black-White List) {
   Add algorithm ① to FFL
}
IF(Include image-tags in a body of (mail)i) {
   Get text(s) from image using OCR
}
IF(Is encoding type supported?) {
   reject (mail)i and return
}ELSE {
   Add algorithm ② to FFL
}
IF(Is encoding type not supported language) {
   Marking (mail)i with "no-Validation" and return }
Add algorithm ③ to FFL
remove Syntax Noise(s)
retrieve Bayesian-keyword
IF(Bayesian-keyword > 0) {
   Add algorithm ④ to FFL
}
return FFL
```

그림 2. Fit-FA Finder를 이용한 스팸 메일 필터링 알고리즘
Fig 2. The spam filtering algorithm using of Fit-FA Finder

2.2 SMBC 플랫폼 설계

SMBC(Spam Mail Blocking Center)는 Proxy Server 기반의 스팸 차단 시스템 프레임 레이어로 설계된다. Proxy Server 기반의 차단 프레임은 물리적으로 임의의 위상(Topology)로 구축 가능하여, 플랫폼 구현시 유연한 모듈/구성 레이어 개발이 가능하다. 그림 3는 SMBC 플랫폼의 전체 구성도를 나타내고 있다.

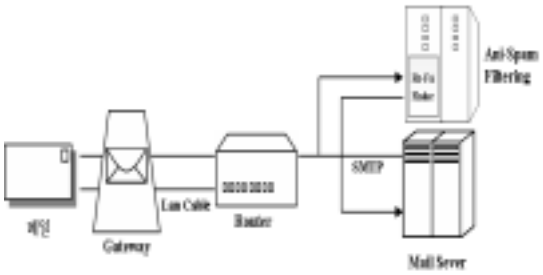

그림 3. SMBC 플랫폼의 전체 구성도
Fig 3. The block of SMBC platform

SMBC는 터널형 스팸 메일 차단 프레임에 비해 필요시 스팸 메일로 분류된 메일이 복구가 가능하다. 그리고 스팸 차단 프레임과 메일 트래픽이 상호 독립되어 네트워크 스트림 채널을 분산시켜 병목현상을 최소화시킬 수 있다.

2.3 기본적인 SMBC 플랫폼 구조 및 처리

본 논문에서 설계되는 SMBC 플랫폼은 크게 스팸 메일 차단 서버와 스팸 메일 보관 시스템, 그리고 스팸 메일 판정을 위한 H/W 가속기로 구분된다. 각 구성 모듈은 S/W와 H/W로 구성되며, 상호 비동기적 연동 체계를 통해 작업을 분산처리 한다. 그림 4는 스팸 메일 처리 구성도를 나타내고 있다.

메일 차단 서버는 TCP/IP 프로토콜을 통해 메일 전송 관리자(MDA)가 메일 수신 관리자에게 메일을 전달한다. 메일 수신 관리자는 전달된
메일을 메일 보관함(Mail Pool)에 보관한다. 스팸 메일 차단기는 수신된 스팸 메일을 검사하고 처리 결과에 따라 스팸 메일 보관함이나 메일 보관함으로 메일을 이동 보관 처리한다. 처리 결과는 스팸메일 필터링 DB에 보관하여 지식 정보를 훈련(Training)시킨다.

신속한 스팸 메일 처리를 위한 포맷 해석 및 판정 관련 처리를 위해 스팸 메일 차단기는 필터링가속기(FA: Filtering Accelerator)와 연동한다. 지식기반의 스팸 판정 정보를 누적하여 특정 메일의 송신자를 추후 메일 송신 가능 여부를 저장하기 위해 Opt-Marker를 이용하여 판정 정보 패턴을 추출 저장한다. 그림 5은 스팸 메일 차단 구성도를 나타내고 있다.

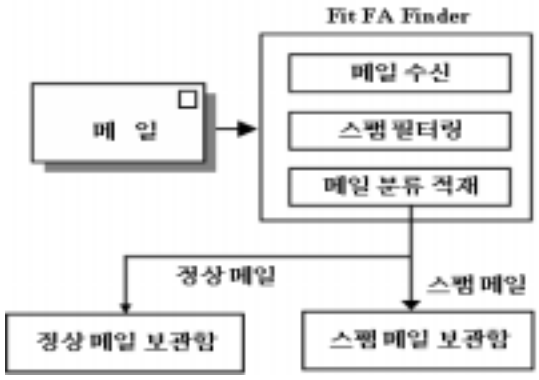

그림 4. 스팸메일 처리 구성도
Fig 4. The processing diagram of SPAM mail

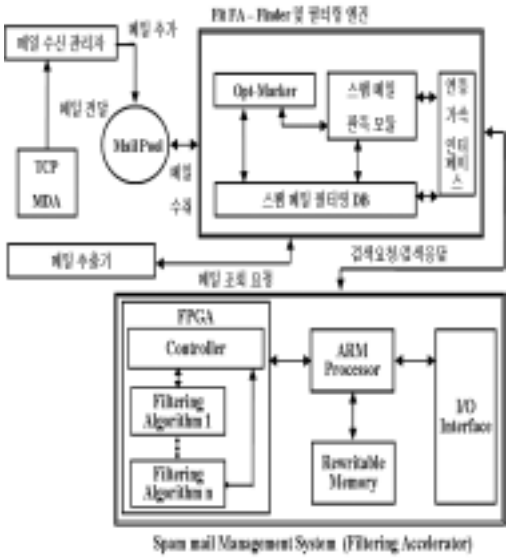

그림 5. 스팸메일 차단 구성도
Fig 5. The filtering block diagram of SPAM mail

III. 결론

본 논문에서는 스팸 메일의 유형에 따라 대응 가능한 최적의 스팸 메일 필터링 알고리즘을 검색하여 적용하는 Fit-FA(Filter Algorithm) Finder를 설계하고 이를 SMBC 플랫폼에 확장 적용 하였다.

적용에 따른 기존 방식의 설계 구조와의 비교를 표 5에서 나타내고 있다.

표 5. 기존 방식과의 비교
Table 5. comparison with referenced method

구분	기존 방식 (①)	[8]의 방식 (②)	본 논문의 방식(③)
네트워크 접속횟수	1회	3회	1회
메일 수신시간	단일 메일 전송	각 메일 subject + body 전송	메일 전송 + 스팸 필터링
스팸공격 대응방식	관리자 직접 감시	Subject 누적만으로 Disk Full에 의한 시스템 마비 시간 지연	Heuristic Rule 적용으로 공격 차단
스팸메일 수신범위	일방적 수신	Subject 만 수신	일방적 수신
허위 주소에 의한 접근성	Open Relay 가능	Open Relay 불가	Open Relay 1회 이상 불가
서버접근의 용의성	직접 접근	메일 Body 추가 접근	Fit FA Finder를 경유한 접근
시스템 확장성	범용화에 따른 유연한 확장	기존 시스템에 메시지 전달 시스템 추가	Fit FA Finder를 기존 시스템에 추가
메시징 인프라 유지 비용 (트래픽, 서버자원)	스팸량에 따라 처리 비용 비례	스팸량이 커질수록 처리 비용 감소	①에 비해 네트워크 트래픽은 동일하나, 서버 자원 유지 보수 비용 감소

기존 스팸 메일 처리 방식(표 4의 ① 방식)은 개별적인 스팸 메일에 대해 입체적 검사를 수행하지 않아 스팸 필터링 정확도가 떨어진다. [8]의 방식(표 4의 ② 방식)은 3번 이상의 네트워크 접속하며, 스팸 공격에 따른 시스템 불능 상태에 빠질 수 있다. 본 논문의 방식(표 4의 ③ 방식)은 입체적 검사를 통한 스팸 필터링 정확도를 높이고, Heuristic Rule 적용을 통해 스팸 공격에 따른 시스템 불능 상태를 막을 수 있으나, Fit-FA Finder 장치 추가로 자원 추가 비용 및 확장성이 떨어진다. 그러나, 다양한 스팸 유형에 대해 입체적 검사 기법을 추가 적용하고 있는 추세[7]에 따라 유사 처리 시스템에 비해 스팸 메일 검사시 메일 유형 분석을 통해 불필요한 검사 과정을 크게 줄일 수 있다. 그리고, 스

팸 메일 검사 필터 적용도 타입 분석에 따라 가장 적절한 알고리즘을 대응 시킬 수 있으므로 차단 성능 및 신뢰도를 향상 시킬 수 있다.

향후 연구 과제는 Fit-FA Finder를 이용한 SMBC 플랫폼 구현 및 성능 평가이다.

참 고 문 헌

[1] 김자경, 이광수"스팸 메일 차단 방법론비교·분석" 한국정보과학회 2004 가을 학술발표논문집 (1),pp.487-489, [2004]

[2] 고주영,심재창,김현기"비밀단어의 회신을 이용한 스팸메일 차단 시스템의 구현"멀티미디어학회논문집 제8권1호.

[3] 김현준, 정재은, 조근식 "가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템" 한국정보과학회논문지, pp.1092-1100, [2004]

[4] 명순희, 조형근, 김인철 "하이브리드 다중 모델 학습 기법을 이용한 자동문서 분류"한국정보과학회 제29권 2호

[5] 이지행, 조성배 "전자우편 문서의 자동 분 류를 위한 다중 분류기 결합" 한국정보과학 회 논문지 소프트웨어 및 응용 제29권 3호

[6] Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz, "A Bayesian Approach to Filtering Junk E-Mail", Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, August 2001

[7] 주덕규, 스팸메일의 현황 및 대책, 한국정보 보호진흥원, 정보통신윤리, 2003.6

[8] 김진만, "스팸메일 차단 시스템 설계 및 구현" 석사학위논문, [2004]

저 자 소 개

박 노 경

1984년 고려대학교 전자공학과 졸업
1990년 고려대학교 전자공학과 공학박사
1999년 3월~2000년2월 OSU ECE 연구교수
2005년 3월~현재 차세대 반도체 연구소장
2006년 6월~현재 RIC 부소장
1988년~현재 호서대학교 공과대학 정보통신공학과 정교수
<주관심분야 : IPTV, ASIC Design, SoC>

한 성 호

1991년 호서대학교 정보통신공학과 공학석사 졸업
2005년 호서대학교 정보통신공학과 공학박사 졸업
2005년~현재 서울 호서 전문학교 전임강사
2006년 3월~현재 호서대학교 정보통신공학과 겸임교수
<주관심분야 : SoC, DMB, ASIC Design>

서 상 진

1999년 부경대학교 전자계산학과 이학사 졸업
2001년 부경대학교 전자계산학과 이학석사 졸업
2001년~현재 서울 호서전문학교 모바일 미디어과 학과장
2005년~현재 호서대학교 정보통신공학과 박사과정
2006년 3월~현재 (주)코모넷 상무
<주관심분야 : 임베디드 시스템, 멀티미디어 정보처리>

진 현 준

1986년 고려대학교 전자공학과 졸업
1998년 미국 리하이대학교 전산학 박사
1998년~현재 호서대학교 정보통신공학과 부교수
<주관심분야 : 시스템 프로그램, 멀티미디어 정보처리>