

# 향상된 차단 성능 지원을 위한 SMBC 플랫폼 개발<sup>☆</sup>

## A Development of the SMBC platform for supporting advanced performance of blocking spam-mails

서 상 진\*      진 현 준\*\*      박 노 경\*\*\*  
Sang-Jin Sso      Hyun-Joon Jin      Noh-kyung Park

### 요 약

현재 새로운 스팸 메일 차단 시스템과 다양한 스팸 차단 기술에 대한 연구가 계속되고 있다. 그렇지만, 새로운 유형의 스팸 메일이 등장하면서 스팸 메일 차단률(Spam mail Filtering Rate)과 오인된 메일(False-positive mail) 발생률은 점차적으로 늘어나고 있다. 하지만 기존에 제안된 스팸 메일 필터링 알고리즘은 새로운 유형의 스팸 메일에 대응하기 위해 적용될 알고리즘 수의 증가와 효율적인 스팸 메일 필터링 알고리즘의 대응 관계에 대한 연구 부족으로 인하여 스팸 메일 차단 시스템의 처리 부하는 증가하고 이에 대한 신뢰성은 반감되고 있다. 본 논문에서는 스팸 메일 차단 시스템의 부하 처리 성능 및 신뢰성을 증가시키기 위해 Fit-FA Finder와 Privacy 기반의 오인된 메일을 복구시키는 SMBC 플랫폼을 개발하고 성능을 분석하였다.

### Abstract

Even though lots of research have been doing about spam mail blocking technologies and their systems, the emergence of spam mails of new types causes the spam mail filtering rate to decrease and the occurrences of false-positive mails to increase. Therefore, existing spam mail filtering algorithms suffer from increasing load to be processed and decreasing reliability in spam mail blocking systems due to the shortage of newly developed algorithms and their research. This paper presents the Fit-FA Finder which is able to select appropriate algorithms to be applied and their procedures, and the development of the SMBC platform. The Fit-FA Finder is developed and implemented in the SMBC platform in which recovering process based on privacy information is employed for false-positive mails

## 1. 서론

현재 전자우편은 적은 비용으로 사용자간 의사 소통의 수단을 제공하고 있으며, 전자우편의 사용은 더욱 더 늘어날 전망이다. 이와 같이 보편화된 전자우편을 통해 전달하고자 하는 의사를 쉽게 송수신할 수 있지만, 원하지 않는 스팸 메일도 수신되는 문제점이 있다[1,3].

그리고 새로운 유형의 스팸 메일의 등장으로 인해 스팸 메일 차단률(Spam mail Filtering Rate)과 오인된 메일(False-positive mail) 발생률은 점차적으로 늘어나고 있다[2]. 하지만 기존에 제안된 스팸 메일 필터링 알고리즘은 다양한 유형의 스팸 메일에 대응하기 위해 지속적으로 적용될 알고리즘의 수가 늘어나고 있다. 그러나, 효율적인 스팸 메일 필터링 알고리즘의 대응 관계에 대한 연구가 부족하여 스팸 메일 차단 시스템의 처리 부하는 증가하고 신뢰성은 반감되고 있다.

본 논문에서는 스팸 메일 차단 시스템의 처리 부하를 줄이고 신뢰성을 증가시키기 위해 Fit-FA Finder와 Privacy 기반의 오인된 메일을 복구시키는 SMBC (Spam Mail Blocking Center) 플랫폼을 개발하고 성능을 분석하였다. 본 논문의 구성은

\* 준 회 원 : 호서대학교 정보통신학과 박사과정  
ssjworld@hanmail.net

\*\* 정 회 원 : 호서대학교 정보통신학과 교수  
hjjin@office.hoseo.ac.kr

\*\*\* 정 회 원 : 호서대학교 정보통신학과 교수  
nkpark@office.hoseo.ac.kr

[2006/11/29 투고 - 2006/12/7 심사 - 2006/12/19 완료]

☆ 이 논문은 2005년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임.

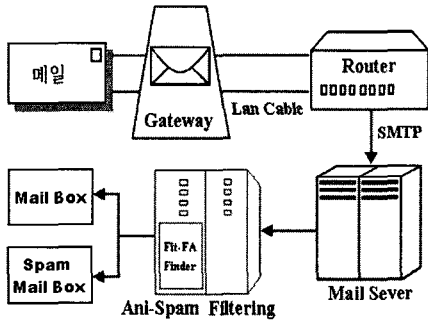
2.1장에서는 임베디드 환경에서의 SMBC를 설계하고, 2.2장에서 구현 및 실험 결과를 분석한다. 그리고 3장에서 결론한다.

## 2. 본 론

### 2.1. SMBC 차단 플랫폼의 설계

#### 2.1.1 SMBC의 개요

스팸 메일 차단 플랫폼인 SMBC는 Proxy Server 기반의 스팸 차단 프레임 레이어로 구성되어 있으며, 복구를 위한 Privacy 정보를 이용한다 [1,4,5]. 그림 1은 SMBC 플랫폼의 전체 구성도를 나타내고 있다.



(그림 1) SMBC 플랫폼의 전체 구성도

SMBC 플랫폼의 스팸 메일 차단 전체 처리 과정은 수신된 메일에 대해 Fit-FA Finder를 적용하여 스팸 필터링 알고리즘 적용 순서 FFL(Filtering Function List)을 결정한다. FFL의 적용 순서에 따라 스팸 필터링 된 메일 중 정상 메일로 판정된 메일은 정상 메일함(Mail Box)으로 이동된다. 그리고 스팸 메일로 판정된 수신 메일은 스팸 메일함(Spam Mail Box)으로 이동되며, Privacy 기반의 오인된 메일 검사를 통해 오인된 메일을 복구 처리한다.

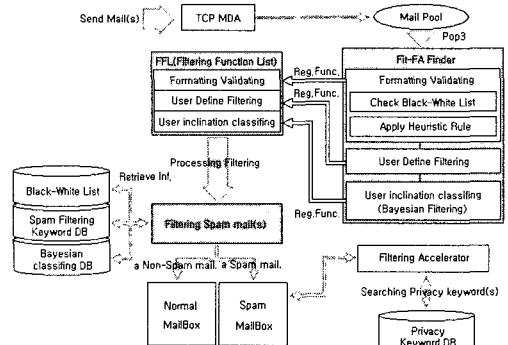
#### 2.1.2 Fit FA-Finder 선택기

스팸 메일 차단 작업은 다양한 스팸 유형에 따

라 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[1].

Fit-FA Finder는 스팸 유형 및 시스템 구조에 따른 차단 절차와 적용 알고리즘을 분석 및 생성해 주는 독립 모듈이다. Fit-FA Finder에 적용할 스팸 메일 필터링 알고리즘은 크게 포맷 검사(Validation) 알고리즘과 사용자의 의도에 따라 차단시킬 필터링 스크립팅(Scripting), 그리고 사용자 취향에 따라 차단 학습을 위한 베이스안 필터링으로 구성된다. 적용 순서는 포맷 검사를 선행하여 사용자의 필터링 스크립팅 적용이 유효하게 처리될 수 있도록 보증한다. 그리고 사용자가 의도한 차단 필터링 스크립팅 적용은 오인된 메일을 발생시킬 수 있으므로, 베이스안 필터링을 적용하여 오인률을 최소화 할 수 있도록 적용 순서를 지정하였다.

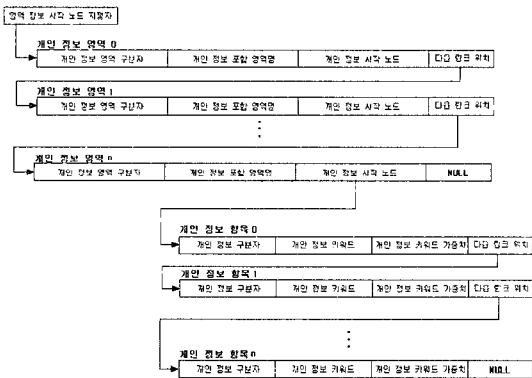
Fit-FA Finder는 SMBC의 Front-Tier에 위치하며, 크게 메일의 포맷 검사(Formatting Validating), 사용자 정의 필터링(User Define Filtering), 사용자 성향 분류(User inclination classifying)를 수행한다. 수행 결과로 생성된 검사 적용 순서 FFL(Filtering Function List)를 이용하여 스팸 필터링을 수행하며, 이와 같은 세부 처리 과정을 그림 2에서 나타내고 있다.



(그림 2) Fit-FA Finder의 세부 처리 세부 구성도

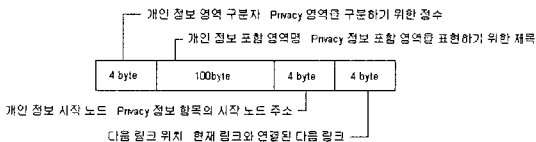
### 2.1.3 FA를 이용한 Privacy 단어 검색

메일 구문 검사시 가장 많은 작업을 수행하는 단어에 대한 문자열 패턴 검사는 문자 단위의 비트 패턴을 검사한다. 본 논문에서는 Unicode 기반의 16비트 문자 검사를 수행하는 FPGA 기반의 FA(Filtering Accelerator)를 이용한다. SMBC 플랫폼은 FA를 이용하여 정상 메일함으로부터 Privacy 정보 항목을 기반으로 수신자 메일 인증을 위해 Privacy 키워드를 저장하는 Privacy Keyword Repository로부터 Privacy 정보를 추출한다. 그림 3에서 Privacy Keyword Repository의 자료 구조를 간략히 나타내고 있다.



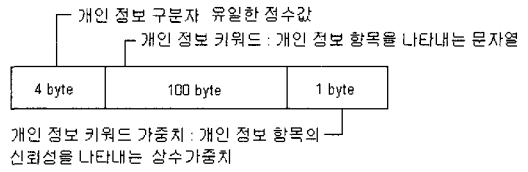
(그림 3) Privacy Keyword Repository의 자료 구조

Privacy Keyword Repository에서 영역 정보를 포함하는 각 노드들의 구성은 그림 4과 같다.



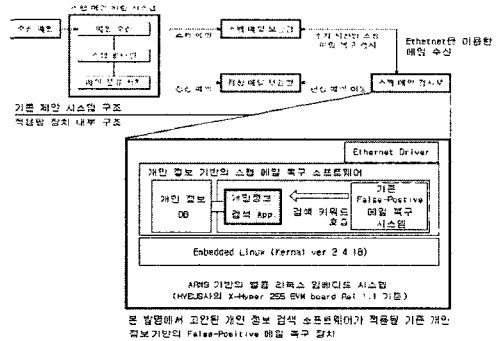
(그림 4) Privacy Keyword 노드의 영역 저장을 위한 노드 구성

Privacy Keyword Repository에서 각 노드들의 구성은 그림 5과 같다.



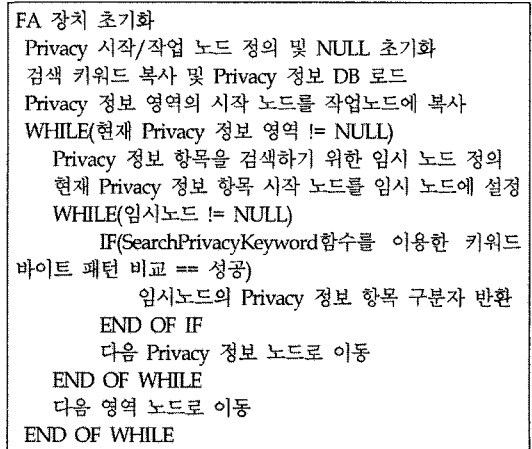
(그림 5) Privacy Keyword 저장을 위한 자료 구조

그림 6은 False-Positive 메일 추출 작업시 FA 내에 Privacy 정보 항목 검색 작업을 위해 연동되는 처리 과정을 나타내고 있다.



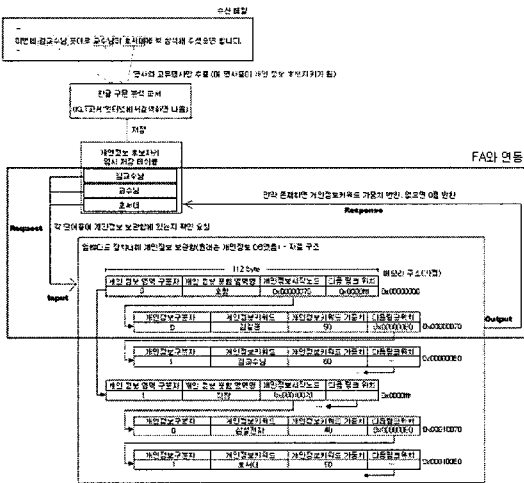
(그림 6) Privacy 노드 검색을 위한 FA와의 연동

FA는 Privacy 키워드를 가중치로 가지는 연결 리스트로 구성되어 있으며, 이더넷을 이용하여 바이트 단위로 비교할 Privacy 키워드를 전송 및 결과를 반환 받는다. 그림 7는 FA 장치를 이용한 검색 처리를 알고리즘으로 나타내고 있다.



(그림 7) Privacy 정보 항목 검색 알고리즘

Privacy 정보 항목 검색 알고리즘 처리를 통해 문자 패턴을 비교하며, 패턴 비교시 FPGA로 설계된 문자 비교기를 이용한다. 비트 패턴 비교는 ex-nor 연산 로직을 이용하여 수행되며, 비교할 문자 패턴의 비교 횟수는 최대 Privacy 정보 항목의 영역과 각 Privacy 정보 항목 영역에 포함된 Privacy 정보 항목의 수에 비례되며, 알고리즘의 처리 예시가 그림 8에서 나타내고 있다.

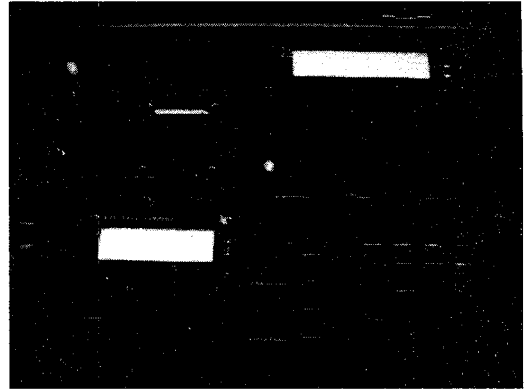


(그림 8) FA 내에 개인 정보 검색 과정 처리 예시

## 2.2. 구현 및 실험 결과

스팸 메일 차단을 위한 기능을 구현하고 성능을 검사하기 위해 기본적인 기능을 이행하는 SMBC 플랫폼을 설계 및 구현하였다. 처리를 검증하기 위한 Simulation Tool은 MAX+plusI와 (주)한백전자 HBE-Combo를 이용하였다.

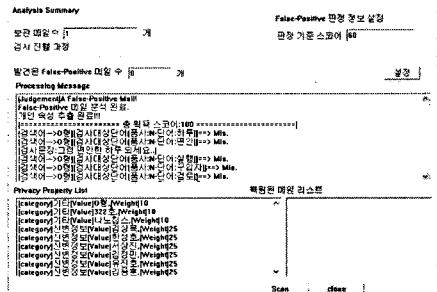
SMBC 플랫폼의 구현 기능들 간의 연동 검증 및 개별 스팸 메일 차단 기능 검사를 위해, Privacy 키워드의 가중치가 70이하인 샘플 메일을 작성하여 그림 9와 같은 예상된 처리 결과를 얻을 수 있었다.



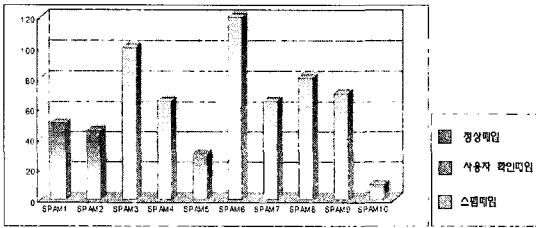
(그림 9) 가중치 합계 70 이하인 샘플 메일 처리 결과

실험을 통해 FA 처리가 정상적으로 이루어지고 있음을 알 수 있었다. 이와 같은 SMBC 플랫폼 구성 모듈의 구성 모듈의 기능이 검증됨에 따라, Privacy 정보 기반의 오인된 메일 복구 처리를 수행하는 Prototype을 구현 및 실험하였다.

SMBC Prototype의 S/W 구성 모듈은 윈도우 2003 Server 환경에서 Visual Studio 6.0, KLT 5.0 파서를 이용하여 구성하였다. 스팸메일 100개당 1개의 오인된 메일이 발생 하는데, 10개의 오인된 메일을 얻기 위해서 1000개의 Spam 메일을 이용하여 유도한 의도에 따라 10개의 오인된 메일을 실험에 사용 하였다.판정 기준은 0-39, 40-59, 60 이상인 가중치에 대해 각각, spam, unsure, normal로 판정한다. 그림 10에서 SMBC의 처리 과정을 나타내고 있으며, 그림 11에서 1000개의 샘플 메일 중 10개의 False-Positive 메일 검사 결과를 나타내고 있다.



(그림 10) SMBC 플랫폼 처리 화면



(그림 11) SMBC 플랫폼의 스팸 메일 차단 실험 결과

본 논문에서 구현된 SMBC Prototype은 오인된 메일로 유도된 10개중 6개를 정상 메일, 2개를 스팸 메일, 그리고 사용자 확인(unsure) 메일로 2개의 수신 메일을 판정하였다. 오인된 4개의 오인된 메일은 실험에서 사용된 한국어 파서인 KLT 구분 분석기의 형태소 중심의 분석 성능 및 관심 영역의 키워드 추출을 위한 단순 학습 방법에 기인되었다.

### 3. 결론

본 논문에서는 SMBC의 성능 향상을 위해 Fit-FA Finder를 설계하고, FA 연동형 SMBC 플랫폼을 개발하여 성능을 분석하였다.

기존 논문[1,2]에서 제안된 차단 시스템의 S/W 기반의 차단 필터를 H/W 기반의 연동 방식으로 전환하여 차단 처리 부하를 분산시켰다. 그리고 Fit-FA Finder의 적용을 통해 스팸 필터링 처리 부하 감소, 레이어 기반의 시스템 구성을 통해 모듈별 독립하였다. 이에 따라, 오인된 메일 복구 성능 실험을 통해 판정 오인률(0.001%-10%)은 6%

축소되어 스팸 메일 차단률(Filtering rate + Recovery rate)과 성능이 향상되었다.

향후 연구 과제로 FA, Fit-FA Finder, Privacy 기반의 오인된 메일 복구 모듈 추가로 인해 기능 확장성, 이식성, 처리 복잡도, 지식 정보의 크기 등의 처리 비용을 더욱 줄이기 위해 Privacy 기반의 오인된 메일 검색 SoC를 설계한다.

### 참고 문헌

- [1] 박노경, 서상진, 진현준, 한성호, 스팸 메일 차단 신뢰도 향상을 위한 SMBC 플랫폼 설계, 한국국통신학회 논문지, 30권 11B, 2005년 7월, pp. 730-736
- [2] 김현준, 정재은, 조근식, 가중치가 부여된 베이저안 분류자를 이용한 스팸 메일 필터링 시스템, 한국정보과학회논문지, 2004년 xx월, pp.1092-1100
- [3] 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
- [4] T. Saito, K. Umesawa, and H.G. Okuno, A Privacy-Enhanced Access Control, 日本電子情報通信學會論文誌, 2001년 11월
- [5] 서상진, 진현준, 박노경, Privacy 속성 기반의 오인된 메일 복구 알고리즘 개발, 전기전자학회 논문지 Vol. 9. No. 2, pp. 108-114

## ● 저 자 소개 ●



### 서 상 진(Sang-Jin Sso)

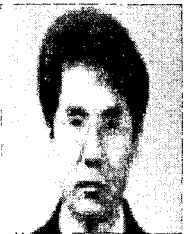
1999년 부경대학교 전산학과 졸업(학사)

2001년 부경대학교 일반대학원 전산학과 졸업(석사)

2005~현재 호서대학교 정보통신학과 박사과정

관심분야 : 임베디드 시스템, 멀티미디어 응용, 모바일 응용, etc.

E-mail : ssjworld@hanmail.net



### 진 현 준(Hyun-Joon Jin)

1984년 고려대학교 전자공학과 졸업(학사)

1986년 고려대학교 전자공학과 졸업(석사)

1998년 Lehigh 대학교 전산학과 졸업(박사)

1998~현재 호서대학교 정보통신학과 교수

관심분야 : 시스템프로그래밍, 멀티미디어 프로토콜, etc.

E-mail : hjjin@office.hoseo.ac.kr



### 박 노 경(Noh-kyung Park)

1984년 고려대학교 전자공학과 졸업(학사)

1986년 고려대학교 전자공학과 졸업(석사)

1990년 고려대학교 전자공학과 졸업(박사)

1988~현재 호서대학교 정보통신학과 교수

관심분야 : 회로 및 시스템설계, SoC 설계, etc.

E-mail : nkpark@office.hoseo.ac.kr