userID 기반의 빠른 메일 차단 알고리즘

심재창*, 고주영**, 김현기**
안동대학교 컴퓨터 공학과*, 멀티미디어 공학과**

A fasrter Spam Mail Prevention Algorithm on userID based

Jae-Chang Shim*, Joo-Young Ko**, Hyun-Ki Kim**
*Computer Engineering, **Multimedia Engineering, Andong National Univ.
{jcshim, sonice, hkkim}@andong.ac.kr

요 약

스팸메일로 인한 피해가 크게 늘어나고 있어 스팸 필터링과 차단에 관한 연구가 활발하다. 스팸메일 차단에 이메일 주소 대신 userID(사용자아이디)를 비교하여 처리 속도를 빠르게 하는 방법을 제안한다. userID가 중복되어 스팸메일이 통과하는 경우가 2% 정도 발생하는데 해당 도메인을 불량 도메인 목록에 등록해서 차단한다.

제안된 방법은 이메일 주소를 비교하는 방법 보다 DB용량도 줄어 들고, 문자의 비교에서 약 3.7배 속도가 향상된다. userID의 자동등록을 위해 등록되지 않는 메일이 수신되면 비밀단어를 반송하는 방법을 적용하였다.

ABSTRACT

The problem of unsolicited e-mail has been increasing for years, so many researchers has studied about spam filtering and prevention. In this article, we proposed a faster spam prevention algorithm based on userID instead of full email address. But there are 2% of false-negatives by userID. In this case, we store those domains in a DB and filter them out.

The proposed algorithm requires small DB and 3.7 times faster than the e-mail address comparison algorithm. We implemented this algorithm using SPRSW(Spam Prevention using Replay Secrete Words) to register userID automatically in userID DB.

Kewwords

spam, e-mail, userID, domain, white e-mail, black e-mail, spam filtering, spam prevention

I. 서 론

인터넷의 중요한 구성 요소의 하나인 이메일은 네트워크의 발달과 더불어 사용이 증가되었고, 기능이 편리해서 전화처럼 우리생활의 한 부분으로 자리 잡았다. 그러나 광고나 음란스팸메일로 이메일의 관리에 어려움이 많고 서버의 용량도 증가된다[1].

스팸메일이란 인터넷을 통하여 대량으로 전달되는 원하지 않는 상업성 이메일(UCE : Unsolicited Commercial E-mail)을 통칭한다. 스팸메일을 차단하는 여러가지 기술적인 방법에는 메일의 헤더나 본문내용의 키워드 기반 필터, blacklist 차단방법, 복합적인 내용 검색방법등이 있다.

키워드 기반 필터는 계속적으로 내용을 업데이트 해야하며 스팸메일 정보를 기준으로 차단하는 방법은 스팸메일 정보가 계속 증가하여 DB의 양이 계속 증가하는 단점이 있다.

최근에 Paul Graham의 베이지안 필터[2]는 많은 신뢰도를 얻고 있으며 이후에 여러 연구에서 베이지안 스팸필터에 대한 연구[3-5]가 진행 중이다. 베이지안 스팸 필터는 스팸메일을 차단하는 효과가 비교적 크지만 정상메일을 스팸메일로 분류하는 오류(false-positive)가 역시 발생한다. 이러한 이유로 스팸메일로 분류된 메일들을 항상 재 확인해야 하는 번거로움이 있다.

II. 이메일 주소 기반의 메일 차단 알고리즘

이메일 주소기반으로 스팸메일을 차단하는 기본적인 방법으로 양호 이메일 주소(white e-mail address)만 수신하는 방법과 불량 메일 주소(black e-mail address)를 차단하는 방법으로 시스템을 구성할 수 있다. 스팸메일은 주소를 자주 바꾸므로 불량메일 차단 보다는 양호 이메일 통과가 DB활용면과 처리 속도측면에서 효과적이다.

양호 이메일 주소 통과 시스템에서 userID와 도메인이 합쳐진 이메일 주소를 전체를 비교한다. 이 방법은 도트(.)와 @이 포함된 평균 21개의 문자열로 구성된 이메일 주소를 모두 비교해야 하므로 시간이 오래 걸리므로 DB에 등록된 userID와 수신된 메일의 userID를 추출하여 비교하는 방법으로 비교되는 문자의 수를 줄여 빠르게 처리하는 방법을 제안한다.

III. userID 비교 기반의 메일 차단 알고리즘

이메일 주소는 userID@domain-name.xxx 등으로 구성이 된다. @앞 부분은 사용자아이디이며 @ 다음은 도메인이다. 수신하기 원하는 이메일 주소를 white e-mail address DB에 저장하는 대신 사용자 아이디 만으로 구성된 white userID DB를 만들고 수신된 메일에서 userID를 추출하여 DB의 userID와 비교를 하는 방법이다.

white userID를 통과시키면 userID는 중복될 수 있어 스팸메일이 차단되지 않을 수 있다. 양호 이메일 주소 userid@domain-name.com으로 인해 userid가 등록되면 userID는 같고 도메인이 다른 userid@bestsex.com이 통과 될 수 있다. 메일을 확인하는 과정에서 이런 메일이 있는 경우 도메인을 black domain DB에 등록한다. 이 과정은 자동으로 처리 될 수 없어 수동으로 웹에서 사용자 인터페이스를 통해 직접 저장하도록 하였다. userID 중복으로 인한 스팸인 경우는 약 2%가 발생한다. 그림 1은 제안된 userID 기반의 빠른 메일 차단 알고리즘이다.

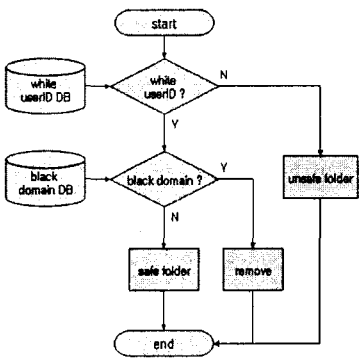

그림 1. 제안된 white userID와 black domail으로 스팸메일을 차단하는 과정

IV. 시스템의 구현 및 고찰

제안된 userID 기반의 스팸 차단 알고리즘을 procmail을 이용하여 구현하였다. 인터넷의 웹 브라우저에서 설정파일들을 변경할 수 있도록 php 언어로 사용자 인터페이스를 구현하였다. 수동으로 userID를 등록하려면 어렵기 때문에 자동으로 userID를 등록하는 방법으로 반송 비밀번호를 활용하는 방법(SPRSW)[6]을 적용하였다. 그림 2는 구현된 스팸메일 차단 시스템의 전체 구성도 이다.

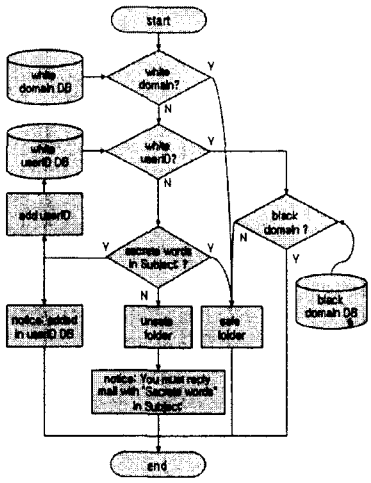

그림 2. 제안된 시스템의 흐름도

스팸 차단 시스템은 Linux 시스템에서 sendmail과 procmail이 실행 되는 환경에서 .forward 파일과 .procmailrc 파일이 이용되었다. 또한 IMAP와 연동하여 서버에 있는 폴더를 로컬 폴더에서 자유롭게 연결하여 사용할 수 있도록 하였다. 그림 3은 userID 기반의 빠른 스팸메일 차단 시스템의 전체 구성도이다.

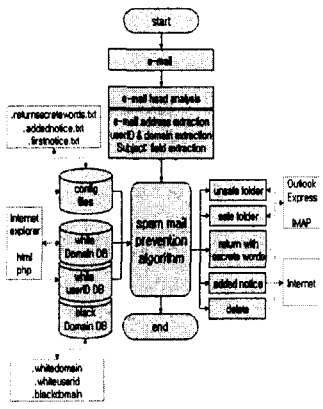

그림 3. userID 기반의 스팸 메일 차단 시스템

php 언어로 구현된 설정파일 변경을 위한 사용자 인터페이스 창은 그림 4와 같다. 회신 비밀단어와 메일이 처음 도착하였음을 알리는 알림 글과 비밀단어를 제목에 적어 보낸 경우 userID가 등록되었음을 알리는 메시지를 입력할 수 있다. 발신자가 이메일을 보냈을 경우 비밀단어는 사용자가 한 개 또는 여러 개 지정할 수 있다. 여러 개로 지정을 할 때 단어 사이를 '|'로 구분 하도록 하였다.

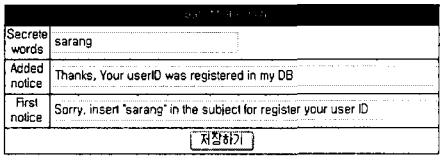

그림 4. 설정파일 변경 사용자 인터페이스

발신자가 메시지를 받고 반송비밀단어를 입력한 다음 다시 메일을 보내면 스팸메일 차단시스템이 비밀단어가 맞는 경우 수신 이메일 주소 목록에 자동 저장된다. 그림 5는 사용자가 등록 되었음을 알리는 메시지이다.

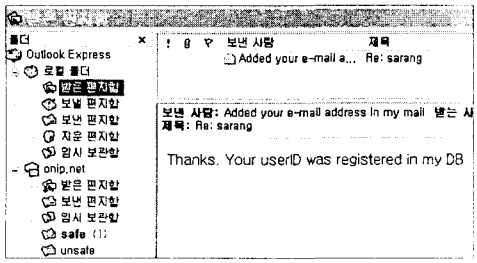

그림 5. 사용자가 등록 되었음을 알림

수신 이메일 주소 목록에 등록된 발신자가 메일을 보냈을 경우 메일이 바로 safe folder로 전달되고 발신자에게 메일이 전달되었음을 알린다.

제안된 방법은 자신이 원하는 이메일만 받을 수 있고 주소가 등록되지 않은 이메일이 도착 했을 때에도 시스템이 자동으로 비밀단어를 포함한 반송메일을 보내어 확인함으로 주소를 등록할 수 있으며 스팸메일로 잘못 분류하는 오류 율을 최소화 할 수 있다. 그림 6은 white domain 관리 창이다.

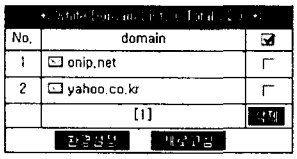

그림 6. white domain DB 확인 창

그림 7은 white userID 관리 창이다. 수동 등록도 가능하며 여러 개를 선택해 지울 수 있다.

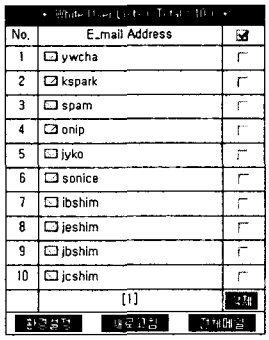

그림 7. white userID DB 확인 창

제안된 방법을 2개의 도메인에 대해서 예제로 적용해 보았다. 표 1에서와 같이 652명이 등록된 palgong.kyungpook.ac.kr의 userID의 평균 문자 수는 5.98 자이며, 876명이 등록된 andong.ac.kr에 대한 userID의 평균 문자 수는 5.46자이다. 평균 userID의 문자 수는 5.72 자이며, 도메인의 문자 수는 점을 포함하여 14.5개이다. 이메일은 @을 포함하므로 21.22개의 문자열로 구성된다. 그러나 userID와 도메인의 길이는 실험대상에 따라 달라 질 수 있다.

표 1. 이메일 주소에서 userID와 도메인의 문자 수

userID & Domain	palgong.knu.ac.kr	andong.ac.kr	Avg.
Avg. characters in userID	5.98	5.46	5.72
Avg. characters in Domain	17.00	12.00	14.50
Total characters	22.98	17.46	20.22
Include @			21.22

100개의 white e-mail address를 가지고 있는 경우의 예이다. userID가 중복된 불량 메일이 2% 이므로 100개의 메일 중에 2개가 중복된다. 100개의 메일이 도착한 경우 따라서 black domain DB에는 2개의 도메인이 등록 된다. 표 2는 이메일 비교 방법과 userID 비교 방법에서 비교되는 문자의 수를 나타낸다.

표 2. 이메일 비교 방법과 userID 비교 방법에서 비교되는 문자의 수

Methods	Comparison characters	Totals	%
white e-mail address	100x21.22x100	2,122,000	3.69
white userID & black domain	100x5.72x100	574,900	1.00
	2x14.50x100		

제안된 방법은 white e-mail address 차단이나 black e-mail address filtering에도 적용이 가능하

다. 이와 같이 제안된 방법은 수신 목록을 등록하여 비교함으로 DB의 용량을 줄일 수 있고 등록되지 않은 사용자를 비밀단어 회신을 통하여 자동으로 등록시키고 스팸메일 차단 처리 속도를 증가 시킬 수 있다.

V. 결 론

본 논문은 스팸메일 필터링 또는 차단에서 이메일 주소 대신 userID를 비교하는 방법을 제안하였다. DB의 용량이 줄어 들고 이메일 주소를 비교하는 경우 보다 문자 비교에서 약 3.7배 비교 량을 줄일 수 있었다. 제안된 알고리즘은 개인의 스팸메일 차단뿐만 아니라 대량의 메일이 통과되는 메일 서버 시스템에 적용하기에 적합하다.

제안된 알고리즘을 스팸메일 차단 시스템으로 구현하였다. userID가 등록되지 않은 메일을 수신하면 시스템에서 회신 비밀단어를 보내고, 제목에 비밀 단어를 넣어 보내어 자동으로 등록하는 반송 비밀번호 시스템(SPRSW)을 적용하였다.

또한 사용자가 웹 브라우저에서 설정 파일 사용자 인터페이스를 관리할 수 있어 DB 관리에 편리하며 IMAP와 연동해서 Outlook Express에서 사용할 수 있어 사용자 환경에 구애받지 않고 사용할 수 있다.

참고 문헌

[1] 2001년 정보화 역기능 실태조사서, 한국 정보 보호원, 12월 2001

[2] Paul Graham, "Better Bayesian Filtering," Proceedings of the 2003 Spam Conference, 2003.

[3] Provost, J. "Naive-Bayes vs. Rule-Learning in Classification of Email," The University of Texas at Austin, Artificial Intelligence Lab. Technical Report AI-TR-99-284, 1999.

[4] W. Yu, "New anti-spam filter based on data mining and analysis of email security," Proceedings of SPIE Data Mining and Knowledge Discovery: Theory, Tools, and Technology, Vol.V, pp.147-154, 2003.

[5] 민도식외, "SVM 분류 알고리즘을 이용한 스팸메일 필터링", 한국정보과학회 03 봄 학술발표논문집(B) pp.552-554, 2003.

[6] 고주영 외, "비밀단어의 회신을 이용한 스팸메일 차단시스템의 구현", 16회 신호처리 합동학술대회 논문집, 2003.