Corso di Laurea in Informatica
Progetto per l'esame di Programmazione 2
Appelli estivi a.a. 2006-07

Obiettivo

Si vuole costruire un sistema antispam che permetta di bloccare la posta indesiderata.

Premessa

Spam *(da Wikipedia)*

Lo spamming (detto anche fare spam) è l'invio di grandi quantità di messaggi indesiderati (generalmente commerciali). Può essere messo in atto attraverso qualunque media, ma il più usato è Internet, attraverso l'e-mail.

Il termine trae origine da uno sketch comico del Monty Python's Flying Circus ambientato in un locale nel quale ogni pietanza proposta dalla cameriera era a base di Spam (un tipo di carne in scatola). Man mano che lo sketch avanza, l'insistenza della cameriera nel proporre piatti con "spam" ("uova e spam, uova pancetta e spam, salsicce e spam" e così via) si contrappone alla riluttanza del cliente per questo alimento, il tutto in un crescendo di un coro inneggiante allo "spam" da parte di alcuni Vichinghi seduti nel locale.

I Monty Python prendono in giro la carne in scatola Spam per l'assidua pubblicità che la marca era solita condurre, in un certo periodo la Spam era ovunque, da quì lo sketch dei Monty's e successivamente l'adattamento informatico alla pubblicità non desiderata.

Il principale scopo dello spamming è la pubblicità, il cui oggetto può andare dalle più comuni offerte commerciali a proposte di vendita di materiale pornografico o illegale, come software pirata e farmaci senza prescrizione medica, da discutibili progetti finanziari a veri e propri tentativi di truffa. Uno spammer, cioè l'individuo autore dei messaggi spam, invia messaggi identici (o con qualche personalizzazione) a migliaia di indirizzi e-mail. Questi indirizzi sono spesso raccolti in maniera automatica dalla rete (articoli di Usenet, pagine web) mediante spambot ed appositi programmi, ottenuti da database o semplicemente indovinati usando liste di nomi comuni.

Per definizione lo spam viene inviato senza il permesso del destinatario ed è un comportamento ampiamente considerato inaccettabile dagli Internet Service Provider (ISP) e dalla maggior parte degli utenti di Internet. Mentre questi ultimi trovano lo spam fastidioso e con contenuti spesso offensivi, gli ISP vi si oppongono anche per i costi del traffico generato dall'invio indiscriminato.

Sondaggi hanno indicato che al giorno d'oggi lo spam è considerato uno dei maggiori fastidi di Internet; l'invio di questi messaggi costituisce una violazione del contratto "Acceptable Use Policy" (condotta d'uso accettabile) di molti ISP e pertanto può portare all'interruzione dell'abbonamento (account) del mittente. Un gran numero di spammer utilizza intenzionalmente la frode per inviare i messaggi, come l'uso di informazioni personali false (come nomi, indirizzi, numeri di telefono) per stabilire account disponibili presso vari ISP. Per fare questo vengono usate informazioni anagrafiche

false o rubate, in modo da ridurre ulteriormente i loro costi. Questo permette di muoversi velocemente da un account a un altro appena questo viene scoperto e disattivato dall'ISP. Gli spammer usano software creato per osservare connessioni Internet con scarsa sicurezza, che possono essere facilmente dirottate in modo da immettere i messaggi di spam direttamente nella connessione dell'obiettivo con il proprio ISP. Questo rende più difficile identificare la posizione dello spammer e l'ISP della vittima è spesso soggetto di aspre reazioni e rappresaglie da parte di attivisti che tentano di fermare lo spammer. Entrambe queste forme di spamming "nascosto" sono illegali, tuttavia sono raramente perseguiti per l'impiego di queste tattiche.

I mittenti di e-mail pubblicitarie affermano che ciò che fanno non è spamming. Quale tipo di attività costituisca spamming è materia di dibattiti, e le definizioni divergono in base allo scopo per il quale è definito, oltre che dalle diverse legislazioni. Lo spamming è considerato un reato in vari paesi e in Italia l'invio di messaggi non sollecitati è soggetto a sanzioni.

I più grandi ISP come America OnLine riferiscono che una quantità che varia da un terzo a due terzi della capacità dei loro server di posta elettronica viene consumata dallo spam. Siccome questo costo è subito senza il consenso del proprietario del sito, e senza quello dell'utente, molti considerano lo spam come una forma di furto di servizi. Molti spammer mandano i loro messaggi UBE attraverso gli open mail relay. I server SMTP, usati per inviare e-mail attraverso internet, inoltrano la posta da un server a un altro; i server utilizzati dagli ISP richiedono una qualche forma di autenticazione che garantisca che l'utente sia un cliente dell'ISP. I server open relay non controllano correttamente chi sta usando il server e inviano tutta la posta al server di destinazione, rendendo più difficile rintracciare lo spammer.

Un punto di vista "ufficiale" sullo spamming può essere trovato nel RFC 2635.

Lo spamming è a volte definito come l'equivalente elettronico della posta-spazzatura (junk mail). Comunque, la stampa e i costi postali di questa corrispondenza sono pagati dal mittente - nel caso dello spam, il server del destinatario paga i costi maggiori, in termini di banda, tempo di elaborazione e spazio per immagazzinamento. Gli spammer usano spesso abbonamenti gratis, in modo tale che i loro costi siano veramente minimi. Per questa ricaduta di costi sul destinatario, molti considerano questo un furto o un equivalente di crimine. Siccome questa pratica è proibita dagli ISP, gli spammer spesso cercano e usano sistemi vulnerabili come gli open mail relay e server proxy aperti. Essi abusano anche di risorse messe a disposizione per la libera espressione su internet, come i remailer anonimi. Come risultato, molte di queste risorse sono state disattivate, negando la loro utilità agli utenti legittimi. Molti utenti sono infastiditi dallo spam perché allunga i tempi che usano per leggere i loro messaggi di e-mail.

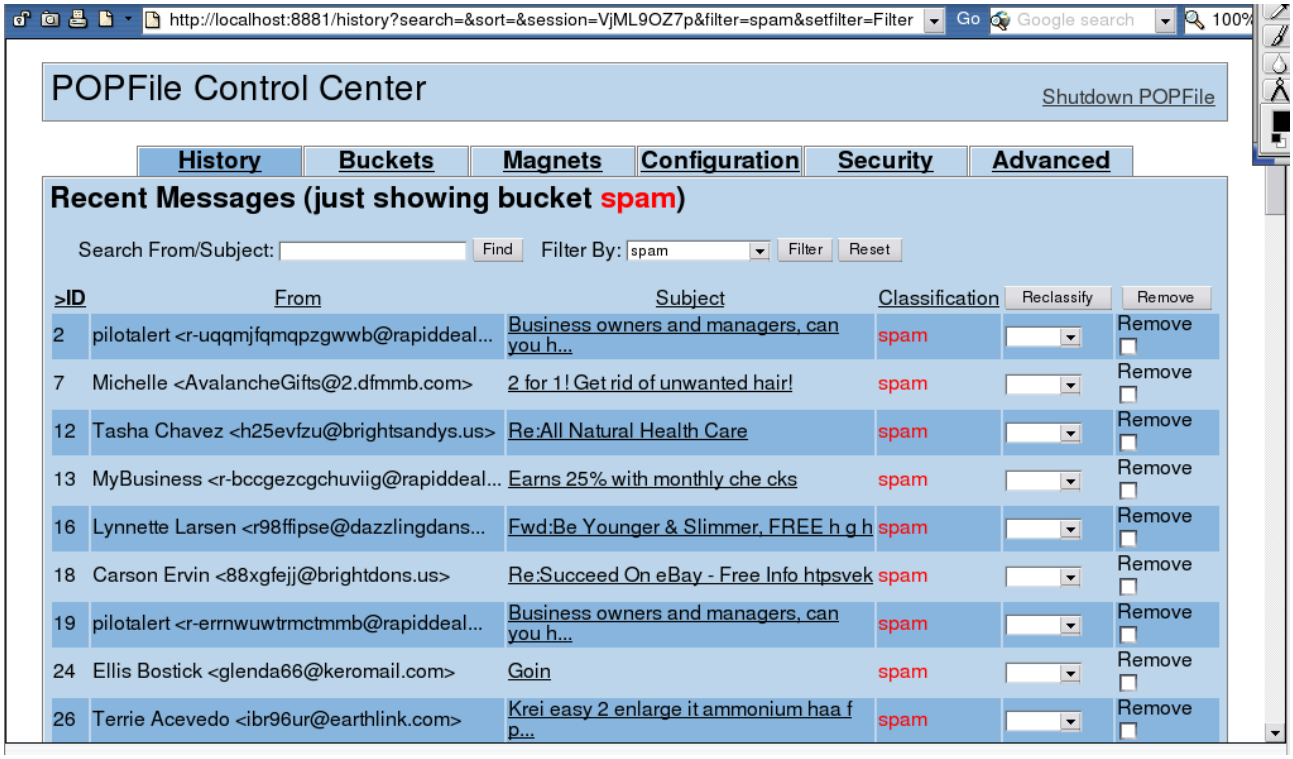
Siccome lo spam è economico da inviare, un ristretto numero di spammer possono saturare internet con la loro spazzatura. Nonostante solo un piccolo numero dei loro destinatari sia intenzionato a comprare i loro prodotti, ciò consente loro di mantenere questa pratica attiva. Inoltre, sebbene lo spam appaia per una azienda rispettabile una via economicamente non attuabile per fare business, è sufficiente per gli spammer professionisti convincere una piccola porzione di inserzionisti ingenui che è efficace per fare affari.

Una tecnica di difesa: filtraggio statistico ed euristico

Fino a poco tempo fa, le tecniche di filtraggio facevano affidamento agli amministratori di sistema che specificavano le liste di parole o espressioni regolari non permesse nei messaggi di posta. Perciò se un server riceveva spam che pubblicizzava "herbal Viagra", l'amministratore poteva inserire queste parole nella configurazione del filtro. Il server avrebbe scartato tutti i messaggi con quella frase. Lo svantaggio di questo filtraggio "statico" consiste nella difficoltà di aggiornamento e

nella tendenza ai falsi positivi: è sempre possibile che un messaggio non-spam contenga quella frase. Il filtraggio euristico, come viene implementato nel programma SpamAssassin, si basa nell'assegnare un punteggio numerico a frasi o modelli che si presentano nel messaggio. Quest'ultimo può essere positivo, indicando che probabilmente contiene spam o negativo in caso contrario. Ogni messaggio è analizzato e viene annotato il relativo punteggio, esso viene in seguito rifiutato o segnalato come spam se quest'ultimo è superiore ad un valore fissato. In ogni caso, il compito di mantenere e generare le liste di punteggi è lasciato all'amministratore. Il filtraggio statistico, proposto per la prima volta nel 1998 nel AAAI-98 Workshop on Learning for Text Categorization, e reso popolare da un articolo di Paul Graham nel 2002 usa metodi probabilistici, ottenuti grazie al Teorema di Bayes, per predire se un messaggio è spam o no, basandosi su raccolte di email ricevute dagli utenti.

PopFile (http://popfile.sourceforge.net/wiki/) è un eccellente sistema per la classificazione delle e-mail in arrivo che permette di separare, ad esempio, le e-mail provenienti dai propri contatti fidati da quelle assolutamente indesiderate (spam). Non solo. Il software è estremamente flessibile in quanto, se l'utente lo desidera, si incarica di effettuare anche una suddivisione in base al contenuto delle varie e-mail. Il programma, per esempio, può essere impiegato non solo per scartare le e-mail di spam ma anche per organizzare, in cartelle differenti, le e-mail di lavoro da quelle personali.



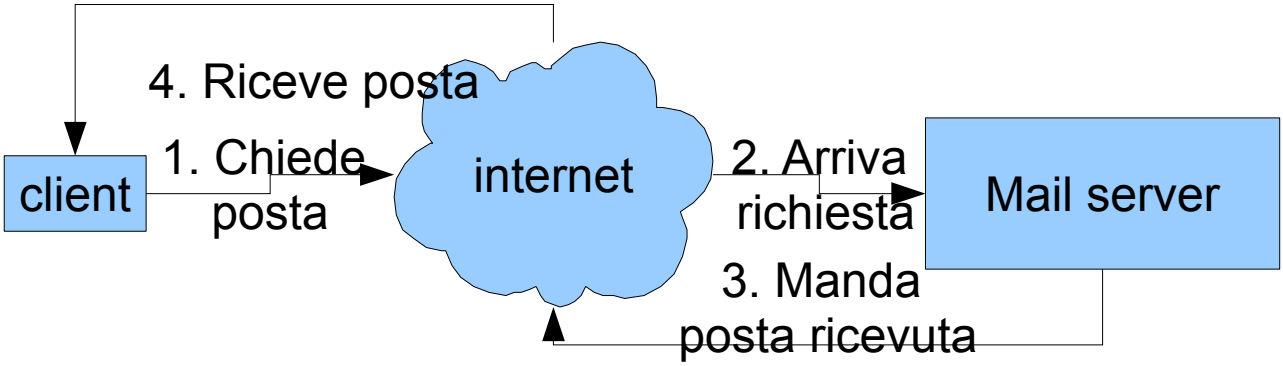
PopFile si frappone tra il client di posta elettronica installato sul personal computer ed il server del provider Internet che gestisce la casella e-mail. La funzione di PopFile è quindi assimilabile a quella di un server proxy: il client di posta (Outlook, Outlook Express, Mozilla Thunderbird, Eudora,…) richiede la ricezione delle e-mail in arrivo a PopFile; quest'ultimo si collega al server POP3 del provider e provvede ad effettuare il download della posta elettronica infine analizza il contenuto di ogni messaggio marcandolo nel modo più opportuno (spam; comunicazione di lavoro; e-mail personale e così via).

Va tuttavia sottolineato come, alla prima installazione, PopFile non sappia come debbano essere trattati i vari messaggi di posta in arrivo: il software non conosce il significato delle varie cartelle previste (“spam”, “lavoro”, “personale”,…). Sarà l’utente a dover istruire il programma sul comportamento da tenere nei vari casi. PopFile è però in grado di imparare assai rapidamente: le sue abilità nel riconoscere la posta indesiderata cresceranno proporzionalmente alla durata del periodo di “addestramento” iniziale.

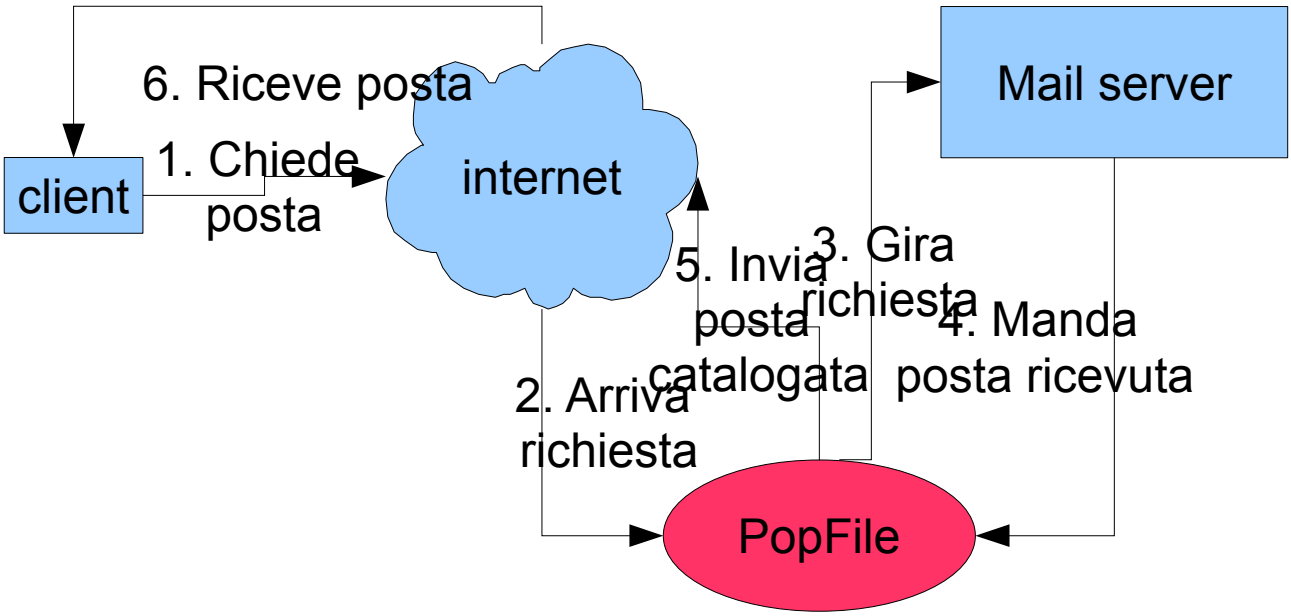
Descrizione del progetto

Sfruttando le capacità di catalogazione della posta di PopFile, vogliamo costruire un filtro antispam.

La normale interazione tra client e server di posta avviene tramite il protocollo Pop3 ed è schematizzata in figura:

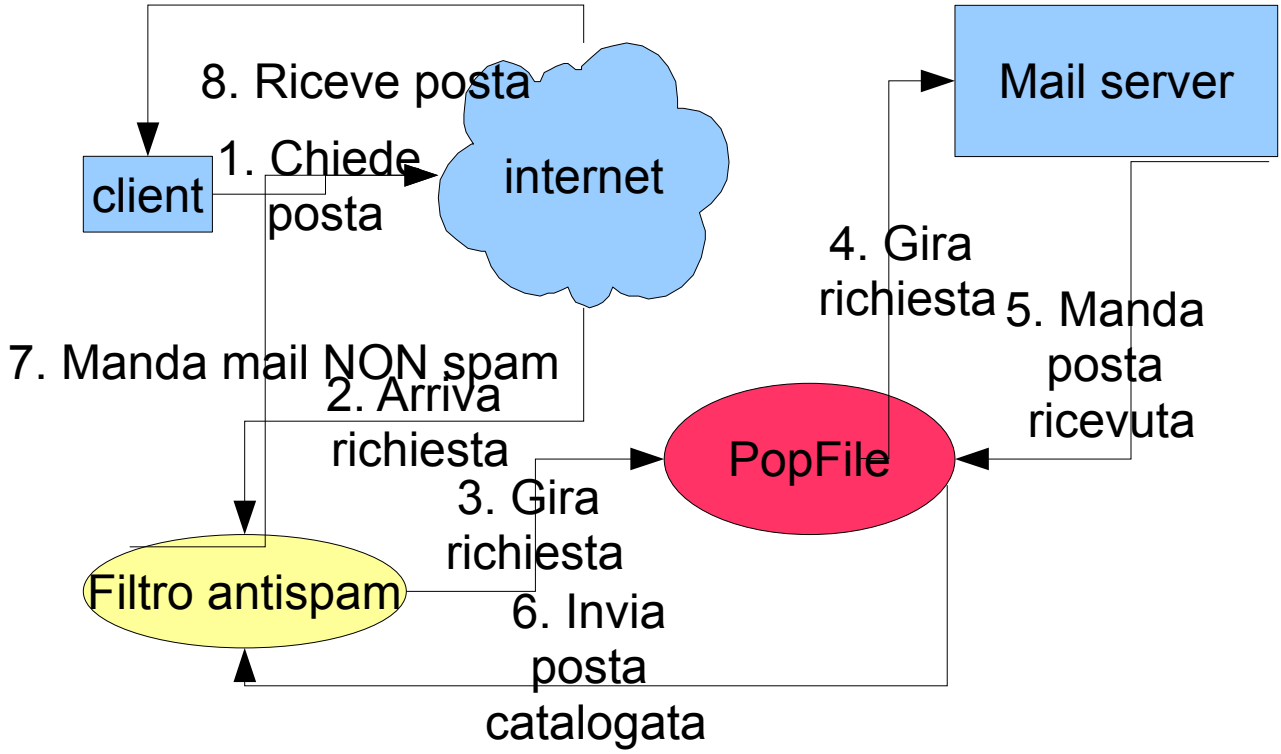


L'introduzione del filtro di PopFile avviene tra il MailServer ed il client:



Si può impostare PopFile così da anteporre la categoria del messaggio nell'oggetto della mail.

Il filtro antispam viene inserito prima di PopFile:



Il filtro si comporrà di 2 programmi:

1. un programma batch **PB**
2. una web application **WA**

PB

PB si comporta come un server Pop3 nei confronti del client e riceve le sue richieste. Le gira a PopFile (sempre con il protocollo Pop3) ed attende di ricevere le email da questo catalogate. Se riceve mail il cui oggetto inizia con [spam]:

1. la salva in un database
2. invia una mail al mittente segnalando che la mail è stata considerata spam e pertanto non inoltrata al destinatario. L'avviso contiene un link alla WA (contenente la chiave del record) per modificare la catalogazione ed inoltrare il messaggio originale al destinatario

Le mail che non rientrano in questa categoria vengono inoltrate al destinatario.

Le mail memorizzate nel database devono essere cancellate periodicamente.

WA

WA è una normale pagina ASP. Estrae la chiave del record dall'url e provvede ad inoltrare la mail che era stata catalogata come spam, cancellandola dal database.

Funzionalità opzionali

1. WA potrebbe riclassificare la mail in PopFile
2. rendere parametriche la porta di collegamento al filtro antispam, al popfile, l'intervallo di tempo per la cancellazione dal database, ecc.

Linguaggio e piattaforma

L'applicazione deve essere sviluppata in ambiente .Net in C# e ASP.