

SPAM

Sistemi di Elaborazione:
Sicurezza Su Reti A.A. 2005/2006
Prof. Alfredo De Santis

A cura di:
De Pascale D.
Zappullo D.

INDICE

- **Introduzione** ←
- La definizione di SPAM
- L'origine del termine SPAM
- Da chi viene effettuato
- Perché lo SPAM è un problema?
- Come prevenire lo SPAM
- Metodi per combattere lo SPAM
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- Come opera uno spammer
- Lo SPAM è illegale?

Introduzione

A chi non è mai capitato di ricevere nella propria mail box messaggi indesiderati? Se la risposta è sì allora stiamo parlando di Spam.

Illustriamo il problema nei seguenti aspetti:

- L'origine del termine spam e cosa esso rappresenta
- Le problematiche relative allo spam e come prevenirlo
- Le tecniche più comuni per combattere lo spam
- Le tecniche più usate dagli spammer
- La normativa in materia di spam

Definizione del termine Spam

Esistono diverse correnti di pensiero:

- Ricezione di e-mail non richieste
- Ricezione di e-mail commerciali non richieste
- "Junk mail" ossia e-mail "spazzatura"
- Ricezione di grosse quantità di e-mail non richieste

Tuttavia quale tra queste definizioni sia corretta è ancora una discussione aperta

L'origine del termine Spam

L'origine del termine non è noto da dove deriva.

Molti credono che l'origine derivi da una scenetta televisiva americana intitolata "Monty Python's".

Tuttavia il significato vero e proprio sta a significare "carne in scatola"



Da chi viene effettuato

- Colui che invia simultaneamente una massiccia quantità di e-mail a scopo pubblicitario senza l'esplicito consenso da parte del destinatario, viene definito Spammer.
- Lo spammer può essere una persona fisica o una persona giuridica che intende pubblicizzare un proprio prodotto, oppure il prodotto di terzi

Statistiche sullo Spam

Le categorie più diffuse di spam sono:

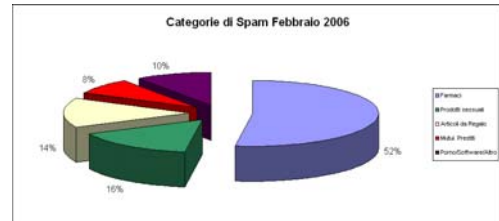
- Farmaci oltre il 50% (generalmente Viagra e Cialis)
- Prodotti sessuali 16%
- Articoli da regalo con il 14%
- Mutui e prestiti con l'8%
- Altri prodotti al 10%

Rispetto agli stessi mesi del 2005:

- Farmaci + 20%
- Articoli da regalo + 7%
- Mutui e prestiti - 9%
- Altre categorie +0%

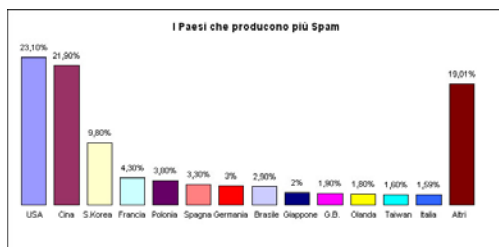
Fonte: Commtouch Software Ltd, aggiornate al febbraio 2006

Statistiche sullo Spam



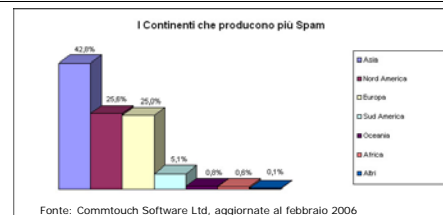
Fonte: Commtouch Software Ltd, aggiornate al febbraio 2006

Statistiche sullo Spam



Fonte: Commtouch Software Ltd, aggiornate al febbraio 2006

Statistiche sullo Spam



Fonte: Commtouch Software Ltd, aggiornate al febbraio 2006

INDICE

- Introduzione
 - La definizione di SPAM
 - L'origine del termine SPAM
 - Da chi viene effettuato
- **Perchè lo SPAM è un problema?** ←
- Come prevenire lo SPAM
- Metodi per combattere lo SPAM
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- Come opera uno spammer
- Lo SPAM è illegale?

Perchè lo SPAM è un problema?

- Implicazioni di natura economica
- Tempo perduto dai destinatari per scaricare, verificare e cancellare il messaggio
- Costo di banda sostenuto dall'ISP e dall'utente
- Intasamento della Rete
- Truffe e virus
- Pornografia in casa e ufficio

INDICE

- Introduzione
 - La definizione di SPAM
 - L'origine del termine SPAM
 - Da chi viene effettuato
- Perché lo SPAM è un problema?
- Come prevenire lo SPAM ←
- Metodi per combattere lo SPAM
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- Come opera uno spammer
- Lo SPAM è illegale?

Come prevenire lo Spam

- Non mettere l'indirizzo e-mail nelle pagine web e nei forum, blog ecc.
- Non rispondere mai allo spam
- Dare l'indirizzo e-mail soltanto alle persone strettamente indispensabili
- Non immettere l'indirizzo e-mail nel browser
- Scegliere un nome utente lungo almeno dieci caratteri
- Usare e fare usare sempre la "copia carbone nascosta"
- Alterare il proprio indirizzo e-mail quando lo si inserisce in una pagine web

INDICE

- Introduzione
 - La definizione di SPAM
 - L'origine del termine SPAM
 - Da chi viene effettuato
- Perché lo SPAM è un problema?
- Come prevenire lo SPAM
- Metodi per combattere lo SPAM ←
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- Come opera uno spammer
- Lo SPAM è illegale?

Metodi per combattere lo Spam

- Soluzioni poco efficaci
 - Presentare denunce al Garante della privacy
 - Denunciare il caso al proprio provider
 - Usare "liste bianche" o "liste nere"
 - Mandare messaggi che sembrano indicare che il vostro indirizzo è inesistente
- Soluzioni efficaci
 - Cambiare il proprio indirizzo e-mail quando possibile
 - Usare filtri antispam
 - Processo di Language Classification

Metodi per combattere lo Spam Filtri Antispam

- Filtro Statico
- Filtro Euristico
- Filtro Bayesiano

Analizziamo ora gli aspetti più importanti di ognuno di essi

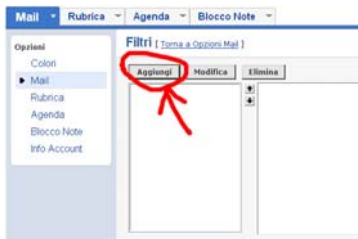
Metodi per combattere lo Spam Filtro Statico

- Il filtro statico funziona confrontando il messaggio in arrivo con un elenco di parole chiave, impostate dall'utente, che sono tipiche dello spam. Se il messaggio contiene una di queste parole viene classificato come spam
- Vantaggi: facile da implementare
- Svantaggi: elevato numero di falsi positivi

Vediamo un esempio di come realizzare tale filtro con una mailbox di yahoo.

Metodi per combattere lo Spam Filtro Statico: Esempio

Per creare un filtro cliccare sul link **Opzioni Mail**.
1. Cliccare su **Aggiungi**



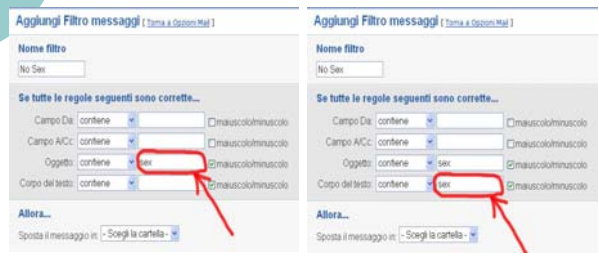
Metodi per combattere lo Spam Filtro Statico: Esempio (continuazione)

2. Scegliere un nome per il filtro (es. No Sex)



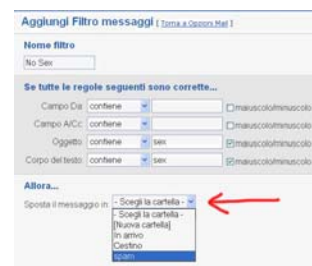
Metodi per combattere lo Spam Filtro Statico: Esempio (continuazione)

3. Scegliere il criterio in base al quale creare la corrispondenza. Ad esempio nel campo oggetto si sceglie l'opzione "contiene", immettendo la stringa di testo da confrontare.



Metodi per combattere lo Spam Filtro Statico: Esempio (continuazione)

4. Scegliere quindi la cartella a cui si desidera che il messaggio venga consegnato. Ad esempio la cartella spam.



Metodi per combattere lo Spam Filtro Statico: Esempio (continuazione)

5. Se si creano più filtri l'ordine è molto importante perché il filtro 1 verrà applicato per primo a un messaggio in arrivo, seguito dal filtro 2 e così via. Se il messaggio non corrisponde ad alcun filtro, verrà consegnato semplicemente nella cartella "In arrivo".



Metodi per combattere lo Spam Filtro Euristico

- Si basa nel creare una lista di parole o frasi, tipiche dello spam, ad ognuna delle quali viene assegnato un punteggio numerico.
- Quando arriva un messaggio esso viene analizzato ricercando al suo interno le parole chiave definite dall'amministratore, il risultato di tale analisi è un valore numerico calcolato facendo la somma di ogni punteggio di frase o parola trovata.
- Se il risultato ottenuto supera il valore di soglia, fissato in precedenza dall'amministratore, il messaggio viene classificato come spam.
- Vantaggi: efficaci se adeguatamente implementati
- Svantaggi: necessità di continui aggiornamenti

Metodi per combattere lo Spam Filtro Bayesiano

- Si basa sul principio che molti eventi sono dipendenti e che la probabilità di un evento futuro si può dedurre dagli eventi passati.
- Si parte con un insieme di messaggi di spam e un insieme di messaggi che non sono spam.
- Si effettua una scansione contando il numero di volte che ogni termine appare in ogni insieme, ottenendo così due database (Spam e Ham).
- Un terzo database contiene per ogni termine la probabilità che esso sia spam.

Metodi per combattere lo Spam Filtro Bayesiano (continuazione)

La probabilità che il termine X sia spam è calcolata come segue:

$$Pr(X) = \frac{(OS S)}{(OS S + OH H)}$$

dove:

S = numero di e-mail di tipo SPAM

OS = numero di volte in cui X è presente nei messaggi di SPAM

H = numero di e-mail di tipo HAM

OH = numero di volte in cui X è presente nei messaggi di HAM

Metodi per combattere lo Spam Filtro Bayesiano (continuazione)

La probabilità che un messaggio sia spam si calcola nel seguente modo:

- Si sommano le probabilità di tutti i termini contenuti nel messaggio, sia P tale valore
- Ai termini non presenti nel database delle probabilità si assegna loro una probabilità compresa tra 0 e 1. Tale valore viene stabilito in precedenza dall'amministratore.
- Si calcola la media aritmetica di P, sia M tale valore

Se M è maggiore di un certo valore di soglia, per esempio 0.7, allora l'e-mail è classificata come spam.

Metodi per combattere lo Spam Filtro Bayesiano (continuazione)

- Vantaggi: identifica fino al 99% di spam
- Svantaggi: necessita di una fase di inizializzazione di parecchi mesi prima che la sua efficacia sia massima

Metodi per combattere lo Spam Filtro Bayesiano: Esempio

Esempio:

numero di messaggi di SPAM (S) = 224

numero di messaggi di HAM (H) = 112

Termine	OS	OH	Probabilità
fun	19	9	0.5135
girlfriend	4	0	1
numbers	0	7	0
cell	8	30	0.1176
the	96	48	0.5000
vehicle	11	3	0.6470
viagra	20	1	0.9090

dove:

OS = numero di volte in cui il termine compare in un messaggio di spam

OH = numero di volte in cui il termine compare in un messaggio di ham

Metodi per combattere lo Spam Language Classification

- In tale processo si identificano parole e termini contenute all'interno di una e-mail per stabilire se si tratta di possibile Spam oppure no
- Il filtro Bayesiano si basa sul processo di Language Classification

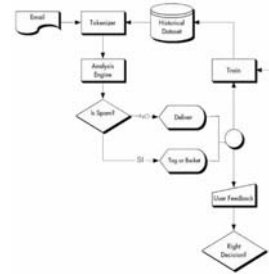
Consiste di 3 componenti principali

- Historical Dataset
- Tokenizer
- Analysis Engine

Language classification (2)

- Historical Dataset: è un database di tutti i termini rilevati all'interno di un'e-mail
- Tokenizer: ha il compito di scomporre l'e-mail in token, utilizzati successivamente dall'Analysis Engine
- Analysis Engine: stabilisce se l'e-mail è spam oppure no attraverso una matrice di decisione costruita inserendo al suo interno tutti i token individuati e il relativo valore di probabilità

Language Classification (3)



Schema del processo di Language Classification

Language Classification (4)

- Quando arriva un messaggio il tokenizer lo scompone in token.
- L'Analysis Engine prenderà la decisione di classificare il messaggio.
- Se il messaggio viene riconosciuto come Spam allora si può scegliere di etichettarlo (tag) o eliminarlo (bucket), altrimenti viene riconosciuto come messaggio legittimo.
- Successivamente si passa alla fase di train che istruisce il filtro

Language Classification (5)

Esistono diversi metodi per la fase di train:

- Train-Everything (TEFT)
- Train-on-Error (TOE)
- Train-Until-Mature (TUM)
- Train-Until-No-Errors (TUNE)

Language Classification Metodi per la fase di train

- **Train-Everything (TEFT)**
 - considera tutti i messaggi in arrivo
 - è implementato aggiungendo al database tutti i token contenuti nei messaggi
 - **Vantaggio:** il database viene costruito in base ai messaggi che riceve l'utente, infatti è in grado di adattarsi rapidamente anche a nuove tipologie di messaggio
 - **Svantaggio:** se si ricevono grandi quantità di spam e poche mail legittime, il filtro apprende la maggior parte dei termini come spam anche quando essi non lo sono.

Language Classification Metodi per la fase di train

- **Train-on-Error (TOE)**
 - viene utilizzato quando si ricevono grandi quantità giornaliere di e-mail
 - costruisce il database solo in presenza di errori
 - **Vantaggio:** Pochi cambiamenti del database
 - **Svantaggio:** Non adatto se si ricevono e-mail di diverse tipologie

Language Classification Metodi per la fase di train

- **Train-Until-Mature (TUM)**
 - si basa sia su TEFT che su TOE
 - apprende nuovi tipi di token da ogni messaggio finché non raccoglie un numero sufficiente di dati per effettuare al meglio una valutazione
 - apprende quando si verifica un errore
- **Train-Until-No-Errors (TUNE)**
 - istruisce il database finché non ci sono errori
 - richiede uno sforzo eccessivo da parte dell'utente in termine di tempo necessario a verificare se il filtro commette errori

Language Classification Esempio

```
From: "Julie Ellison" <gcq@swaniggy@bcglobal.net>
Reply-To: "Julie Ellison" <gcq@swaniggy@bcglobal.net>
Subject: Don't Pay For Name Brand Drugs
Date: Sun, 11 Apr 2004 10:21:05 -0600
Content-Type: text/plain

CANADIAN GENERICS NOW HAS VALUE!

Know where to find discounted Prescriptions? Buy your personal prescription
drugs on the internet and save! Allergies, Weight Loss, Muscle and Pain Relief
Men and Womens Health, heartburn, migraines, Impotence Get meds from Canada
here:

http://sacribbieheterozygous.zsstrom2.com/gp/default.asp?id=qm03
Order Some HERE
```

Language Classification Esempio (continuazione)

1° Passo: Tokenizzare il messaggio

From:Julie	From:Ellison	From:gcq@swaniggy	From:bcglobal
From:Julie	Reply-To:Julie	Reply-To:Ellison	Reply-To:Julie
Subject:Don't	Subject:Pay	Subject:For	Subject:Name
Subject:Brand	Subject:Drugs	Date:Sun	Date:Apr
Content-Type:text	Content-Type:plain	Url:Http	Reply-To:bcglobal
Url:Http:bcglobal.net	sgour	Url:zstrom2	Url:None
Url:gp	Url:default	Url:asp	Url:Id
Url:qm03	Reply-To:gcq@swaniggy	CANADIAN	GENERICS
NOWHAS	VALUUM	Case	where
Weight	and	discounted	Prescriptions
Buy	your	order	from
drugs	on	the	internet
and	save	Allerges	to
Loss	HERE	Muscle	Pain
Relief	Men	Womens	Health
heartburn	impotence	Impotence	Get
meds	from	Canada	here
Order	Some	personal	

Language Classification Esempio (continuazione)

2° Passo: Costruzione e valutazione della matrice di decisione

TOKEN	PROBABILITA'
NOW	0.99750
drugs	0.967430
meds	0.998432
Loss	0.999900
Date:Apr	0.999876
here	0.991238
Relief	0.999900
Weight	0.995478
Url:asp	0.938004

Content-Type:text	0.932180
and	0.885617
personal	0.127251
HERE	0.856461
Url:gp	0.851932
Pain	0.830545

Language Classification Esempio (continuazione)

3° Passo: train

- Dipende dal tipo di train che si vuole utilizzare (TEFT, TOE, TUM, TUNE).

4° Passo: Correzione errori

- In questa fase l'utente può correggere eventuali errori di decisione da parte del filtro.

INDICE

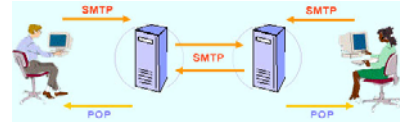
- Introduzione
 - La definizione di SPAM
 - L'origine del termine SPAM
 - Da chi viene effettuato
- Perché lo SPAM è un problema?
- Come prevenire lo SPAM
- Metodi per combattere lo SPAM
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- **Come opera uno spammer** ←
- Lo SPAM è illegale?

Cenni sul funzionamento della posta elettronica

Prima di analizzare le modalità di azione degli spammer è utile fornire una breve descrizione di come funziona la posta elettronica.

- I computer coinvolti nella spedizione di un messaggio di posta elettronica sono:
 - computer del mittente
 - computer del destinatario
 - i server e-mail
- Il server e-mail è un computer su cui risiede fisicamente la casella di posta elettronica dell'utente

Cenni sul funzionamento della posta elettronica



- Il mittente non contatta direttamente il server e-mail del destinatario, in quanto il client di posta è configurato in modo da dialogare sempre con lo stesso server
- Sarà il server a preoccuparsi di inoltrare il messaggio al server e-mail del destinatario.

Come opera uno Spammer

- Lo spammer può inviare un alto numero di e-mail nei seguenti modi:
 - Tramite ISP
 - Tramite Realy aperto
 - Tramite Relay multihop

Come opera uno Spammer Tramite ISP

- Ci sono ISP consenzienti i quali mettono a disposizione i loro server per l'invio di e-mail di spam dietro pagamento.
- Tuttavia l'ISP potrebbe non essere consenziente, ma, un po' per indolenza e un po' per necessità, non intervengono in maniera decisa in quanto lo spam è una fonte di guadagno per quegli ISP che vendono soluzioni antispam.

Come opera uno Spammer Tramite Relay aperto

- Sfruttano server SMTP mal configurati che consentono di trasmettere messaggi di posta a chiunque senza che il mittente ed il destinatario dell'e-mail appartengano al sistema locale

Come opera uno Spammer Tramite Relay multihop

- Simile al Relay aperto ma sfrutta server SMTP secondari
- Solitamente i server SMTP sono utilizzati all'interno di complesse reti aziendali.
- Lo spammer riesce ad inviare le proprie e-mail sfruttando un problema di relay aperto presente in un server secondario.
- Il server principale, il quale ha il compito di gestire le comunicazioni tra l'interno e l'esterno, invia l'e-mail in quanto la richiesta avviene da un host autorizzato.

Come fa uno Spammer a reperire gli indirizzi e-mail?

- Generando gli indirizzi in modo automatico
- Utilizzando i programmi chiamati "spider"
- Acquistando gli indirizzi direttamente dagli ISP

INDICE

- Introduzione
 - La definizione di SPAM
 - L'origine del termine SPAM
 - Da chi viene effettuato
- Perché lo SPAM è un problema?
- Come prevenire lo SPAM
- Metodi per combattere lo SPAM
 - Soluzioni meno efficaci
 - Filtri Antispam
 - Processo di Language Classification
- Come opera uno spammer
- Lo SPAM è illegale? ←

Lo Spam è illegale?

- Chi intende utilizzare le e-mail per comunicazioni commerciali e promozionali senza mettere in atto comportamenti illeciti deve tenere presente che:
 - è necessario il consenso informato del destinatario
 - il consenso è necessario anche quando gli indirizzi e-mail sono formati ed utilizzati automaticamente mediante un software
 - il consenso del destinatario deve essere chiesto prima dell'invio e solo dopo averlo informato chiaramente sugli scopi per i quali i suoi dati personali verranno usati
 - non è ammesso l'invio anonimo di messaggi pubblicitari
 - chi detiene i dati deve sempre assicurare di far valere i diritti riconosciuti dalla normativa sulla privacy
 - chi acquista banche dati con indirizzi di posta elettronica è tenuto ad accertare che gli interessati abbiano effettivamente dato il consenso all'invio di materiale pubblicitario
 - la formazione di appositi elenchi di chi intende ricevere e-mail pubblicitarie o di chi è contrario (le cosiddette "black list") non deve comportare oneri per gli interessati

Lo Spam è illegale?

Le sanzioni per chi viola le disposizioni di legge:

- Multa, in particolare per omessa informativa all'utente (fino a 90mila euro)
- Sanzione penale qualora l'uso illecito dei dati sia stato effettuato al fine di trarne un profitto o per arrecare ad altri un danno (reclusione da 6 mesi a 3 anni).
- Sanzione accessoria della pubblicazione della pronuncia penale di condanna o dell'ordinanza amministrativa di ingiunzione.

Bibliografia

- <http://www.garanteprivacy.it> (Normativa sulla privacy in materia di spam)
- <http://www.commtouch.com> (Statistiche sullo spam)
- <http://www.attivissimo.net> (Come difendersi dallo spam)
- <http://www.spin.it/spam> (Specifiche sullo spam)
- <http://www.sophos.it> (statistiche sullo spam)
- Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification by Jonathan A. Zdziarski