**Web Spam Taxonomy**

Web spamming si riferisce ad azioni volte a indurre in errore i motori di ricerca in classifica alcune pagine superiore che meritano. Recentemente, la quantità di web spamming ha aumentato drasticamente, portando ad una degradazione di ricerca risultati. Questo articolo presenta un taxonomy di tecniche di spam attuali, che riteniamo può aiutare a sviluppare contromisure adeguate.

**1 Introduzione**

Come sempre più persone si affidano sulla ricchezza di informazioni zione disponibile on-line, una maggiore esposizione sul World Wide Web può produrre notevoli guadagni finanziari per individui o organizzazioni. Più spesso, i motori di ricerca sono gli ingressi al Web; è per questo che alcune persone cercano di ingannare i motori di ricerca, in modo che le loro pagine diventino di alto rango nei risultati di ricerca, e, quindi catturare l'attenzione degli utenti.

Proprio come con le e-mail, si può parlare del fenomeno di *spamming* il Web. La conseguenza primaria di web spamming è che la qualità dei risultati di ricerca diminuisce. Per esempio, al momento della scrittura di questo articolo, il secondo risultato restituito da una ricerca importante motore per la query "farmacia Kaiser" è stata una pagina sul sito web di spam techdictionary.com. Questo sito contiene solo poche righe di informazioni utili (soprattutto alcune definizioni termine, probabilmente copiato da un vero dizionario), ma è costituito da migliaia di pagine, ciascuna ripetendo lo stesso contenuto e che punta a decine di

altre pagine. Tutte le pagine sono state probabilmente create per aumentare la classifica di alcuni altri, e nessuna di loro sembra essere particolarmente utile per chiunque cerchi farmacie convenzionate con Kaiser-Permanente.

La conseguenza secondaria di spamming è che gli indici dei motori di ricerca sono gonfiati con pagine inutili, aumentando il costo di ogni query eseguita.

Per fornire a basso costo, servizi di qualità, è fondamentale per motori di ricerca per indirizzo web spam. I motori di ricerca attualmente combattono lo spam con una varietà di tecniche, spesso manuali, ma per quanto ne sappiamo, non hanno ancora pienamente un efficace insieme di strumenti per combatterla. Crediamo che il primo passo nella lotta contro lo spam è la *comprensione* di essa, cioè, analizzando le tecniche che gli spammer utilizzano per ingannare i motori di ricerca. Una corretta comprensione della spamming può guidare lo sviluppo di contromisure adeguate.

A tal fine, in questo lavoro si organizzano tecniche di web Spamming in una tassonomia che può fornire un quadro per la lotta antispam. Forniamo anche una panoramica delle statistiche pubblicate su web spam che sottolinea l'entità del problema.

Ci sono stati brevi discussioni di spam nella letteratura scientifica [3, 6, 12]. Si può anche trovare i dettagli per diverse tecniche specifiche sul Web stesso (ad esempio, [11]). Tuttavia, riteniamo che questo documento offra la prima tassonomia completa di tutte le più importanti tecniche di spamming note finora. Per costruire la nostra tassonomia, abbiamo lavorato a stretto contatto con gli esperti in una delle principali società di Motore di ricerca, basandosi sulla loro esperienza, mentre allo stesso tempo indagando numerosi casi di spam per conto nostro.

Alcuni lettori potrebbero mettere in discussione la saggezza di rivelare segreti spamming, preoccupati che questo possa incoraggiare spamming supplementare. Assicuriamo lettori che nulla in questo lavoro è segreto per gli spammer; sono gli utenti del web più che altro a che non hanno familiarità con le tecniche presentate qui. Noi crediamo che pubblicizzando queste

tecniche di spamming sensibilizziamo la ricerca d’interesse della comunità.

**2 Definizione**

L'obiettivo di un motore di ricerca è quello di fornire alta qualità di risultati identificando correttamente tutte le pagine web che sono *rilevanti* per una query specifica, e la presentazione

l'utente con alcuni dei più *importanti* di questi pagine rilevanti. Rilevanza è di solito misurata attraverso la somiglianza testuale tra la query e una pagina. Le pagine possono dare una specifica query, un numero rilevante di punteggio; più alto è il numero, e più rilevante è la

pagina alla query. Importanza riferisce alla popolarità globale (Query-indipendente) di una pagina, come spesso dedotto dalla struttura di link (ad esempio, pagine con molti link in entrata sono più importanti), o forse altri indicatori. In pratica, i motori di ricerca di solito combinano rilevanza e importanza, calcolando un punteggio combinato che viene utilizzato per ordinare i risultati della query presentati per l'utente.

Usiamo il termine *spamming* (anche, *spamdexing*) per riferire a qualsiasi azione umana deliberata, che ha lo scopo di innescare una rilevanza ingiustamente favorevole o importanza

distanza di qualche pagina web, considerando la propria vera pagina valore. Useremo lo *spam* aggettivo per contrassegnare tutti coloro oggetti Web (elementi di contenuto pagina o link) che sono la risultato di una qualche forma di spamming. Le persone che svolgono spamming sono chiamati *spammers*.

Si può trovare sul World Wide Web una manciata di altre definizioni di web spamming. Per esempio, alcuni delle definizioni (ad esempio, [13]) sono vicini al nostro, affermando che qualsiasi modifica effettuata ad una pagina solo perché esistono motori di ricerca è spamming. Organizzazione specifica zioni o gruppi di utenti web (ad esempio, [9]) definiscono spamming da enumerare alcune delle tecniche che vi presentiamo nelle sezioni 3 e 4.

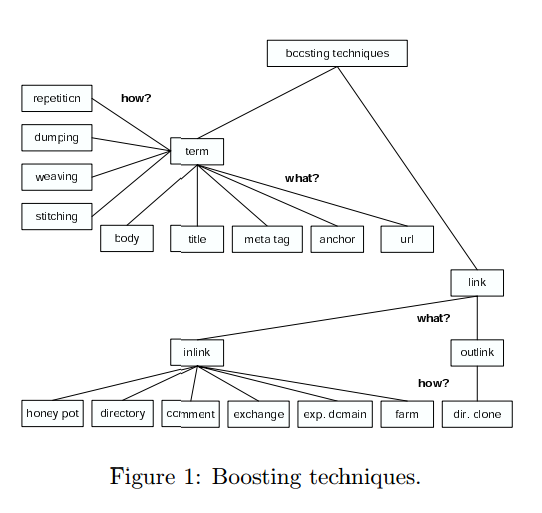
Una voce importante nell'arena web di spam è quella del motore di ricerca ottimizzatori (SEO), tale come SEO Inc. (Www.seoinc.com) o Bruce Clay (Www.bruceclay.com). L'attività di alcuni SEO benefici tutta la comunità web, in quanto aiutano autori creano ben strutturati, pagine di alta qualità. Come-mai, la maggior parte dei SEO impegnarsi in pratiche che noi chiamiamo Spamming. Per esempio, ci sono SEO che definiscono Spamming esclusivamente come sempre maggiore rilevanza per le query non legati al tema (s) della pagina. Questi SEO approva e le tecniche pratiche che hanno un impatto su importanti punteggi, per ottenere quello che chiamano posizionamento di pagina web "etico" o di ottimizzazione. Si prega di notare che in accordo con la nostra definizione, tutti i tipi di azioni destinate ad aumentare ranking (o rilevanza, o l'importanza, o entrambi), senza migliorare il vero valore di una pagina, sono considerate spamming.

Ci sono due categorie di tecniche associate con web spam. La prima categoria comprende il Boost , vale a dire, i metodi attraverso i quali si cerca per ottenere un'elevata rilevanza e / o l'importanza per alcune pagine. La seconda categoria comprende le tecniche di nascondigli,

metodi che di per sé non influenzano la ricerca

motore di algoritmi di ranking, ma che vengono utilizzati per nascondere

le tecniche di amplificazione adottate dagli occhi degli utenti di Internet. Le due sezioni seguenti discutono ognuna di queste due categorie in modo più dettagliato.



**3 Tecniche di Boosting**

In questa sezione presentiamo tecniche di spamming che tecniche atte a influenzare gli algoritmi di ranking utilizzati dai motori di ricerca.

La figura 1 illustra la tassonomia, al fine di guidare la nostro discussione.

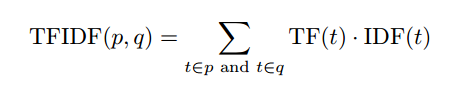
**3.1 Termine Spamming**

Nel valutare la pertinenza del testo, i motori di ricerca considerano dove su un termini di ricerca pagina web si verifica. Ogni tipo di posizione viene denominata *campo*. I campi di testo comuni per una pagina *p* sono il corpo del documento, il titolo, i meta tag nell'intestazione HTML e URL della pagina di *p*. In aggiunta, i testi di ancoraggio associati URL che puntano a

*p* sono considerati appartenenti alla pagina *p* (testo di ancoraggio campo), dal momento che spesso descrivono molto bene il contenuto di *p*. I termini in campi di testo di *p* vengono utilizzati per determinare il rilevanza di p rispetto ad una query specifica (un gruppo di termini di ricerca), spesso con diversi pesi dati a diversi campi. Termine spamming si riferisce alle tecniche che misura il contenuto di questi campi di testo al fine di rendere le pagine di spam rilevanti per alcune query.

**3.1.1 Algoritmi di destinazione**

Gli algoritmi utilizzati dai motori di ricerca per classificare web pagine basate sui loro campi di testo utilizzano varie forme di la metrica TFIDF fondamentale utilizzato in recupero di informazioni [1]. Dato un campo di testo specifica, per ogni termine t che è comune per il campo di testo e una query, TF (t) è la frequenza di tale termine nel campo di testo. Per esempio, se il termine "mela" appare 6 volte nel corpo del documento che si compone di un totale di 30 termini, TF ("Apple") è 6/30 = 0,2. La frequenza inversa del documento IDF (t) di un termine t è correlato al numero dei documenti della collezione che contengono t. Per istanza, se appare "apple" in 4 dei 40 documenti nella collezione, la sua IDF ("apple") punteggio sarà 10. Il Punteggio TFIDF di una pagina p rispetto ad una query q è quindi calcolato su tutti i termini t comuni:



Con punteggi TFIDF in mente, gli spammer possono avere due obiettivi: o per fare una pagina rilevante per un numero elevato di query (cioè, per ricevere un non-zero TFIDF score), o per fare una pagina molto importante per una query specifica (cioè, per ricevere un punteggio alto TFIDF). Il primo obiettivo può essere ottenere con l'inclusione di un gran numero di termini distinti in un documento. Il secondo obiettivo può essere raggiunto ripetendo alcuni termini "mirati". (Possiamo supporre che spammer non possono avere un controllo reale sui punteggi IDF dei termini. Inoltre, alcuni motori di ricerca ignorano IDF punteggi complessivamente. Così, il modo principale di aumentare i punteggi TFIDF è aumentando la frequenza di termini all'interno dei campi di testo specifiche di una pagina.)

**3.1.2 Tecniche**

Tecniche Term spamming possono essere raggruppati in base il campo di testo in cui si verifica l'spamming. Pertanto, distinguiamo:

* *Body Spam.* In questo caso, i termini di spam sono incluso nel corpo del documento. Questa tecnica di spamming è tra la comparsa semplice e più quelli Ular, ed è quasi vecchia quanto i motori di ricerca stessi.
* *Title spam*. I motori di ricerca di oggi di solito danno un maggiore peso a termini che compaiono nel titolo un documento. Quindi, ha senso includere la termini di spam nel titolo del documento.
* *Meta tag spam*. I meta tag HTML che appaiono nell'intestazione del documento sono da sempre il target ottenere di spamming. A causa del pesante spamming, motori di ricerca attualmente danno scarsa priorità a questi tags, o addirittura li ignorano completamente. Qui è un semplice esempio di un meta tag keywords di spam: <Meta name = "keywords" content = "comprare, a buon mercato, macchine fotografiche, lenti, accessori, Nikon, Canon ">
* *Anchor text spam*. Come con il titolo del documento, i motori di ricerca assegnano maggiore peso per ancorare termini di testo, in quanto si suppone di offrire un festival Estate mary del documento punta. Pertanto, lo spam termini sono talvolta inclusi nel testo di ancoraggio i collegamenti ipertestuali HTML di una pagina. Si prega di notare che questa tecnica spamming è differente da quelli precedenti, nel senso che i termini di spam sono Non aggiunto a una pagina di destinazione stessa, ma l'altro pagine che puntano al bersaglio. Come testo di ancoraggio viene indicizzato per entrambe le pagine, lo spamming ha impatto per la classifica sia l'origine e la destinazione pagine. Una di spam semplice testo di ancoraggio è: <a href="target.html"> gratuiti, offerte, a buon mercato, economico, poco costoso, gratis </a>
* *URL spam*. Alcuni motori di ricerca anche abbattere l'URL di una pagina in un insieme di termini che vengono utilizzati per determinare la rilevanza della pagina. Per sfruttare questo, gli spammer a volte creare URL lunghi che includono sequenze di termini di spam. Per esempio, si potrebbe incontrare URL di spam come: buy-canon-rebel-20d-lens-case.camerasx.com, buy-nikon-d100-d70-lens-case.camerasx.com, . . . Alcuni spammer anche andare al punto di istituire un server DNS che risolve qualsiasi nome host all'interno un dominio.

Spesso, le tecniche di spamming sono combinate. Per istanze, testo di ancoraggio e l'URL spam è spesso incontrato insieme con il collegamento di spam, che sarà discusso nella sezione 3.2.2.

Un altro modo di raggruppare tecniche di spamming termine è basato sul tipo di termini che vengono aggiunti ai campi di testo. Corrispondentemente, abbiamo:

* *Repetition* di uno o alcuni termini specifici. Da questa parte, spammer ottenere una maggiore rilevanza per un documento rispetto ad un piccolo numero di interrogazione termini.
* *Dumping* (scarico) di un gran numero di termini correlati, spesso persino interi dizionari. In questo modo, gli spammer fare una certa pagina rilevante per molti diversi query. Dumping è efficace contro le query che includere relativamente rari, termini oscuri: per tale query, è probabile che solo un paio di pagine sono, quindi, anche una pagina di spam utili a bassa rilevanza / importanza appare tra i primi risultati.
* *Weaving* (Tessitura) di termini di spam in contenuto copiato. A volte gli spammer duplicati testo corpora (ad esempio, articoli di giornale) disponibili sul Web e inserto termini di spam in loro in posizioni casuali. Questa tecnica è efficace se l'argomento dell'originale testo reale era così rara che solo un piccolo numero di Esistono pagine pertinenti. Weaving è utilizzato anche per diluizione, cioè, per nascondere alcuni termini di spam ripetuti all'interno del testo, in modo che gli algoritmi dei motori di ricerca che filtra la ripetizione semplice sarebbe fuorviato. Un breve esempio di tessitura spam è: Ricordate, non solo biglietto aereo per dire il piano giusto biglietti cosa al posto giusto, ma lontano viaggi a basso costo più difficile ancora, per lasciare camere d'albergo non dette la cosa sbagliata al momento di vacanza allettante.
* *Phrase stitching* (Frase cuciture) viene utilizzata anche dagli spammer per creare contenuti in modo rapido. L'idea è quella di incollare insieme sentenze o frasi, forse provenienti da fonti diverse; la pagina di spam potrebbe quindi presentarsi per le query su qualsiasi degli argomenti delle frasi originali. Per esempio, uno spammer utilizzando questo documento come fonte potrebbe venire con il seguente collage: L'obiettivo di un motore di ricerca è quello di fornire alta risultati di qualità identificando correttamente. Ingiustificatamente Tecniche di amplificazione favorevoli, vale a dire, attraverso i metodi che si cerca si basa sulla individuazione di alcune caratteristiche comuni di pagine di spam.

**3.2 Link Spamming**

Oltre rilevanza metriche basate sui termini, i motori di ricerca contano anche su informazioni di collegamento per determinare l'importanza distanza di pagine web. Pertanto, gli spammer spesso creano strutture di collegamento che sperano aumenterebbe l'importanza di uno o più dei loro pagine.

**3.2.1 Target Algorithms (Algoritmi di destinazione)**

Per la nostra discussione degli algoritmi di mira da collegamento lo spam, si adotterà il seguente modello. Per un Spammer, ci sono tre tipi di pagine sul Web:

1. pagine *inaccessibili* sono quelle che uno spammer può Non modificare. Queste sono le pagine fuori portata; lo spammer non può influenzare i loro link in uscita. (Si noti che uno spammer può ancora puntare a inaccessibili pagine.)
2. pagine *accessibili* sono mantenuti da altri (presumibilmente non affiliati con lo spammer), ma può ancora essere modificato in modo limitato da uno spammer. Ad esempio, uno spammer può essere in grado di inviare un commento a un post di blog, e che possono commentare contenere un link a un sito di spam. Come infiltrazione ac- pagine spet- è di solito non è semplice, lasciare noi dire che uno spammer ha un budget limitato di m pagine accessibili. Per semplicità, assumiamo che al massimo un collegamento in uscita può essere aggiunto a ciascun pagina accessibile.
3. pagine proprie sono mantenuti dal spammer, che quindi ha il pieno controllo sui loro contenuti. Chiamiamo il gruppo di pagine proprie una Σ farm spam. A Spam- L'obiettivo di mer è quello di aumentare l'importanza di una o più delle sue proprie pagine. Per semplicità, dire vi è una sola pagina di destinazione t. C'è una certa costi di manutenzione (registrazione del dominio, web host- ing) associati con proprie pagine di uno spammer, così abbiamo può assumere che uno spammer ha un budget limitato di n tali pagine, non compresa la pagina di destinazione.

Con questo modello in mente, discutiamo i due ben algoritmi noti utilizzati per calcolare i punteggi importanza sulla base delle informazioni di collegamento.

**HITS**. L'algoritmo HITS originale è stato introdotto in [7] per classificare le pagine su un argomento specifico. E 'più com- mon, tuttavia, utilizzare l'algoritmo su tutte le pagine del Web assegna hub e autorità punteggi globali per ogni pagina. Secondo la definizione di circolare HITS, im- pagine hub portanti sono quelle che puntano a molti importanza pagine autorità impor-, mentre importanti pagine di autorità sono quelli puntato da molti hub. Un motore di ricerca che utilizza l'algoritmo HITS per classificare le pagine restituisce come interrogazione risultato una fusione delle pagine con il più alto mozzo e autorità punteggi.

Punteggi Hub possono essere facilmente vittime di spam aggiungendo uscente zione collegamenti a un gran numero di ben noto, stimabile pagine, ad esempio www.cnn.com o www.mit.edu. Così, un spammer dovrebbe aggiungere molti link in uscita verso il bersaglio

t pagina di aumentare il suo punteggio hub.

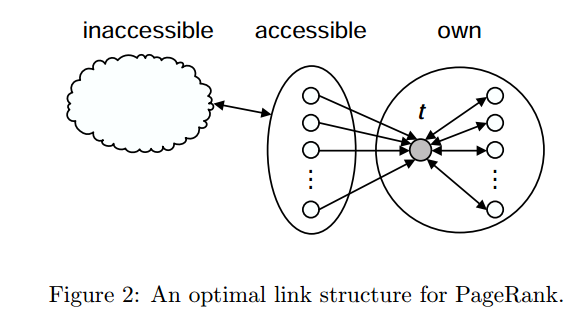
Ottenere un punteggio alto autorità è più compliance cato, in quanto implica avere molti link in entrata da mozzi presumibilmente importanti. Uno spammer potrebbe aumentare i punteggi del mozzo delle sue pagine n (ancora una volta, con l'aggiunta di molti link in uscita a loro) e poi fare quelle pagine puntano al bersaglio. Link da importante accessorio mozzi bili potrebbero aumentare l'autorità punteggio del bersaglio ancora di più. Pertanto, la regola qui è "il più il migliore ": entro i limiti del budget, la spammer dovrebbe avere tutte le pagine proprie e accessibili puntare al bersaglio. Non bersaglio proprie pagine dovrebbe anche Scegliere come molti altri (noti importante) autorità possibile.

**PageRank**. PageRank, come descritto in [10], usi informazioni di collegamento in entrata da assegnare importanza globale colonne sonore a tutte le pagine sul Web. Si presuppone che il

numero di link in entrata a una pagina è correlata a quella La popolarità di pagina tra gli utenti di Internet media (persone sarebbe puntare a pagine che trovano importante). Il intuizione dietro l'algoritmo è che una pagina web è importante se diverse altre pagine web importante punto

ad esso. Corrispondentemente, Page Rank si basa su un mutuale rinforzo tra le pagine: l'importanza di un certo le influenze di pagina e viene influenzato dal importanza di alcune altre pagine.

Una recente analisi dell'algoritmo [2] ha mostrato che il punteggio PageRank totale PR (Γ) di un gruppo Γ di pagine inaccessibile



(All'estremo, una singola pagina) dipende da quattro fattori:



dove PR statico è il componente punteggio causa della statica punteggio di distribuzione (salto casuale); PR in è il ri- punteggio vuto attraverso i link in entrata da pagine esterne; PR fuori

è il punteggio lasciando Γ attraverso i link in uscita a pagine esterne; e PR

*… da sistemare ….*

lavello

è la perdita a causa di punteggio

quelle pagine all'interno del gruppo che non hanno in uscita

links.

Per il nostro modello di azienda agricola di spam, i cavi formula precedenti

a una classe di strutture di collegamento ottimali che sono stati dimostrato

per massimizzare il punteggio della pagina di destinazione [4]. Uno di questi

struttura ottimale è presentato in figura 2; ha la

proprietà discutibilmente desiderabili che (1) che fa tutto proprio

pagine raggiungibili da quelli accessibili (in modo che

potrebbe essere strisciato da un motore di ricerca), e (2) che fa

ciò utilizzando un numero minimo di collegamenti. Possiamo osservare

come la struttura presentata massimizza totale Page-

Classifica punteggio della fattoria di spam, e di t pagina in parti-

lare:

1. Tutti disponibili pagine proprie n sono parte dello spam

farm, massimizzando il punteggio statico PR

statico

(Σ).

2. Tutte le pagine accessibili m puntano alla fattoria di spam,

massimizzando l'ingresso punteggio PR

in

(Σ).

3. Link che puntano al di fuori della fattoria di spam sono sostegno

pressato, rendendo PR

fuori

(Σ) pari a zero.

4. Tutte le pagine all'interno della fattoria hanno qualche uscita

link, rendendo un PR pari a zero

lavello

(Σ) componente punteggio.

All'interno del podere di spam, il punteggio della pagina t è

massima perché:

1. Tutte le pagine accessibili e proprio punto direttamente al

obiettivo, massimizzando la sua entrata punteggio PR

in

(T).

2. L'obiettivo punta a tutte le altre pagine proprie. Senza

tali collegamenti, t sarebbero avevano perso una parte significativa della

il suo punteggio (PR

lavello

(T)> 0), e le proprie pagine sarebbe

era stato irraggiungibile dall'esterno fattoria spam.

Si noti che non sarebbe saggio per aggiungere link da

l'obiettivo di pagine esterne alla fattoria, come quelli

diminuirebbe il PageRank totale dello spam

farm.

Come si può vedere in figura 2, la "più è meglio" regola

vale anche per PageRank. E 'vero che la creazione di

sofisticate strutture di collegamento all'interno di un'azienda agricola di spam fa

non di migliorare la classifica della pagina di destinazione. Tuttavia, un

spammer può raggiungere alto PageRank accumulando

molti link in entrata da pagine accessibili, e / o di

la creazione di grandi aziende di spam con tutte le pagine di puntamento

al bersaglio. I corrispondenti tecniche di spamming

sono presentati il prossimo.

3.2.2 Tecniche

Noi Link Group tecniche di spamming in base al fatto

aggiungono numerosi link in uscita a pagine popolari o

si riuniscono molti link in entrata a una sola pagina di destinazione

o un gruppo di pagine.

Link in uscita. Uno spammer potrebbe aggiungere manualmente un

numero di link in uscita a pagine ben note, sperando

per aumentare il punteggio hub della pagina. Contemporaneamente,

il metodo più diffuso per la creazione di una massiccia

numero di link in uscita è la clonazione directory: Si può

ricerca World Wide Web un numero di directory

siti, alcuni più grandi e meglio conosciuto (ad esempio, il DMOZ

Open Directory, dmoz.org, o la directory Yahoo,

dir.yahoo.com), altri più piccoli e meno famosi

(Ad esempio, indice del Bibliotecario a Internet, lii.org).

Queste directory organizzano i contenuti web su temi

e argomenti secondari, e Lista siti rilevanti per ciascuno. Spam-

mers poi spesso semplicemente replicare alcuni o tutti i

pagine di una directory, e creare così massiccia outgoing-

strutture di collegamento rapido.

Link in entrata. Per accumulare un mero

BER di link in entrata a una singola pagina o un set di

pagine, uno spammer potrebbe adottare alcuni dei seguenti

strategie:

• Creare un vaso di miele, una serie di pagine che forniscono

alcune risorse utili (ad esempio, copie di alcuni Unix

pagine di documentazione), ma che hanno anche (HID

den) collega alla pagina di destinazione di spam (s). Il miele

pot poi attrae la gente per puntare a esso, Boost-

ing indirettamente la classifica della pagina di destinazione (s).

Si prega di notare che il già citato direzione

cloni Tory potrebbero fungere da vasi di miele.

• Infiltrati una directory web. Diversi web directory

consentono ai webmaster di postare link ai loro siti sotto

qualche argomento nella directory. Può accadere che

la redazione di tali elenchi non controllano e

verificare aggiunte collegamento strettamente, o ottenere ingannato da un

spammer qualificati. In questi casi, gli spammer

può essere in grado di aggiungere directory pagine link che

indicare le loro pagine di destinazione. Come directory tendono

avere sia alto PageRank e mozzo colonne sonore, questo

tecnica di spamming è utile per stimolare sia la

PageRank e autorità decine di pagine di destinazione.

• link post sul blog, bacheche non moderati,

libri degli ospiti, o wiki. Come accennato in precedenza in Sezione

zione 3.2.1, gli spammer possono includere URL per loro

pagine di spam come parte del apparentemente innocente

commenti / messaggi che inviano. Senza un edi-

tor o un moderatore per sorvegliare tutto com- presentato

menti / messaggi, pagine del blog, message board,

o libro degli ospiti finiscono link a spam. Anche se

vi è un editor o un moderatore, potrebbe essere non

banale rilevare commenti spam / messaggi come

potrebbe impiegare alcune delle tecniche nascondigli pre-

sentato nella sezione successiva. Ecco un semplice ex

ampio di un commento nel blog di spam che caratterizza sia

collegamento e il testo di ancoraggio spamming:

Bella storia.

Leggi il mio <a href = "http: //

bestcasinoonlinever.com "> Las Vegas </a> casino

viaggio.

E 'importante ricordare che commento nel blog

spamming sta guadagnando popolarità, e non è solo

un problema per i motori di ricerca, ma ha anche una forte

influenze dirette sulla vasta comunità di milioni

leoni di blogger: per gli utenti di Internet con la propria

blog, commentare spamming rappresenta un fastidio

simile a e-mail spamming. Recentemente, un mero

BER di strumenti e iniziative sono state avviate per frenare

commento spamming. Per esempio, alcuni blogger

mantenere elenchi di nomi di dominio che appaiono in

URL di spam [8].

• Partecipare a scambio di link. Spesso, un gruppo

di spammer istituito una struttura di scambio di link, in modo da

che i loro siti puntano l'uno all'altro.

• Comprare domini scaduti. Quando un nome di dominio estre-

Pires, gli URL su vari altri siti web

punto di pagine all'interno indugiare dominio scaduto

per qualche tempo. Alcuni spammer acquistare scaduti

domini e li popolano con lo spam che prende

vantaggio del falso pertinenza / rilevanza con-

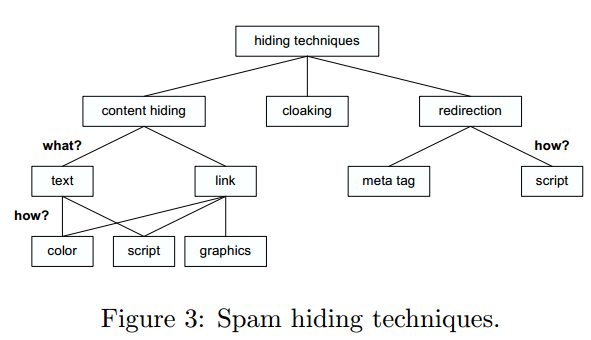
veyed dal pool di vecchi collegamenti.

• Creare fattoria spam. In questi giorni gli spammer possono

controllare un gran numero di siti e creare arbitrato

strutture di collegamento contrario che aumentare il ranking

di alcune pagine bersaglio. Mentre questo approccio è stato

 proibitivamente costose a pochi anni fa, oggi

è molto comune in quanto i costi di registrazione del dominio

e web hosting sono diminuiti drasticamente.

4 tecniche Nascondere

E 'normale per gli spammer di nascondere i segni rivelatori

(ad esempio, termini ripetute, lunghe liste di link) della loro atti-

atti-. Usano una serie di tecniche per nascondere la loro

abuso da parte degli utenti di Internet regolari visitare le pagine di spam, o

dai redattori di società del motore di ricerca che cercano di

individuare i casi di spam. Questa sezione offre una panoramica

dello spam più comune nascondere le tecniche, anche Estate di

marized in Figura 3.

4.1 Nascondere Content

Termini o link di spam in una pagina possono essere resi invisibili

quando il browser esegue il rendering della pagina. Uno tecno- comuni

nique sta usando combinazioni di colori appropriati: termini nella

corpo di un documento HTML non sono visibili se sono

visualizzato nel stesso colore dello sfondo. Colore

schemi possono essere definiti sia nel documento HTML

o in un foglio di stile CSS allegata (CSS). Mostriamo

un semplice esempio HTML seguente:

<Body background = "bianco">

<Font color = "bianco"> testo nascosto </ font>

. . .

</ Body>

In modo simile, i link di spam possono essere nascosti

evitando testo di ancoraggio. Invece, gli spammer spesso creano

piccoli, 1 × immagini ancoraggio 1-pixel che sono o traspa-

ent o background-color:

<a href="target.html"> <img src = "tinyimg.gif"> </a>

Uno spammer può anche utilizzare gli script per nascondere alcuni dei

elementi visivi nella pagina, per esempio, impostando

lo stile HTML visibile attributo su false.

4.2 Cloaking

Se gli spammer possono chiaramente identificare i client web crawler,

essi possono adottare la seguente strategia, denominata cloak-

ing: data una URL, web server di spam restituiscono uno specifico

Documento HTML a un browser web, mentre

restituire un documento diverso da un web crawler. Da questa parte,

gli spammer possono presentare il contenuto in definitiva destinazione

agli utenti del web (senza tracce di spam nella pagina),

e, allo stesso tempo, inviare un documento di spam

il motore di ricerca per l'indicizzazione.

L'identificazione di web crawler può essere fatto in

due modi. Da un lato, alcuni spammer mantengono

elenco degli indirizzi IP utilizzati dai motori di ricerca, e iden-

duare web crawler in base alle loro IP corrispondenti. Sul

D'altra parte, un server web può identificare l'applicazione

richiede un documento in base al campo user-agent in

il messaggio di richiesta HTTP. Ad esempio, nel follow

ing semplice messaggio di richiesta HTTP il nome utente-agent

è quello utilizzato dal Microsoft Internet Explorer 6

Browser:

GET pagine / db / members.html HTTP / 1.0

Host: www-db.stanford.edu

User-Agent: Mozilla / 4.0 (compatible; MSIE 6.0; Window

dows NT 5.1)

I nomi utente-agente non sono strettamente standardizzati,

ed è davvero l'applicazione richiede che cosa

da includere nel campo del messaggio corrispondente. Nostante

theless, crawler dei motori di ricerca si identificano per

un nome distinto da quelli usati da web tradizionale

applicazioni browser. Questo viene fatto al fine di consentire

webmaster di bloccare l'accesso ad alcuni dei contenuti,

parametri di traffico di rete di controllo, o addirittura eseguire

alcune buone intenzioni, ottimizzazioni legittimi. Per in-

posizione, alcuni siti servono ai motori di ricerca le versioni di

le pagine che sono privi di link di navigazione, indirizzi

vertisements e altri elementi visivi relativi alla

presentazione, ma non il contenuto. Questo tipo di ac-

tività potrebbe anche essere gradito da alcuni ricerca

motori, in quanto contribuisce indicizzare le informazioni utili.

4.3 reindirizzamento

Un altro modo di nascondere il contenuto di spam su una pagina è da

reindirizzando automaticamente il browser ad un altro URL

non appena la pagina viene caricata. In questo modo la pagina ancora

viene indicizzato dal motore di ricerca, ma l'utente non sarà

mai vedere da-pagine con reindirizzamento agiscono come intermediari

(o proxy, portoni) per gli obiettivi finali, che

gli spammer cercano di servire a un utente di raggiungere i loro siti

attraverso i motori di ricerca.

Reindirizzamento può essere ottenuto in vari modi.

Un approccio semplice è quello di approfittare del rinfresco

meta tag nell'intestazione di un documento HTML. Da

impostare il tempo di refresh a zero e l'URL di aggiornamento per

la pagina di destinazione, gli spammer può raggiungere reindirizzamento come

appena la pagina viene caricata nel browser:

<Meta http-equiv = "refresh" content =

"0; url = target.html">

Mentre l'approccio precedente non è difficile da attuare

zione, i motori di ricerca può facilmente identificare quali redirec-

zione tenta analizzando i meta tag. Altro sofisticatezza

spammer corticato raggiungere reindirizzamento come parte di alcuni

script nella pagina, come script non vengono eseguiti dal

crawler:

<Script language = "javascript"> <! -

location.replace ("target.html")

-> </ Script>

5 Statistics

Mentre abbiamo una buona comprensione di spamming

tecniche, i dati statistici a disposizione del pubblico de-

scribing l'importo e la natura del web spam è molto

limitata. In questa sezione passiamo in rassegna alcuni di ciò che è

noto.

Due documenti discutono la prevalenza di web spamming, pre-

che rappresen- deriva da tre esperimenti. Fetterly et al.

[3] valutata manualmente pagine di esempio di due diversi

insiemi di dati.

La prima serie di dati (DS1) rappresentato 150 milioni

URL che sono stati sottoposti a scansione più volte, una volta alla settimana

nel corso di un periodo di 11 settimane, dal novembre 2002 al feb-

braio 2003. Gli autori mantenuto lo 0,1% di tutti i cingolati

pagine, scelti in base a un hash degli URL. A Man-

Ispezione sessuale di 751 pagine campionato dal set di ri-

pagine nuto produssero 61 pagine di spam, che indica una preva-

lenza del 8,1% dello spam nel set di dati, con una fiducia

Intervallo di 1,95% al di confidenza del 95%.

Il secondo insieme di dati (DS2) è il risultato di un sin-

ricerca in ampiezza GLE iniziato a casa Yahoo!

pagina, condotto tra luglio e settembre 2002.

La ricerca ha riguardato circa 429 milioni di pagine. Durante

una valutazione manuale più tardi, da un campione casuale di

1.000 URL, gli autori sono stati in grado di scaricare 535

pagine, di cui 37 (6,9%) erano di spam.

Un terzo, insieme indipendente di statistiche è fornita da

Gyongyi et al. [5]. In questo caso, gli autori hanno utilizzato la

set completo di pagine strisciato e indicizzato dal Al-

motore di ricerca taVista a partire da agosto 2003. La diversi

miliardi di pagine web sono stati raggruppati in circa 31

milioni di siti Web (DS3), ciascuna corrispondente grosso modo

un host web individuale. Invece di campionamento casuale,

la seguente strategia è stata adottata: gli autori segmenti

mentato l'elenco dei siti in ordine decrescente PageRank

in 20 secchi. Ciascuno dei secchi conteneva una renza

numero ferenti di siti, con punteggi PageRank sommando

fino a 5 percento del PageRank totale. Di conseguenza,

il primo secchio conteneva i 86 siti con il più alto

Punteggi PageRank, secchio 2 il prossimo 665, mentre l'ultima

secchio contenente 5 milioni di siti che sono stati assegnati il

più bassi punteggi PageRank. La parte superiore della figura 4

mostra le dimensioni di ciascun segmento su una scala logaritmica.

Innanzitutto, un campione iniziale di 1000 siti stata costruita

selezionando 50 siti a caso da ciascun segmento.

Poi, il campione è stato ridotto a 748 siti esistenti

potrebbero essere classificati in modo chiaro. Un display di controllo manuale

coperto che 135 (18%) di questi siti erano spam. Il

parte inferiore della figura 4 presenta la frazione di spam

ciascun segmento. È interessante notare che quasi il 20%

del secondo PageRank secchio è spam, che indica che

alcuni spammer sofisticati possono raggiungere alte importanza

punteggi tanza. Inoltre, notare che vi è un alto preva-

lenza dello spam (quasi il 50%) in secchi 9 e 10. Questo

Infatti sembra indicare che gli spammer "medi" può

generare una notevole quantità di spam con mid-range

PageRank logaritmica.

La tabella 1 riassume i risultati dei tre pre-

Esperimenti presentate. Le differenze tra il ri-

figure di prevalenza portati potrebbe essere dovuto ad un gioco

di diversi fattori:

• Le scansioni sono state eseguite in tempi diversi. Esso

is possible that the amount of spam increased over

tempo.

• sono stati utilizzati differenti strategie di crawling

• There could be a difference between the fraction

of sites that are spam and the fraction of pages

that are spam. In other words, it could be the

case that the average number of pages per site is

different for spam and non-spam sites.

• Classification of spam could be subjective; indi-

viduals may have broader or narrower definition

of what constitutes spam.

Despite the discrepancies, we can probably safely esti-

mate that 10-15% of the content on the Web is spam.

As the previous discussion illustrates, our statistical

knowledge of web spam is sparse. It would be of inter-

est to have data not only on what fraction of pages or

sites is spam, but also on the relative sizes (as measured

in bytes) of spam and non-spam on the Web. Questo

would help us estimate what fraction of a search en-

gine's resources (disk space, crawling/indexing/query

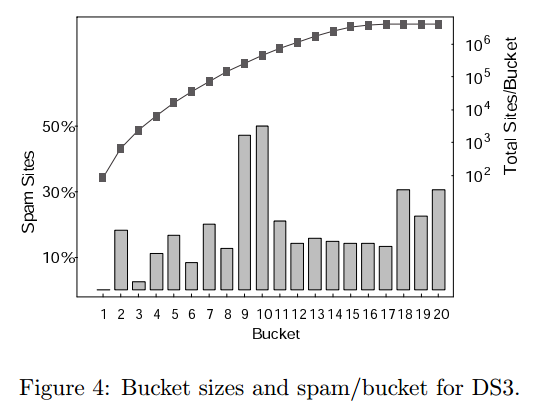
processing time) is wasted on spam. Another im-

portant question is how spam evolves over time. Fi-

nally, we do not yet know much about the relative fre-

quencies of different spamming techniques, and the co-

occurrence patterns between them. It is suspected that

 currently almost all spammers use link spamming, usu-

ally combined with anchor text spamming, but there

are no published research results supporting this hy-

pothesis. It is our hope that future research in the

field will provide some of the answers.

6 Conclusions

In this paper we presented a variety of commonly used

web spamming techniques, and organized them into a

taxonomy. We argue that such a structured discussion

of the subject is important to raise the awareness of

the research community. Our spam taxonomy natu-

rally leads to a similar taxonomy of countermeasures.

Correspondingly, we outline next the two approaches

that a search engine can adopt in combating spam.

On one hand, it is possible to address each of the

boosting and hiding technique presented in Sections 3

and 4 separately. Accordingly, one could:

1. Identify instances of spam, ie, find pages that

contain specific types of spam, and stop crawling

and/or indexing such pages. Search engines usu-

ally take advantage of a group of automatic or

semi-automatic, proprietary spam detection algo-

rithms and the expertise of human editors to pin-

point and remove spam pages from their indexes.

For instance, the techniques presented in [3] could

be used to identify some of the spam farms with

machine-generated structure/content.

2. Prevent spamming, that is, making specific spam-

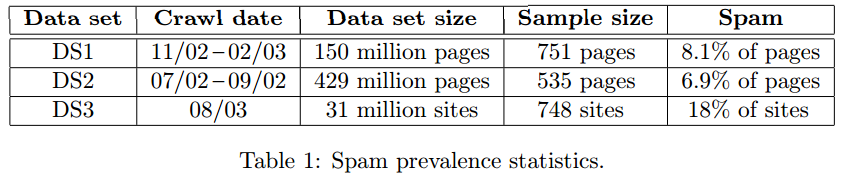
ming techniques impossible to use. Per esempio,

a search engine's crawler could identify itself as a

regular web browser application in order to avoid

cloaking.

3. Counterbalance the effect of spamming. Today’s



search engines use variations of the fundamental

ranking methods (discussed in Sections 3.1.1

and 3.2.1) that feature some degree of spam resilience.

On the other hand, it is also possible to address the

problem of spamming as a whole, despite the differences

among individual spamming techniques. This

approach relies on the identification of some common

features of spam pages. For instance, the spam detection

methods presented in [5] take advantage of the

approximate isolation of reputable, non-spam pages:

reputable web pages seldom point to spam. Thus, adequate

link analysis algorithms can be used to separate

reputable pages from any form of spam, without dealing

with each spamming technique individually.

Acknowledgement

This paper is the result of many interesting discussions

with one of our collaborators at a major search engine

company, who wishes to remain anonymous. We would

like to thank this person for the explanations and examples

that helped us shape the presented taxonomy

of web spam.

<http://airweb.cse.lehigh.edu/2005/gyongyi.pdf>