Capitolo 1

Introduzione

Con questa tesi si ha come obbiettivo di studiare e analizzare le varie tecniche di spam detection ed in particolare analizzare le tecniche online. Le tecninche verrano classificate sulla base dei segnali che utilizzano. Il fattore chiave è che non ci sono, o meglio sono poche, al momento tecniche online di spam detection, ovvero tecninche che rilevano lo spam durante la fase di crawling. Infatti quasi tutti i metodi tentano di fare il crawling dell'intero web e successivamente classificare le pagine in spam oppure buone.

Il feonomeno del web spam è sempre più presente all'interno del web. Questo è dovuto al fatto che gli utenti tendono ad esaminare solo i primi risultati ritornati dai motori di ricerca e quindi se un sito fa parte dei primi risultati di una un motore di ricerca ha un ritorno economico legato alla quantità di traffico che viene generata per quel sito. Uno studio del 2005 descritto in [3], stima che la perdita finanziaria mondiale causata dallo spam e di circa 50 miliardi di dollari e nel 2009 (come descritto in [4]) è salita a 130 miliardi di dollari.

Recentemente tutte le più grandi compagnie di motori di ricerca hanno identificato il recupero di informazioni non pertinenti come una delle priorità da risolvere. Le conseguenze del web spam possono essere [6]:

- la qualità delle ricerche è compromessa penalizzando i legittimi siti

web;

- un utente podrebbe perdere la fiducia sulla qualità di un motore di ricerca e perciò passare con facilità all'utilizzo di un altro;

- inoltre i siti spam possono essere usati come mezzo per malware, pubblicazione di contentuto per adulti e attacchi di tipo "fishing".Un prova tangibile si può vedere in [1], dove gli autori hanno eseguito l'algoritmo di PageRank su 100 milioni di pagine e hanno notato che 11 su i primi 20 risultati erano composti da siti pornogragici.

Queste considerazioni evidenziano che quando si progetta un motore di ricerca bisogna tenere conto delle pagine che potrebbero portare al mal funzionamento del motore stesso.

Il lavoro prodotto sarà utilizzato per essere integrato all'interno di un web crawler distribuito ad alte prestazioni. L'esigenza di tale modulo è sorta a seguito dello sviluppo, presso il Dipartimento, di un crawler chiamato *BubiNG*, altamente configurabile ma privo al momento di qualunque forma di rilevazione di siti e contenuti malevoli. Il problema è estremamente interessante sia dal punto di vista teorico che da quello pratico: infatti, sebbene siano numerose le tecniche descritte in letteratura per la determinazione di spam (usando come segnali sia il contenuto che la struttura dei link), è sorprendentemente scarso l'insieme di tali tecniche che possono essere usate on-line durante il crawl. Il problema diventa ancora più complesso se si aggiungono considerazioni legate ai vincoli di spazio di memoria disponibile e tempo di calcolo. Infatti in lettaratura il processo di spam detection viene eseguito subito dopo la fase di crawling. Ovvero il processo è composto dai seguenti passi:

- crawling dell'intero web;

- fase di spam detection;

- indicizzazione.

Questo modello è utile perché molte delle tecniche utilizzate fanno delle analisi sul grafo che risulta alla fine del processo di crawling. Da queste considerazioni noi proviamo a fare delle analisi per determinare se il processo di spam detection può essere fatto durante la fase di crawling ovvero al momento in cui il crawler esegue il "fetch" di una pagina determinare "on the fly" se la pagina è buona o ha un contenuto malevolo.

1.1 Ranking dei motori di ricerca

Prima di spiegare i vari metodi per fare web spam e successivamente quelli utili ad identificarlo, è necessario capire come i motori di ricerca sono capaci di valutare la rilevanza di una pagina web per una determinta query.

In linea di massima un sistema di reperimento di informazioni ovvero un motore di ricerca è dato da una collezione documentale D (un insieme di documenti) di dimensione N , da un insieme Q di interrogazioni, e da funzione di ranking ($r : Q \times D \mapsto R$) che assegna a ogni coppia formata da un'interrogazione e un documento un numero reale. L'idea è che a fronte di un'interrogazione a ogni documento viene assegnato un punteggio reale: i documenti con punteggio nullo non sono considerati rilevanti, mentre quelli a punteggio non nullo sono tanto più rilevanti quanto il punteggio è alto. In particolare i metodi di ranking si dividono in *endogeni* ed *esogeni*. I primi metodi fanno uso del contenuto del documento per valutarne la rilevanza mentre i secondi fanno uso di un struttura esterna che nel caso del web si rififerisce al grafo composto dai collegamenti ipertestuali tra le pagine. Tra i metodi esogeni sono di maggiore importanza *tf-idf* e *BM25* mentre tra quelli esogeni i più diffusi in letteratura sono *PageRank* e *HITS*.

1.1.1 Metodi di ranking endogeno

L'algoritmo usato dai motori di ricerca per fare il rank delle pagine web basandosi sui campi di testo usa varie forme del *tf-idf*. Il *tf-idf* e un metodo di ranking endogeno che utilizza il contenuto di una pagina per assegnarle un punteggio. Il *tf-idf* è una misura composta da due misure più semplici: la *Term Frequency* e la *Inverse Document Frequency*. Il primo metodo assegna a un documento d il punteggio dato dalla somma dei conteggi dei termini t dell'interrogazione che compaiono nel documento stesso. In questo modo documenti che hanno termini che compaiono più frequentemente avranno un punteggio più elevato. Utilizzare solo questo metodo non conviene in quanto è facilmente manipolabile. Inoltre non tiene cono del fatto che alcuni termini occorono più frequentemente non perché rilevanti, ma perché altamente frequenti all'iterno di *ogni* documento. Ad esempio le congiunzioni. Il secondo metodo è definito come l'inverso del numero di documenti nella collezione che contengono il termine t [5].

$$idf_t = \log \frac{N}{df_t} \tag{1.1}$$

La combinazione del *TF* ed dell' *IDF* produce una misura composta che permette di normalizzare il peso dei termini. Il *TF-IDF* di un documento d rispetto a uan query q è calcolato su tutti i termini t in comune come:

$$tf - idf(d, q) = \sum_{t \in d \; and \; t \in q} tf(t) \cdot idf(t) \tag{1.2}$$

Con il *TF-IDF* gli spammers possono avere due obbiettivi: o creare pagine rilevanti per un gran numero di query o creare pagine molto rilevanti per una specifica query. Il primo obbiettivo può essere finalizato includendo un gran numero di termini distinti in un documento. Il secondo attraverso la ripetizione di determinati nel documento. Ma il più delle volte i motori di ricerca non considerano l'*IDF* e perciò per incrementare il *TF-IDF* conviene incrementare la frequenza dei termini.

Anche se il *tf-idf* riesce a pesare abbastanza bene i vari termini ha molti limiti e per questo che il sistema di pesatura più attualmente usato è *BM25* che è uno schema di pesatura basato sul *modello probabilistico*. Questo schema è il risultato di uno studio puramente euristico.

1.2 Web spam

Con il termine web spamming si fa riferimento a tutti i metodi che tentano di manipolare gli algoritmi di ranking dei motori di ricerca per aumentare il valore di ranking di alcune pagine rispetto ad altre[2]. Dato il numero esorbitante di pagine che vengono create e pubblicate sul web, gli utenti competono per poter far comparire le proprie pagine tra le prime dei risultati di una query. Il fenomedo dello spamming o spamindexing comporta che la qualità delle ricerche decrementa, gli utenti tenderanno ad utilizzare altri motori di ricerca, l'indicizzazione di pagine che non sono utili e l'aumento del costo delle operazioni di query ed infine la causa di malware e reindirizzamento verso contenuto per adulti[6].

L'obbiettivo del motori di ricerca è di ottenere ottimi risultati per identificare tutte le pagine web che sono rilevanti per una specifica query e presentarle secondo l'importanza che esse hanno. Di norma la rilevanza viene misurata attraverso la similarità testuale tra la query e le pagine mentre l'importanza è definita come la popolarità globale della pagina e a volte è inferita dalla struttura dei link [2]. Ci sono due categorie di tecniche associate al web spam [2]:

- **tecniche boost** che cercano di far avere più importanza o rilevanza a delle pagine
- **tecniche hiding** che sono metodi per nascondere le tencinche di boost all'utente dal browser, anche se alcuni autori incorporano queste tecniche facenti parte delle tecniche di boost

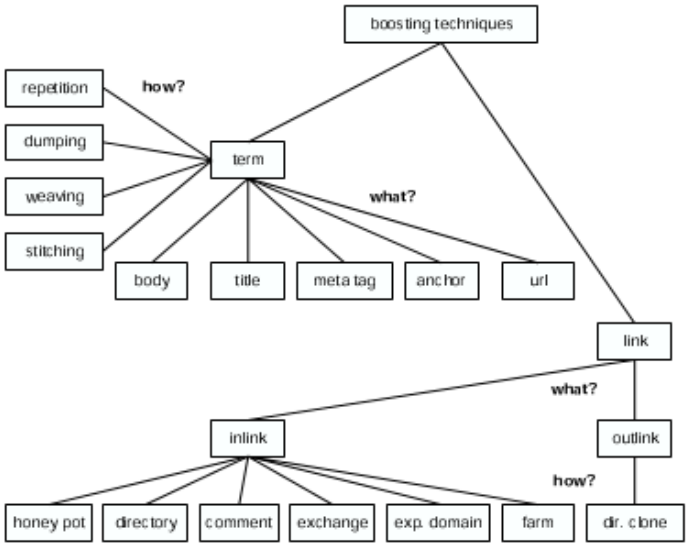

Figura 1.1: Tassonomia delle tecniche boost

1.2.1 Tecniche di boost

Le tecninche di boosting si dividono in: *Term Spamming* e *Link Spamming*. Con l'avvento degli algoritmi di ranking basati sulla struttura del grafo il textitTerm Spaming è stato trascurato. In figura 1.1 è specificata la tassonomia delle tecniche boost [2].

1.2.2 Term Spamming

Nel valutare la rilevanza testuale i motori di ricerca considerano dove i termini di una query compaiono in una pagina. Il tipo di punto all'interno della pagina è chiamato *campo*. I più comuni campi di testo per una pagina p sono: il body della pagina, il titolo, i meta tag nell'header HTML e l'URL della pagina. Inoltre il testo delle ancore (il tag a) associate all'URL che puntano alla pagina p sono considerati dal momento che possono descrivere molto bene il contenuto della pagina. I campi di testo di p sono utilizzati per detrminare la rilevanza di p rispetto ad una query alcune volte con differenti

proporziona. Il term spamming utilizza tecniche di pesatura dei contenuti dei campi di testo in modo tale da creare pagine di spam rilevanti per delle query [2].

Le tecninche di spamming possono essere raggruppate in base ai campi di testo dove viene fatto spamming. In base a questo distinguiamo:

- *Body Spam.* In questo caso lo spam è nel corpo del documento. Questo è lo spam più diffuso.
- *Title Spam.* Molti motori di ricerca danno molta importanza ai termini che compaiono nel titolo.
- *Meta Tag Sapm.* I tag che compaiono nell'header sono soggetti a spam. Per questo i motori di ricerca danno poca importanza a questi campi o non li considerano.
- *Anchor Text Spam.* I motori di ricerca assegnano un peso maggiore al testo nelle ancore. Perciò testo di spam è incluso nel testo delle ancore dei collegamenti HTML di una pagina. In questo caso lo spamming non viene fatto sulla pagina che si vuole far avere un rank più alto ma sulle pagine che puntano ad essa.
- *URL Spam.* Alcuni motori di ricerca dividono l'URL delle pagine in un insieme di termini che sono usati per determinare la rilevanza di una pagina. Per sfruttare questo si creano lunghi URL che includono una grande sequenza di spam di termini.

Queste tecniche possono essere utilizzate insieme.

Un altro modo per raggruppare queste tecniche si basa su che tipo di termini vengono utilizzati nei campi di testo. In questo modo vengono raggrupate in:

- *Ripetizione di uno o più specifici termini.*

- *Inclusione di molti termini generici per creare pagine rilevanti per molte query.*
- *Intreccio di vari termini all'interno della pagina.*
- *Creazione di frasi per creare contenuti veloci attraverso la concatenazione di frasi da fonti diverse.*

1.2.3 Link Spamming

In [2] viene definito che per uno spammer ci sono tre tipi di pagine nel Web: inaccessibili, accessibili(blog), proprietarie. Le inaccessibili sono quelle che uno spammer non può modificare. Le accessibili sono pagine gestite da altri ma che possono essere modificate un po' dallo spammer. Le proprietarie sono pagine degli spammer che hanno il pieno controllo. Il gruppo di pagine proprietarie è chiamato *spam farm* Σ.

Di norma i motori di ricerca utilizzano due algoritmi di ricerca per aumentare l'importanza basandosi sulle informazioni dei link: PageRank e HITS.

Ci sono due categorie principali di link spam: outgoing link spam e incomign link spam.

Outgoing link. E' uno dei metodi più facili da implementare in quanto basta aggiungere dei link a pagine conosciute considerate buone, sperando di poter aumentare il loro hub score. Per fare questo si possono utilizzare delle directory che contengono liste di siti come DMOZ o Yahoo!. Queste directory organizzano i contenuti web in contenuti e in liste di siti relativi.

Incoming link. Ci sono diverse strategie che si possono adottare in modo tale da avere un numero elevato di link in entrata:

- *Honeypot*: ovvero si creano un insieme di pagine che hanno un contenuto interessante (un esempio può essere una documentazione Lin-

ux) ma che hanno link nascosti alla pagina o alle pagine per cui deve aumentare il valore di rilevanza.

- *Infiltrarsi in una directory web*: molte directory web permettono ai webmasters di postare links ai loro siti che honno lo stesso contenuto.
- *Postare link nei blog, forum e wiki*: includere URL a pagine di spam come parte di un commento.
- *Scambio di link*: scambiare link con altre pagine di spam.
- *Comprare domini scaduti*: quando un dominio scade ci sono delle pagine che puntano ancora ad esso. Comprando questi domini e riempirli di spam ha dei vantaggi per la rilevanza che si acquisice dai link che puntano ancora ad essa .
- Creare una spam farm: con l'abbassamento dei costi si possono costruire delle spam farm che hanno come obbietivo di aumentare la rilevanza delle pagine di spam. Molte volte si utilizzano tecninche come il vaso di miele. In questo caso il valore di page rank aggregato delle pagine è propagato alla pagina che hanno come link. Delle forme più aggressive di honeypot è l'hijacking[6]. dove gli spammers prima attaccano un sito con una buona reputabilità e poi usano questo come parte della loro link farm.

1.2.4 Click Spamming

Dal momento che i motori di ricerca utilizzano i dati sul flusso di click per regolare le funzioni di ranking, gli spammers generano clik fraudolenti per manipolare il comportamento di queste funzioni in modo tale da fare avere un migliore rank i loro siti. Vengono fatte delle query e si clicca sulla pagina che si vuole aumentare il rank. Questo viene fatto in modo automatico

attraverso script che girano su diverse macchine per non fare sospettare di tale comportamento.

1.2.5 Tecninche di hiding

Le tecninche di hiding si possono classificare in: content hiding, cloaking, redirection.

Content hiding. I termini o link di spam possono essere nascosti quando il browser visualizza una pagina. Una tecnica è quella di utilizzare lo stesso colore per i termini e lo sfondo. Mentre per i link basta non inserire il testo all'interno delle ancore che indirizzano a una pagina. Un'altra tecninca è quella di utilizzare degli script per nascondere il contenuto.

Cloaking. E' facile identificare quando la richiesta di una pagina è fatta da un crawler o da un browser per questo si utilizza questa tecninca che dato un URL il server spam restituisce un diverso documento HTML a seconda se la richiesta è fatta da un crowler o un browser. Cosi all'utente viene presentata una pagina con del contenuto mentre si manda un documento di spam al motore di ricerca. La rilevazione di crowler può essere fatta in due modi: o si mantiene in memoria una lista di indirizzi di crawler oppure attraverso l'header della richiesta HTTP andando a vedere il campo user-agent se questo è diverso dai pi comuni browser allora può essere un crawler.

Redirection. Un altro modo è quello di reindirizzare il browser ad un altro URL appena la pagina è caricata.

Bibliografia

[1] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 309–318, New York, NY, USA, 2004. ACM.

[2] Zoltan Gyongyi and Hector Garcia-Molina. Web spam taxonomy. Technical Report 2004-25, Stanford InfoLab, March 2004.

[3] Nicholas R. Jennings. The global economic impact of spam. *Ferris Research*, 2005.

[4] Nicholas R. Jennings. Cost of spam is flattening – our 2009 predictions. *Ferris Research*, 2009.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[6] Nikita Spirin and Jiawei Han. Survey on web spam detection: Principles and algorithms. *SIGKDD Explor. Newsl.*, 13(2):50–64, May 2012.