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 sono 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 classi, di norma le classi sono due: 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 calcolati dai motori di ricerca e quindi se un sito fa parte degli n primi risultati, 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. Per questo motivo, 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 [10]:

- la qualità delle ricerche è compromessa penalizzando i legittimi siti web;
- un utente potrebbe 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 con contenuto per adulti.

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 è il risultato della 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 per 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 è il 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 conto del fatto che alcuni termini occorono più frequentemente non perché rilevanti, ma perché altamente frequenti all'interno 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 [6].

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

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 una 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 termini 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* [9] che è uno schema di pesatura basato sul *modello probabilistico*. Questo schema è il risultato di uno studio puramente euristico.

1.1.2 Metodi di ranking esogeno

Uno dei metodi esogeni è *PageRank* descritto in [8]. PageRank usa le informazioni portate dai link in entrata (*inlink*) per determinare un punteggio globale di importanza di una pagina. Esso assume che esiste un legame tra numero di *inlink* di una pagina p e la popolarità della pagina p. Il concetto fondamentale dietro *PageRank* è che una pagina è importante se moltre altre pagine importanti puntano ad essa. Questo concetto è mutualmente rinforzante ovvero l'importanza di una certa pagina influenza ed è influenzata dall'importanza delle altre pagine [2]. In dettaglio *PageRank* è basato sulla passeggiata naturale del grafo del web G. Più precisamente, la passeggiata viene pertubata nel seguente modo: fissato un parametro α tra 0 e 1, a ogni passo con probabilità α si segue un arco uscente, e con probabilità $1 - \alpha$ si sceglie un qualunque altro nodo del grafo utilizzando una qualche distribuzione v, detta vettore di preferenza (per esempio, uniforme). Assumendo che non esistano pozzi, la catena è quindi rappresentata dalla combinazione lineare:

$$\alpha G + (1 - \alpha)1v^T \tag{1.3}$$

dove G è la matrice della passeggiata naturale su G. Il fattore α è detto fattore di attenuazione di norma è impostato a un valore di 0,85.

Un altro metodo usato per il ranking delle pagine è *HITS (Hyperlink-Induced Topic Distillation)* introdotto in [5]. Differentemente da *PageRank* esso assegna due punteggi di importanza a ogni pagina: uno di *hubbines* e uno di *autorevolezza*. L'intuizione dietro a HITS è che invece di un singolo punteggio di importanza esista un concetto di pagina *autorevole*, cioè pagi-

na con contenuto pertinente e interessante, e di *hub*, cioè pagina contenente numerosi collegamenti a pagine autorevoli. I due concetti si rinforzano *mutuamente*: una pagina autorevole è puntata da molte pagine centrali, e una buona pagina centrale punta a molte pagine autorevoli.

Questo approccio considera che nel web ci sono due tipi di pagine: quelle che contengono dei contenuti per un determinato argomento (*authoritative*) e quelle che contengono tanti link a delle pagine *authoritative* che sono chiamate pagine *hub*. Le pagine *hub* sono utili per scoprire le pagine *authoritative* [7].

L'algoritmo parte da un sottografo del web ottenuto a partire da un'interrogazione. La selezione del sottografo può essere fatta in vari modi, un modo è quello di prendere un certo insieme di risultati ottenuto da un motore di base e generare un sottografo sulla base di una query e delle pagine che puntano a quelle ottentute dalla query. Per questo sottoinsieme di pagine otteniamo un matrice di adicenza A. I punteggi di *hub* e *authority* per tutte le pagine del sottoinsieme possono essere formalizzate dalla seguente coppia di equazioni:

$$\begin{cases} \vec{a} = A^T \vec{h} \\ \vec{h} = A \vec{a} \end{cases} \tag{1.4}$$

Può essere dimostrato che la soluzione per il sistema di equazioni 1.4 dopo una serie iterativa di calcoli converge rispettivamente al principale autovettore di AA^T e A^TA [7][10].

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 aumentarne il valore di alcune pagine rispetto ad altre [2]. Dato il numero esorbitante di pagine che vengono create e pubblicate sul web, gli utenti competono per far comparire le proprie pagine tra le prime dei risultati di una query. Il

fenomeno dello spamming o spamindexing ricade sulla qualità delle ricerche causando diversi problemi: indicizzazione di pagine che non sono utili, aumento del costo delle operazioni di query, malware e reindirizzamento verso contenuto per adulti ed infine che gli utenti che si sentono motivati ad utilizzare altri motori di ricerca,[10].

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

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 *Term 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 pag-

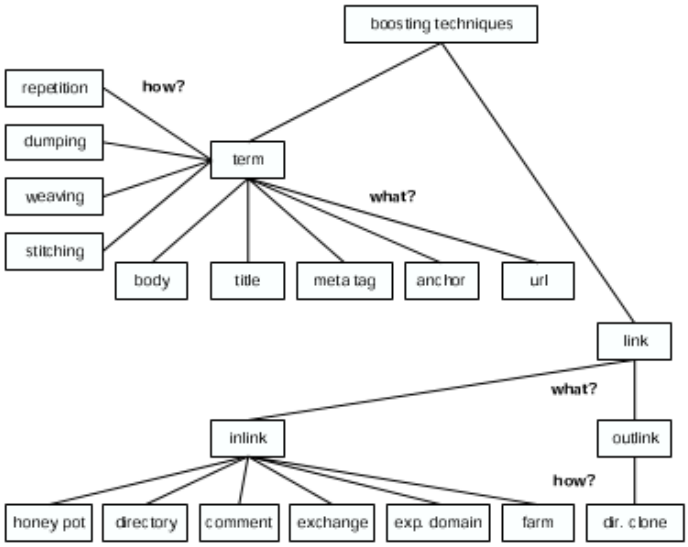

Figura 1.1: Tassonomia delle tecniche boost

ina p sono: il body della pagina, il titolo, i meta tag nell'header HTML e l'URL della pagina. Inoltre viene considerato anche come *campo*, il testo delle ancore (il tag a) associate all'URL che puntano alla pagina p dato che descrivere molto bene il contenuto della pagina . I campi di testo di p sono utilizzati per determinare la rilevanza di p rispetto ad una query (alcune volte i campi vengono pesati sulla base della loro importanza) e perciò chi fa *term spamming* utilizza tecniche di pesatura dei contenuti dei campi di testo in modo tale da aumentare l'efficacia dello spam [2]. Le tecninche di spamming possono essere raggruppate in base ai *campi* di testo dove viene fatto spamming. In base a questo distinguiamo [2]:

- *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. Quindi ha senso includere termini di spam all'interno del titolo della pagina.

- *Meta Tag Spam.* I tag che compaiono nell'header sono molto frequentemente soggetti a spam. Per questo i motori di ricerca danno poca importanza a questi campi o non li considerano. Di seguito viene mostrato un esempio di spam.

```
<meta name="keyword" content="buy, cheap, cameras,
  lens, accessories, nikon, canon">
```

- *Anchor Text Spam.* I motori di ricerca assegnano un peso maggiore al testo nelle ancore perché pensano che esse contengano un riassunto del contenuto della pagina. 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.

```
<a href="target.html">free, great deals, cheap,
  inexpensive, cheap, free</a>
```

- *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 metodo di ranking, gli spammers creano lunghi URL che includono una grande sequenza di termini spam, un esempio può essere: *buy-canon-rebel-20d-lens-case.camerasx.com.*

Queste tecniche possono essere utilizzate insieme o separatamente. Un altro modo per raggruppare queste tecniche si basa sul tipo di termini che vengono utilizzati nei campi di testo [2]:

- 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.

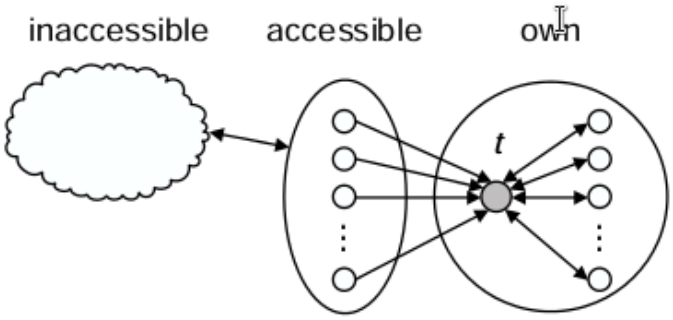

Figura 1.2: Tipi di pagine nel web per uno spammer

- Creazione di frasi di senso compiuto per l'elaborazione di contenuti generati velocemente 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) e proprietarie (fig. 1.2). Le inaccessibili sono quelle che uno spammer non può modificare. Le accessibili sono pagine gestite da altri ma che possono essere modificate livemente dallo spammer attraverso l'immissione di un post in un forum o in un blog o portali di questo genere. Le proprietarie sono pagine degli spammer di cui hanno il pieno controllo. Il gruppo di pagine proprietarie è chiamato *spam farm*.

Di norma i motori di ricerca utilizzano due algoritmi per aumentare l'importanza basandosi sulle informazioni dei link: PageRank e HITS e sulla base di questi due tipi di algoritmi vengon definite due categorie principali di link spam: *outgoing link spam* e *incomign link spam*. L'*Outgoing link* è uno dei metodi più facili da implementare in quanto basta aggiungere dei link nella propia pagina, ad altre pagine che sono considerate buone, sperando di poter aumentare il punteggio di *hub*. Per la ricerca di link da includere nella

pagina per cui si vuole incrementare il punteggio di *hub* 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. Per quanto riguarda *Incoming link*, ci sono diverse strategie che si possono adottare in modo tale da avere un numero elevato di link in entrata [2]:

- *Honeypot*: ovvero si creano un insieme di pagine che hanno un contenuto interessante (un esempio può essere una documentazione Linux) ma che hanno link nascosti alla pagina o alle pagine per cui si deve aumentare il valore di rilevanza.

- *Infiltrarsi in una directory web*: molte directory web permettono ai webmasters di postare links ai loro siti che hanno 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. Questa è una pratica comune tra chi fa spam ed esistono blog completamente finalizzati all'incontro di spammer per lo scambio dei link.

- *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 di un pargina spam detta *target page*, un esempio è mostrato in fig. 1.3. Molte volte si utilizzano tecninche come *honeypot*. In questo caso il valore di page rank aggregato delle pagine è propagato alla pagina target. Una delle forme più aggressive di honeypot è l'*hijacking* [10],

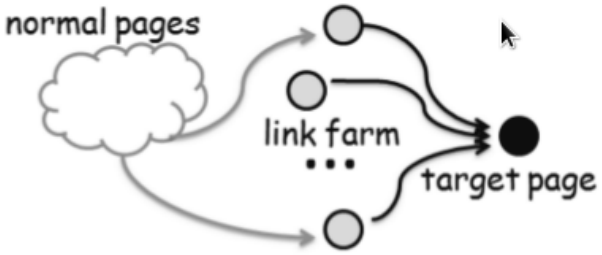

Figura 1.3: Esempio di una spamfarm

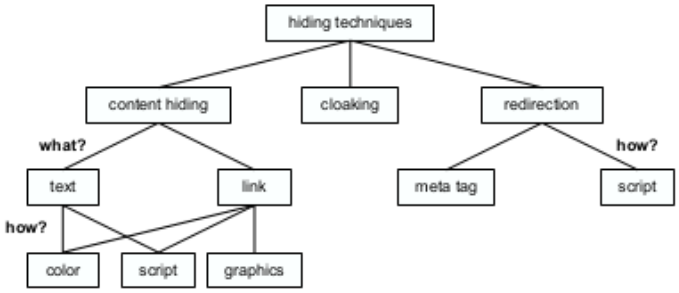

Figura 1.4: Tecniche di hiding

dove gli spammers prima attaccano un sito con una buona reputabilità e poi usano questo come parte della loro link farm.

1.2.4 Tecninche di hiding

Le tecninche di hiding si possono classificare in: *content hiding*, *cloaking*, *redirection* (fig. 1.4) [2]. Nel *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 tecnica è quella di utilizzare degli script per nascondere il contenuto. Il *Cloaking* sfrutta il fatto che è facile identificare quando la richiesta di una pagina è fatta da un crawler o da un browser. Perciò questa tecnica dato un

URL, il server spam restituisce un documento HTML diverso a seconda se la richiesta è fatta da un crawler o da un browser. Quindi vengono distribuiti due contenuti diversi in base se la richiesta al server spam è fatta da un crawler o da un browser. La rilevazione di un crawler 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 più comuni browser allora può essere un crawler. Nell'esempio sotto, lo user-agent della richiesta HTML indica un l'uso del web browser Chrome.

```
Mozilla/5.0 (X11; Linux x86_64)
AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/32.0.1700.102 Safari/537.36
```

La *Redirection* è un'altra tecnica che reindirizza il browser ad un altro URL appena la pagina è caricata. Un esempio di redirection server-side è mostrato di seguito.

```
header("Location: http://www.example.com/");
```

1.2.5 Click Spamming

Un ultimo metodo per fare web spam è il *Click Spamming* [10]. I motori di ricerca utilizzano dati sul flusso dei click per regolare le funzioni di ranking, quindi gli spammers generano clik fraudolenti per manipolare il comportamento di queste funzioni in modo tale da fare avere un rank migliore ai loro siti. Il metodo prevede che vengano fatte delle query e si clicchi sulla pagina che si vuole aumentare il rank. Tale metodo viene eseguito in modo automatico attraverso script che girano su diverse macchine per non fare sospettare delle numerose richieste provenienti da un unica macchina [10].

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] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999.

[6] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. Pages 117–119.

[7] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. Pages 474–476.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

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