Indice

Elenco delle figure

Capitolo 1

Introduzione

Questa tesi ha come obbiettivo lo studio e l'analisi delle tecniche di spam detection attualmente esistenti ed in particolare delle tecniche online. Nella prima parte le tecniche verranno classificate sulla base dei segnali che utilizzano. Successivamente verranno eseguiti dei test per valutare alcuni algoritmi di spam detection offline eseguiti durante la fase di crawling. Ed infine verrano presentati e discussi i risultati ottenuti. Al momento in cui si scrive non ci sono, o meglio sono poche, le tecniche online di spam detection, ovvero tecniche che rilevano lo spam durante la fase di crawling. Infatti quasi tutti i metodi tentano di fare il crawling dell'intera porzione di web di interesse e successivamente classificare le pagine in classi (dove di norma le classi sono due: *spam* oppure *non spam*).

Il fenomeno 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 può avere un ritorno economico legato alla quantità di traffico che viene generata per quel sito. Uno studio del 2005 descritto in [7] stima che la perdita finanziaria mondiale causata dallo spam e di circa 50 miliardi di dollari e nel 2009 (come descritto in [8]) 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 riassunte come segue[16]:

- la qualità delle ricerche è compromessa penalizzando i siti web legittimi;
- 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 contenuto per adulti e attacchi di tipo "fishing". Una prova tangibile si può vedere in [3], dove gli autori hanno eseguito l'algoritmo di *PageRank* su 100 milioni di pagine e hanno notato che 11 sui 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 per il futuro sviluppo di un modulo di spam detection. 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, cioè 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 letteratura 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 con cui si possono creare pagine 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 determinata 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 ($r : Q \times D \rightarrow R$) 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 più 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 ad esempio il grafo composto dai collegamenti ipertestuali tra le pagine web, questo non implica che i metodi esogeni non possono fare uso del contenuto della pagina (per esempio il testo delle ancore). I criteri si dividono ulteriormente in statici (o indipendenti dall'interrogazione) e dinamici (o dipendenti dall'interrogazione). Nel primo caso, il punteggio assegnato a ciascun documento è fisso e indipendente da un'interrogazione q mentre nel secondo il punteggio assegnato a ciascun documento e dipendente da un'interrogazione q.

Tra i metodi endogeni 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

Come già detto i precedenza i metodi di ranking endogeno utilizzano il contenuto di una pagina per assegnarle un punteggio. Possono essere anch'essi statici o dinamici (cioè dipendere o meno da un'interrogazione). 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 dinamico 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 in cui i termini dell'interrogazione 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 occorrono 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 [10]. Più precisamente:

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

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:

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

Con il *tf-idf* gli spammer 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 ottenuto includendo un gran numero di termini distinti in un documento; il secondo, attraverso la ripetizione di determinati termini nel documento.

Un altro metodo di ranking endogeno dinamico è *BM25* [15] che è uno schema di pesatura basato sul *modello probabilistico* ed attualmente è il sistema di pesatura più usato.

1.1.2 Metodi di ranking esogeno

Uno dei metodi esogeni statici è *PageRank* descritto in [14]. PageRank usa le informazioni portate dai link in entrata (*inlink*) per determinare un punteggio globale di importanza di una pagina. Esso assume che esista un legame tra il 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 molte 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 [6]. In dettaglio *PageRank* è basato sulla passeggiata naturale del grafo del web G. Più precisamente, la passeggiata viene perturbata 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 matrice di transizione della catena è quindi rappresentata dalla combinazione lineare:

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

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 esogeno usato per il ranking delle pagine è *HITS (Hyperlink Induced Topic Distillation)* introdotto in [9]. 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 pagina 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* [11].

L'algoritmo lavora su 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 ottenute dalla query. Per questo sottoinsieme di pagine otteniamo un matrice di adiacenza A. I punteggi di *hub* e *authority* per tutte le pagine del sottoinsieme possono essere formalizzate dalla seguente coppia di equazioni:

$$\begin{cases} \vec{a_{t+1}} = A^T \vec{h_t} \\ \vec{h_{t+1}} = A \vec{a_{t+1}} \end{cases} \tag{1.4}$$

Può essere dimostrato che la soluzione ottenuta applicando iterativamente il sistema 1.4 converge rispettivamente al principale autovettore di AA^T e A^TA [11][16].

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 alcune pagine rispetto ad altre [6]. 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; inoltre questo spinge gli utenti ad utilizzare altri motori di ricerca [16].

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

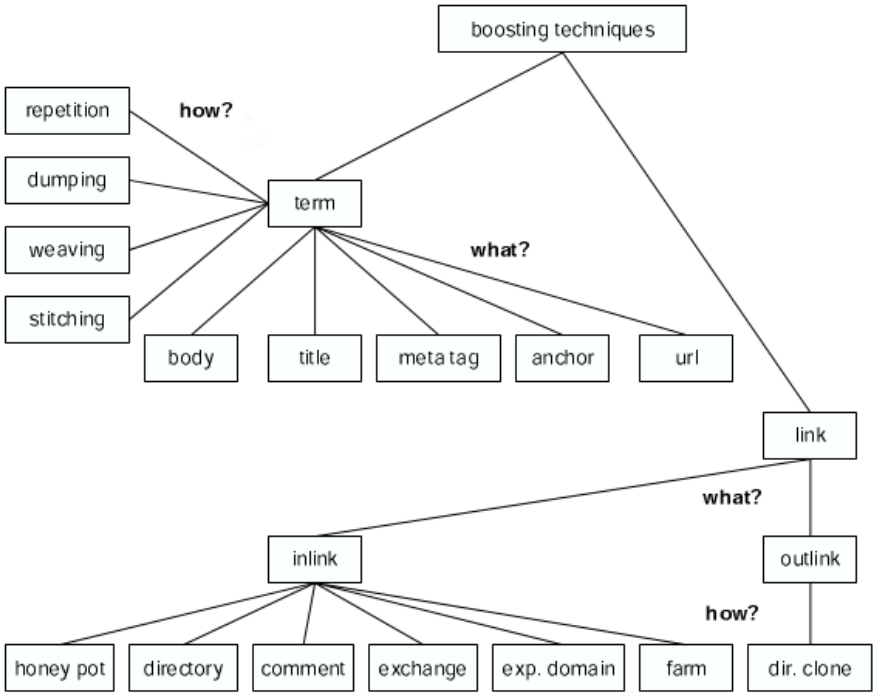

Figura 1.1: Tassonomia delle tecniche boost

popolarità globale della pagina e a volte è inferita dalla struttura dei link [6]. Ci sono due categorie di tecniche associate al web spam [6]:

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

1.2.1 Tecniche di boost

Le tecniche 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 una possibile tassonomia delle tecniche boost [6].

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 viene considerato anche come *campo*, il testo delle ancore (il tag a) associate all'URL che puntano alla pagina p dato che descrive 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 [6]. Le tecniche di spamming possono essere raggruppate in base ai *campi* di testo dove viene fatto spamming. In base a questo distinguiamo [6]:

- *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 questo tipo 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 del 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 cui 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 spammer 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 [6], possiamo avere:

- 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 di senso compiuto per l'elaborazione di contenuti generati velocemente attraverso la concatenazione di frasi da fonti diverse.

1.2.3 Link Spamming

Il *link spamming* è un tipo di spam che fa uso della struttura dei link tra le pagine web per favorire il rank di una pagina target *t*. In [6] si afferma 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 lievemente dallo spammer attraverso l'immissione di un post in un forum o in un blog o portali di questo genere. Le proprietarie sono pagine su cui gli spammer hanno il pieno controllo. Il gruppo di pagine proprietarie è chiamato *spam farm.*

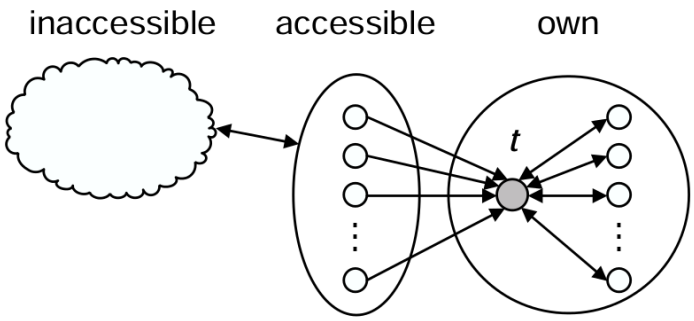

Figura 1.2: Tipi di pagine nel web per uno spammer

Molti motori di ricerca utilizzano due algoritmi per aumentare l'importanza basandosi sulle informazioni dei link: PageRank e HITS; sulla base di questi due tipi di algoritmi vengono definite due categorie principali di *link spamming*: *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 propria 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 [6]:

- *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 link 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.

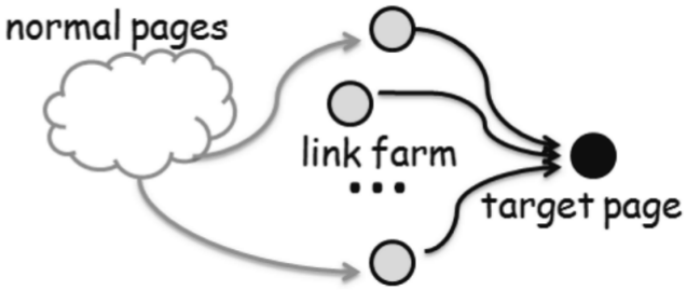

Figura 1.3: Esempio di una spamfarm

- *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. Una tecnica è comprare questi domini e manipolare le pagine in modo tale da fare aumentare il rank di una pagina target.

- *Creare una spam farm*: con l'abbassamento dei costi si possono costruire delle spam farm che hanno come obbiettivo di aumentare la rilevanza di un pagina spam detta *target page*, un esempio è mostrato in fig. 1.3. Molte volte si utilizzano tecniche 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* [16], dove gli spammer prima attaccano un sito con una buona reputabilità e poi usano questo come parte della loro link farm.

1.2.4 Tecniche di hiding

Le tecniche di hiding si possono classificare in: *content hiding*, *cloaking*, *redirection* (fig. 1.4) [6]. 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

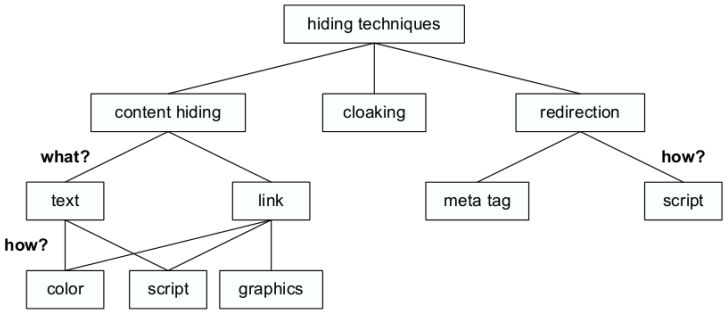

Figura 1.4: Tecniche di hiding

identificare quando la richiesta di una pagina è fatta da un crawler o da un browser: questa tecnica dato un URL, il server spam restituisce un documento HTML diverso a seconda che la richiesta sia fatta da un crawler o da un browser. Quindi vengono distribuiti due contenuti diversi in base al fatto che la richiesta al server spam sia fatta da un crawler o da un browser. La rilevazione di un crawler può essere effettuata in due modi: o si mantiene in memoria una lista di indirizzi di crawler oppure si usa 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 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* [16]. I motori di ricerca utilizzano dati sul flusso dei click per regolare le funzioni di ranking, quindi gli spammer generano clik fraudolenti per manipolare il comportamento di queste fun-

zioni 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 di cui si vuole aumentare il rank. Tale metodo viene eseguito in modo automatico attraverso script che girano su diverse macchine per non fare sospettare il motore di ricerca delle numerose richieste provenienti da un unica macchina [16].

Capitolo 2

Survaey sulle tecniche di spam detection

Il capitolo illustra alcune tecniche di spam detection presenti in letteratura. Le tecniche verranno suddivise sulla base del tipo di segnali che vengono utilizzati, come: il contenuto, il grafo ottenuto dalla fase di crawling ed altri segnali (es. header delle richieste HTTP). Nella prima parte del capitolo verranno illustrate le tecniche basate sul contenuto, nella seconda parte le tecniche che fanno uso di grafi ed infine le tecniche che fanno uso di altri tipi di segnali.

2.1 Tecniche basate sul contenuto

2.1.1 Prime feature per identidicare lo spam

Uno dei primi studi sul web spam è descritto in [4], in questo articolo vengono eseguite delle analisi per confrontare alcune proprietà delle pagine spam e con contenuti in modo tale da estrarre dei valori tramite cui si può stimare se una pagina sia spam. Bisogna precisare che questi valori sono il risultato di studi empirici effettuati su alcuni dataset. Le proprietà vengono classificate come segue:

- Proprietà degli URL. Come descritto in precedenza il *link spam* è una forma di spam dove gli spammer cercano di aumentare il rank derivato dagli algoritmi basati sulla struttura dei link.Quindi uno spammer cerca di creare automatica-

mente tante pagine spam di bassa qualità che puntano a una pagina target p. Alcune analisi sulle proprietà dei link mostrano che gli URL di un HOST sono buone feature per identificare lo spam. In particolare il nome di un HOST con molti caratteri, punti, barre e numeri è un buon indicatore di spam. Perciò un modo semplice per classificare le pagine sarebbe quello di usare un valore di soglia tale per cui superata tale soglia la pagina venga classificata come spam.

- Host name resolution. Alcuni motori di ricerca (es. Google) data una query q, calcolano un rank più alto a un URL u se i termini che compongono il nome dell'host di u combaciano con i termini della query. Gli spammer per sfruttare questo meccanismo popolano gli URL delle pagine spam con termini contenuti in query molto frequentu che sono rilevanti per un certo settore.

- Proprietà del contenuto: le pagine generate automaticamente hanno tutte lo stesso template, ad esempio ci sono numerosi siti di spam che dinamicamente generano pagine che hanno uno stesso numero di parole. Una tecninca per determinare lo spam è quella di clusterizzare le pagine in base alla somiglianza dei template. Visto che le pagine di spam sono molto simili tra loro, identiicando molte pagine con la stessa struttura è probabile che siano di spam.

Oltre a queste proprietà base che servono per identificare lo spam in [13], un lavoro del 2006, vengono descritti una serie di metodi per l'indivuduazione dello spam. Ogni metodo è altamente parallelizabile, può essere eseguito in un tempo proporzionale alla dimensione della pagina e identifica lo spam di ogni pagina scaricata. Inolte questi metodi possono essere combinati con tecniche di machine learning per creare un algoritmo di rilevazione di spam più efficiente. Questo lavoro è un proseguio del lavoro precedentemente descitto [4]. In primis sono stati identificati quali domini e pagine (classidficate sulla base della lingua) contenesserò più spam. Il risultato ha evidenziato che i domini ".biz, .us, e .com" sono quelli con maggiore contenuto di spam mentre per quanto riguarda le pagine contententi più spam sono le francesi, tedesche e inglesi. I risultati sono rappresentati nei due grafici in figura 2.1 e in figura 2.2. Questi risultati si basano sul dataset messo a disposizione degli

autori che è stato ricavato utilizzando MSN Search crawler nell'agosto del 2004, per maggiori informazioni [13].

Una pratica molto comune nel costruire pagine spam è la cosidetta "Keyword stuffing". Durante questo processo il contentuto della pagina aumenta con un numero di parole popolari che sono irrilevanti con il resto della pagina. In molti casi per poter aumentare le probabilità di essere messa in cima al rank di molte query, il contenuto di una pagina spam viene aumentato con tante parole estrane all'argomento della pagina. In figura 2.3 viene plottato la distribuzione delle parole per ogni pagina del data set. Oltre alla distribuzione delle parole viene raffigurata la percentuale di pagine per ogni range di parole che sono considerate spam. Dal grafico viene notato che la prevalenza di spam è più alta nelle pagine con pmolte parole. Perciò c'è una correlazione tra prevalenza di spam e numero di parole. Il conteggio delle parole da solo non è una buona euristica visto che porta un alto tasso falsi positivi.

La tecnica "Keyword stuffing" viene utilizzata anche per la costruzione dei titoli delle pagine spam dal momento che alcuni motori di ricerca assegnano un peso maggiore ai termini della query presenti all'interno del titolo della pagina. Il grafico in fugura 2.4 rappresenta la distribuzione del numero di parole all'interno dei titoli delle pagine. Come per gli altri grafici viene plottata la rispettiva percentuale di pagine spam. Dal grafico si vede che un'eccesso di parole all'interno del titolo è un indicatore (come per il contenuto della pagina) che una pagina è spam. Le parole che vengono utilizzate nel processo di "keyword stuffing" vengono selezionate casualmente o da un ristretto gruppo di query comuni. Per tentare di esaminare il comportamento con cui vengono selezionate e costruite le frasi vengono esaminate le pagine per determinare se sono costituite da un eccesso di parole molto comuni. Prima vengono identificate le n parole più comuni all'interno del corpus poi viene calcolata, per ogni pagina, la frazione delle parole comuni contenute in ogni pagina. Questo processo viene ripetuto per ogni scelta di n. In figura 2.5 viene mostrato il grafico per n=*200*. Il grafico è basato sulla frazione di parole di una pagina che sono contenute nell'insime delle 200 parole più comuni nella porzione delle pagine inglesi del data set utilizzato in [13]. Il grafico ha una caratteristica gaussiana e

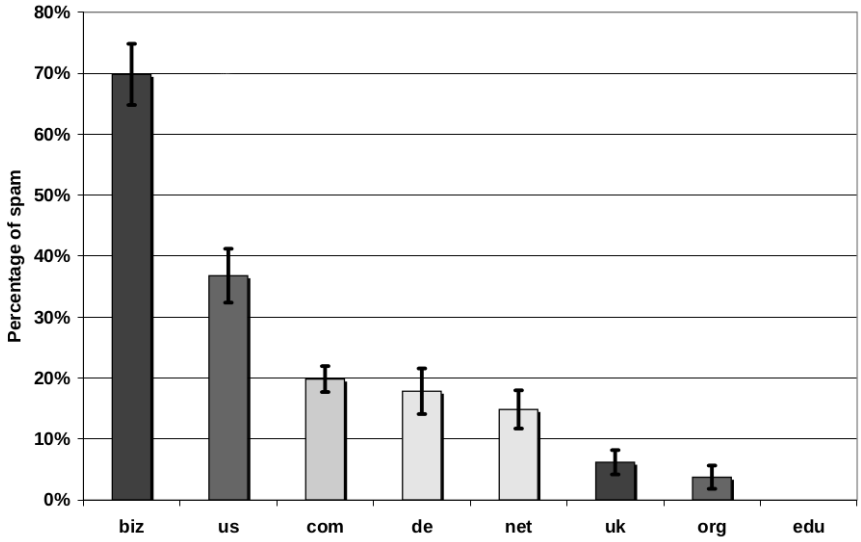

Figura 2.1: Occorrenze dello spam classificate per dominio all'iterno del dataset descritto in [13]

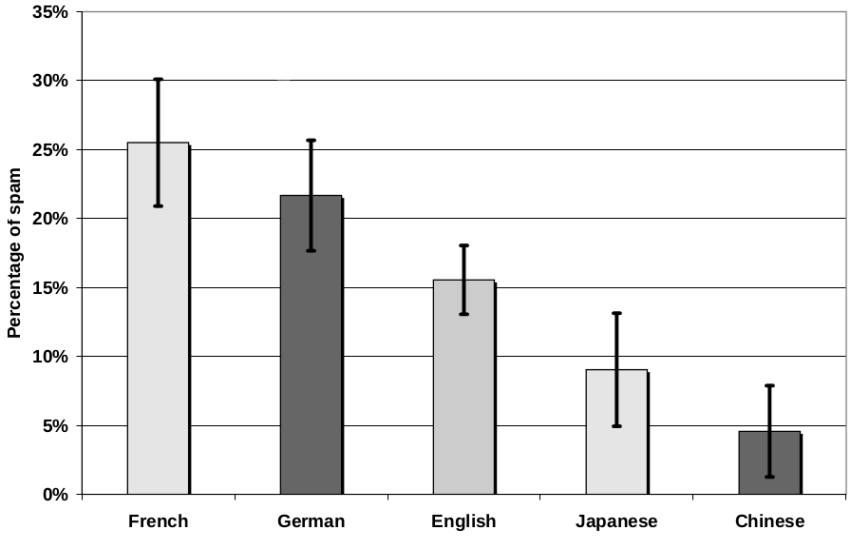

Figura 2.2: Occorrenze dello spam classificate per lingua all'iterno del dataset descritto in [13]

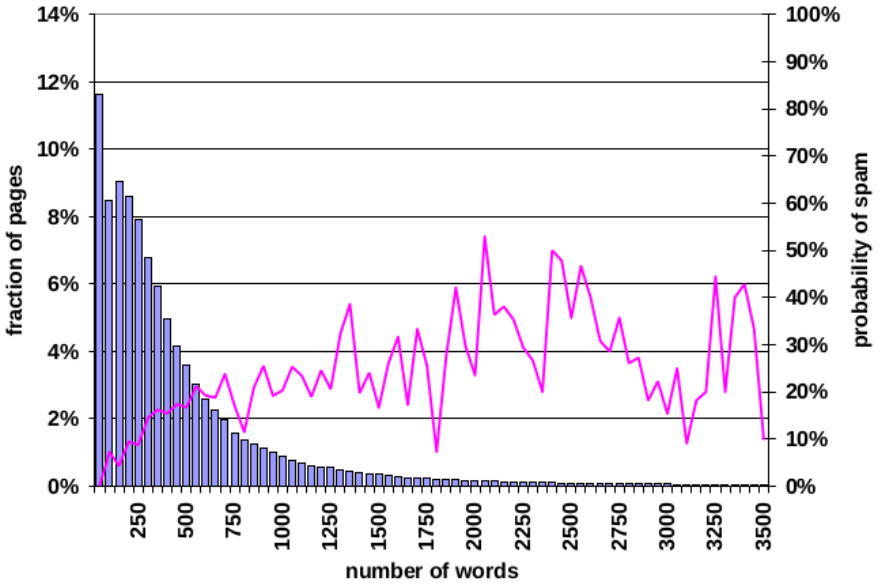

Figura 2.3: Prevalenza di spam sulla base del numero di parole per pagina

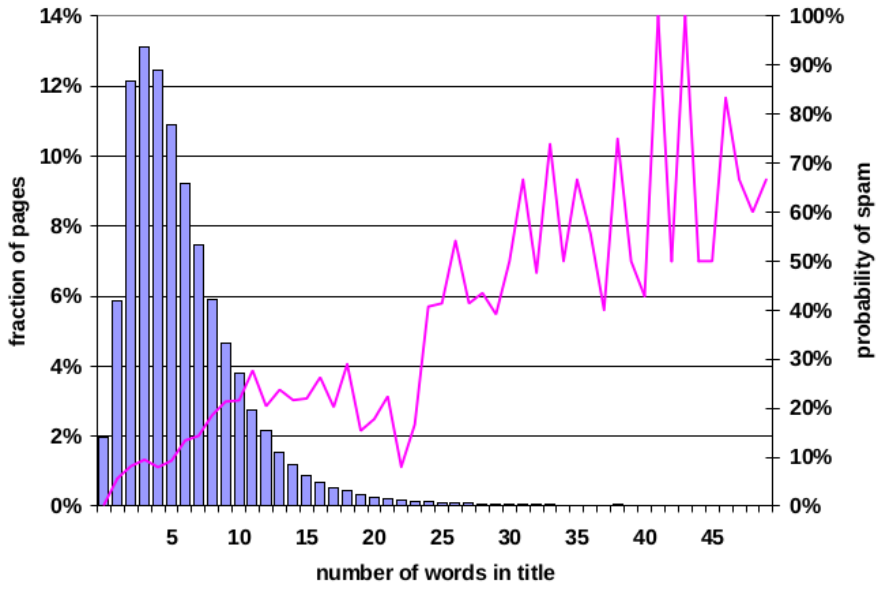

Figura 2.4: Prevalenza di spam sulla base del numero di parole all'interno dei titoli delle pagine

suggerisce che la maggior parte delle pagine di spam sono generate tessendo parole da un dizionare con una scelta casuale.

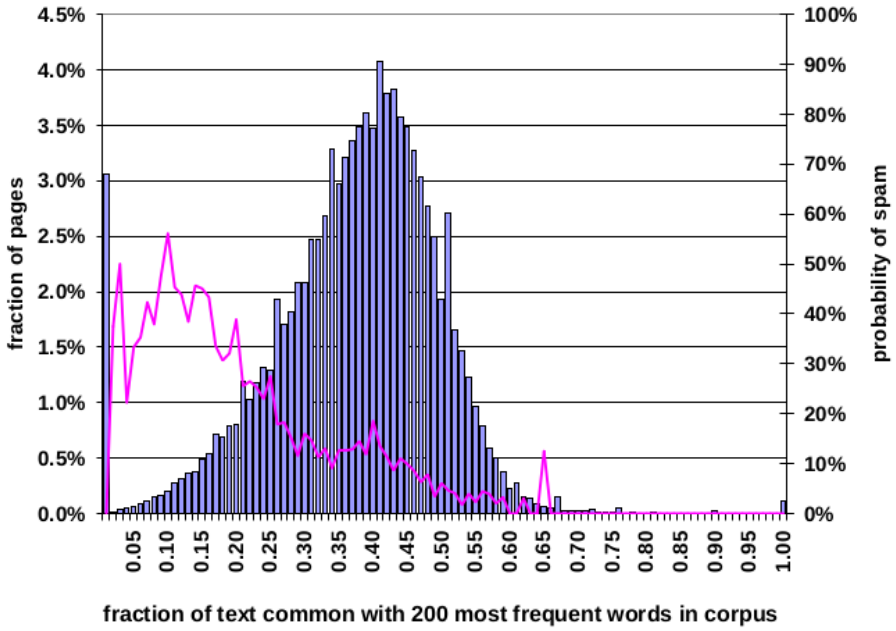

Figura 2.5: Prevalenza di spam sulla base della frazione di parole di una pagina che sono tra le 200 piu frequenti parole nel corpus

Un metodo per identificare pagine spam che sono generate automaticamente è descritto in [5]. Queste pagine spam sono costruite dinamicamente legando insieme frasi grammaticali ben formate. Le pagine create secondo questo metodo vengono generate al volo e cambiano completamente a ogni download. Queste pagine vengono create per essere indicizzate dalla maggior parte dei motori di ricerca. I link di queste pagine puntano ad altre pagine che sembrano essere su un altro host ma vengono risolte da uno stesso IP. Apparte l'uso di differenti host name ma che in realta corrispondono ad unico host per non creare l'illusione di una struttura nepostica questo meccanismo viene usato anche per eludere le politiche di politeness di un web crawler volte al non sovraccarico di qualsiasi host se tali politiche sono basate sul nome dell'host. Infine generando pagine con frasi formate in modo corretto impediscono agli utenti di individuare l'inganno.

Un altro dato trovato dagli autori per determinare se una pagina è spam riguarda la lunghezza media delle parole delle pagine. Dal dataset preso in considerazione in [13] si nota che la distribuzione risultante della lunghezza media delle parole è simile a una gaussiana con moda e mediana corrispondenti a una media di 5.0. La maggior parte delle parole hanno una lunghezza media compresa tra 4.0 e 6.0. Come si nota dal grafico in figura 2.6 le parole con lunghezza media 10 sono certamente spam.

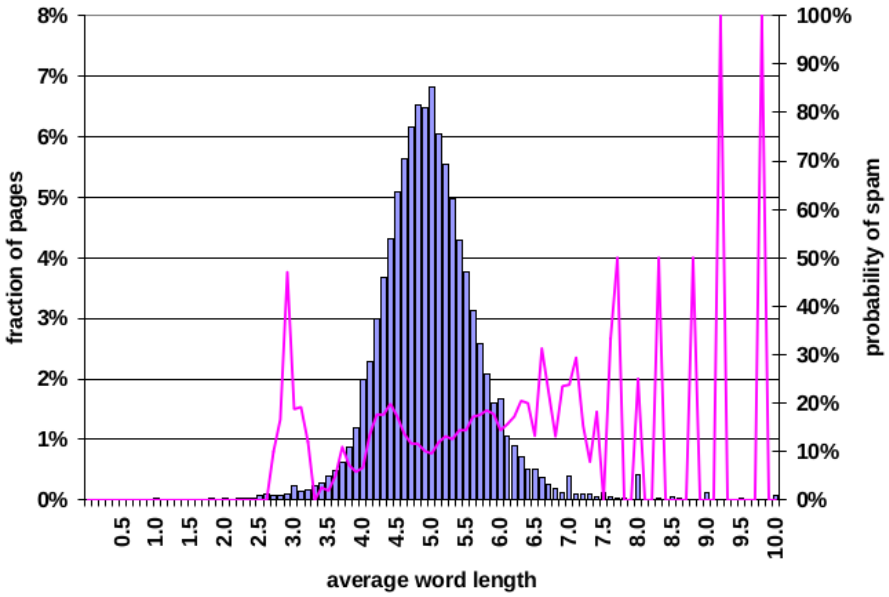

Figura 2.6: Prevalenza di spam sulla base della lunghezza media delle parole per pagina

Un'altra proprietà delle pagine web che consente di stimare se una pagina è spam oppure no è la quantita di testo che è contenuta all'interno delle ancore (il tag "¡a¿" delle pagine web). Infatti una pratica comune dei motori di ricerca è considerare il testo delle ancore dei link in una pagina come annotazioni che descrivono il contenuto della pagina che viene puntata dal link. L'idea prinicipale è che se la pagina a ha un link alla pagina b con testo dell'ancora, ad esempio, "computer" allora potremmo concludere che b parli di computer, anche se questa keyword non compare all'interno della pagina b. Alcuni motori di ricerca tengono conto di questo durante il ranking e potrebbero considerare la pagina b come risultato di una query contenente la keyword "computer". Sfruttando questo meccanismo alcune pagine di spam vengono create solo per contenere del testo all'interno delle ancore per valorizzare il ranking

di altre pagine. Queste pagine di norma sono solo cataloghi di link ad altre pagine. Per capire meglio il fenomeno è stato calcolato la frazione di tutte le parole del testo delle ancore all'interno di una pagina esclusi i markup rispetto al contenuto della pagina. In figura 2.7 viene visualizzato il grafico risultante. Si nota che un'alta frazione di testo delle ancore aumenta la probabilita che la pagina sia spam ma usare questa euristica da sola potrebbe portare un alto numero di falsi positivi.

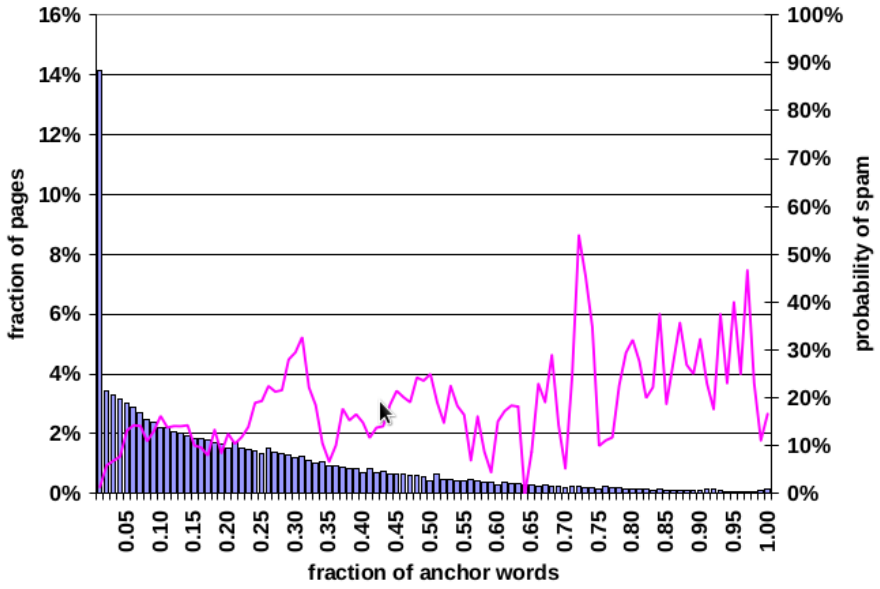

Figura 2.7: Prevalenza di spam sulla base della quantità di testo delle ancore delle pagine

Oltre a queste proprietà anche la quantità di contenuto visibile è utile per capire se una pagina è spam. Infatti calcolando la frazione di contentuto visibile all'interno di una pagina: definita come a lunghezza in termini di byte di tutte le parole non di markup diviso l'interna dimensione della pagina, si nota dal grafico in figura 2.8 della distribuzione delle frazioni di contenuto visibile che le pagine di spam hanno meno markup delle pagine normali. Questo fa intendere che molte pagine spam hanno il solo scopo di dover essere indicizzate dai motori di ricerca e non di essere fruite da un utente.

Come detto in precendenza i motori di ricerca possono dare un maggiore peso a pagine che contengono le keyword contenute nella query più volte all'interno della pagina ad esempio utilizzano come metodo di ranking il "term-frequency". Alcune

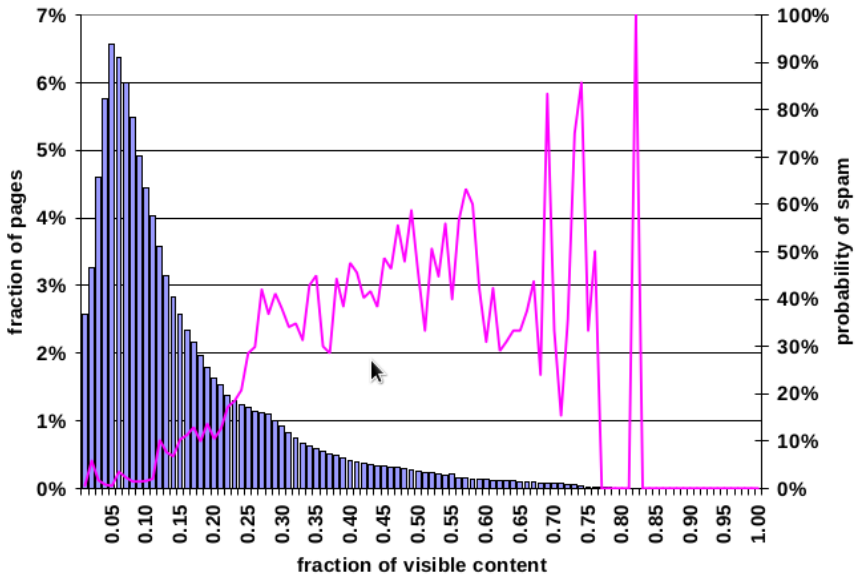

Figura 2.8: Prevalenza di spam sulla base della frazione di contenuto visibile

pagine spam per trarre vantaggio da questo meccanismo replicano i contenuti molte volte per aumentare il rank. Per rilevare la rilevare la ridondanza di contenuti (ottenuta dal processo di replica) e perciò stimare se la pagina sia spam, viene calcolato il rapporto di compressione ovvero la dimensione della pagina non compressa divisa per la dimensione della pagina compressa. In figura 2.9 viene visualizzato la distribuzione del rapporto di compressione e la likelihood che la pagina sia spam. Dal grafico si nota che quanto più il valore di compressione è elevato molto più probabilmente la pagina può essere considerata spam in particolare la probabilità che una pagina sia spam risiede sulla parte destra del grafico, il 70 per cento delle pagine con un rapporto di compressione maggiore di 4.0 sono giudicate spam. Questo è dovuto al fatto che una compressione su una pagina piena di contenuti ridondanti (spam) sarà più efficace che su una pagina caratterizza da conenuti casuali (non spam). E perciò il rapporto tenderà a crescere quanto più la pagina sarà caratterizzata da contenuti ridondanti.

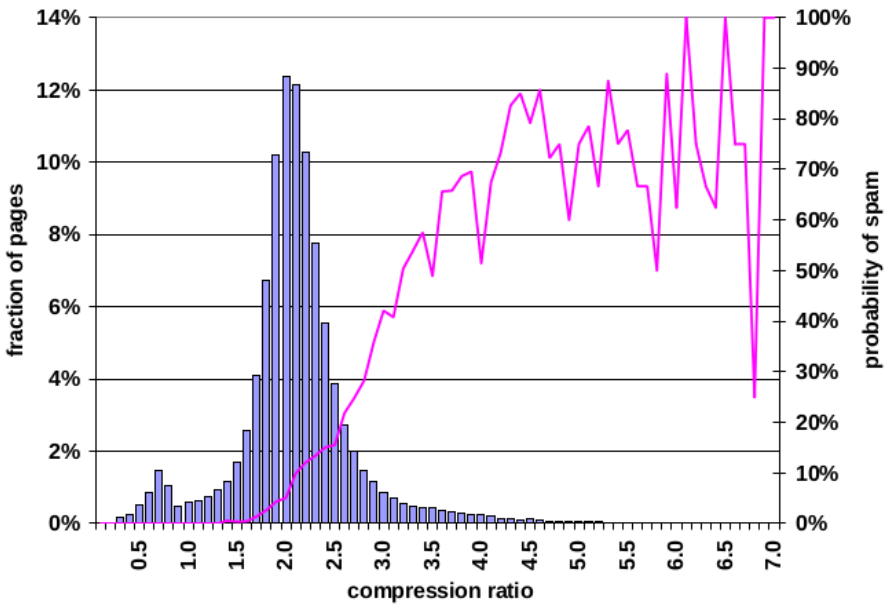

Figura 2.9: Prevalenza di spam sulla base del rapporto di compressione

2.1.2 Utilizzo di un classificatore per combinare le feature

Sempre in [13] le euristiche vengono combinate considerando il problema di rilevamento dello spam come un problema di classificazione. In questo caso viene creato un modello di classificazione il quale data una pagina web userà le caratteristiche della pagina in modo tale di classificarla in una delle due classi: spam o non spam. Costruire un classificatore precede una fase di traning durante la quale i parametri del classificatore sono determinati e una fase di testing durante la quale le perfomance del classificatore sono valutate. Per ogni pagina all'interno del dataset viene calcolato il valore di ogni feature (le varie euristiche) e usiamo questi valori per istruire il classificatore. VUsando un classificatore di tipo "decision-tree", l'algoritmo di classificazione funziona come segue: dato un insieme di dati da traning e un insieme di feature l'algoritmo crea un diagramma di flusso come una struttura ad albero. Ogni nodo dell'albero corrisponde al test da valutare per una particolare feature mentre ogni arco è un valore di uscita del test ed infine le foglie corrispondono alle classi. Per applicare l'albero alle pagine viene controllato il valore della proprietà definita nel nodo radice dell'albero per quella pagina e confrontato con la soglia indicata dai

lati uscenti poi si seguono i lati sulla base dei valori ottenuti finche non si arriva ad assegnare la classe per quella pagina. Un esempio del classificatore è rappresentato in figura 2.10 Per migliorare il classificatore si possono usare tecniche come bagging o boosting. Le tecninche creano un insieme di classificatori che vengono combinati.

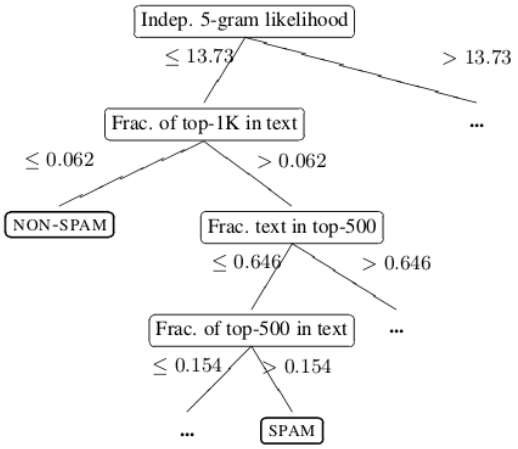

Figura 2.10: Esempio di classificatore

2.1.3 Modelli dei linguaggi per rilevare lo spam

In [12] vengono proposte nuovi tipi di feature per rilevare lo spam e vengono definiti dei modelli dei linguaggi per analizzare le sorgenti estratte da ogni sito in un dataset. Questo metodo di rilevamento delle pagine web spam fa uso di feauture basate su contenuto e struttura dei link in modo combinato per rilevare differenti tipi di web spam. Viene definito un modello per ogni sorgente e calcolato quanto sono differenti i due modelli da ogni altro modello. Le sorgenti di informazioni usate sono:

- testi delle ancore e testo vicino alle ancore della pagine sorgente
- titolo e contenuto della pagina obbiettivo dello spam

Viene utilizzata la Kullback-Leibler Divergence (KLD) per misurare la divergenze tra le distribuzioni di probabilità dei termini di due documenti. La KLD viene applicata a unità di testo della pagina sorgente e di quella linkata. Infatti a Kullback-Leibler

(KL), una misura asimmetrica della divergenza la quale misura quanto male una distribuzione di probabilità M_q riesce a modellare M_d.

$$KLD(T_1||T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)} \tag{2.1}$$

dove in 2.1 $P_{T_1}(t)$ è la probabilità del termine t nella prima unità di testo e $P_{t_2}(t)$ è la probabilità del termine texitt nella seconda unità di testo. In basso sono rappresentati due esempi di KLD applicata tra il testo delle ancore della pagina sorgente e i titoli delle pagine puntate dai link (esempio preso da WEBSPAM-UK2006).

```
KLD(Free Ringtones || Free Ringtones for Your Mobile Phone from
   PremieRingtones.com) = 0.25

KLD(Best UK Reviews || Findabmw.co.uk - BMW Information
  Resource) = 3.89
```

Per determinare se una pagina è spam viene provato a trovare una relazione tra due pagine collegate sulla base del loro valore di divergenza. I valori sono ottenuti calcolando le divergenze con KLD tra una o più sorgenti di informazioni da ogni pagina. In particolare si usano tre tipi di informazione di una pagina sorgente: testo delle ancore, testo intorno alle ancore, termini nell'URL. Sono utilizzate tre tipi di informazione per la pagina che viene linkata dalla pagina sorgente: titolo, contenuto della pagina, meta tag. La combinazione di queste feature possono essere usate per determinare la divergenza tra due pagine. Le feature sono descritte di seguito:

- testo delle ancore - contenuto. Quando una pagina collega un'altra pagina questa ha solo un modo per convincere l'utente di visitare la pagina collegata, mostrare in maniera concisa le informazione relative alla pagina collegata. Perciò una grande divergenza tra questi due pezzi di testo mostra una chiara evidenza di spam. In figura 2.11 è mostrato la divergenza KL tra le sorgenti di informazione, come si vede la curva delle pagine normali è più compatta delle spam. Ma da sola questa feature non è una buona misura.

- testo vicino alle ancore - contenuto. Alcune volte il testo delle ancore è un valore poco descrittivo per ovviare al problema viene usato il testo che circonda

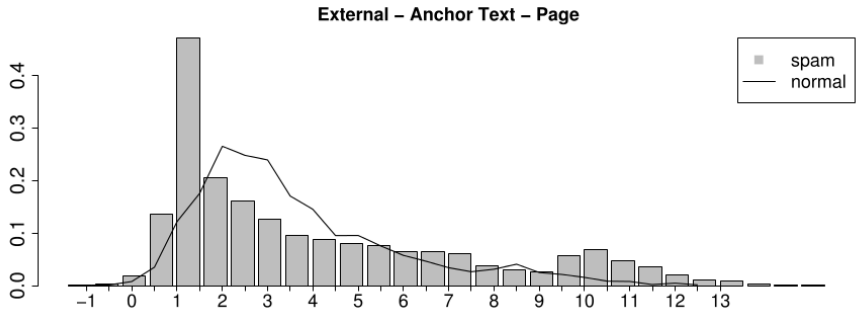

Figura 2.11: Istogramma della divergenza KL tra testo delle ancore e il contenuto della pagina puntata basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

le ancore. Nell'esperimento vengono utilizzate 7 parole per lato. Il risultato è che questa feature riesce meglio a rilevare lo spam. In figura 2.12 viene mostrato che le pagine spam hanno alti valori di divergenza mentre le normali sono concentrate intorno $KL \approx 2.5$.

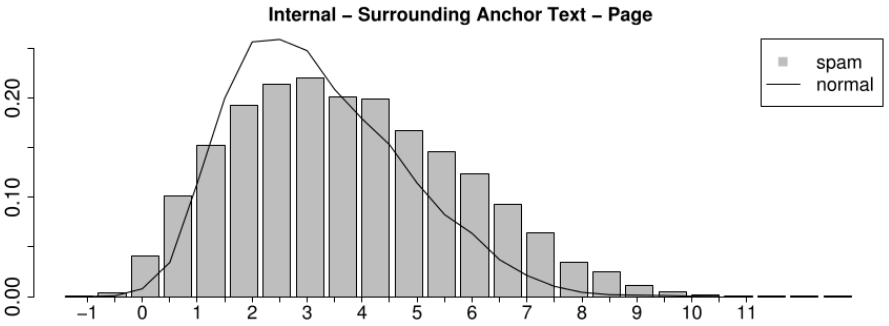

Figura 2.12: Istogramma della divergenza KL tra testo intorno alle ancore e il contenuto della pagina puntata basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

- termini nell'URL - contenuto. Negli URL ci possono essere informazioni che descrivono bene la pagina di destinazione. I motori di ricerca danno molta importanza agli URL per questo un metodo di spam è creare URL come `www.domain.com/viagra-youtube-free-download-poker-online.html` e se visitiamo la pagina questa è magari uno store online. Questa è una delle tecniche per fare spamming. Perciò vengono prelevati i termini più rilevanti da un URL in modo tale da calcolarne la divegenza col contenuto della pagina di destinazione.

La misura finale si vede in figura 2.13 che mostra una grande differenza tra l'istogramma delle pagine normali con quello delle pagine spam.

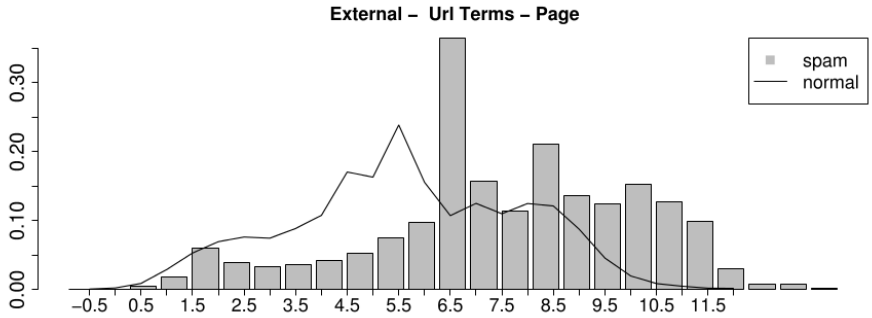

Figura 2.13: Istogramma della divergenza KL tra termini degli URL e il contenuto della pagina puntata basato sul dataset WEBSPAM-UK2007 utilizzato in [12]

- testo delle ancore - titolo. Queste due feature sono molto simili inquanto descrivono la pagina con poche parole. Ma la prima può essere scritta anche da chi non è il proprietario della pagina destinazione. In figura 2.14 notiamo che questa feature da sola non descrimina bene i due tipi di pagine ma è abbastanza efficace utilizzata in congiunzione con altre feature.

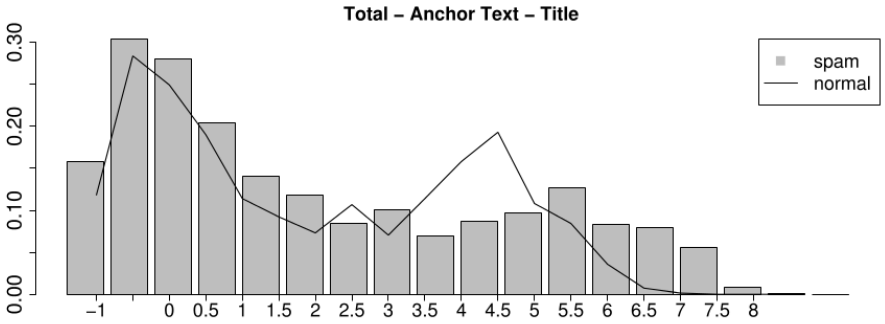

Figura 2.14: Istogramma della divergenza KL tra testo delle ancore e titolo della pagina puntata basato sul dataset WEBSPAM-UK2007 utilizzato in [12]

- testo intorno alle ancore - titolo. Dal grafico in figura 2.15 vediamo che rispetto alla feature precedente questa rileva meglio lo spam infatti molti valori di spam sono concentrati per valori di $KL > 3$.

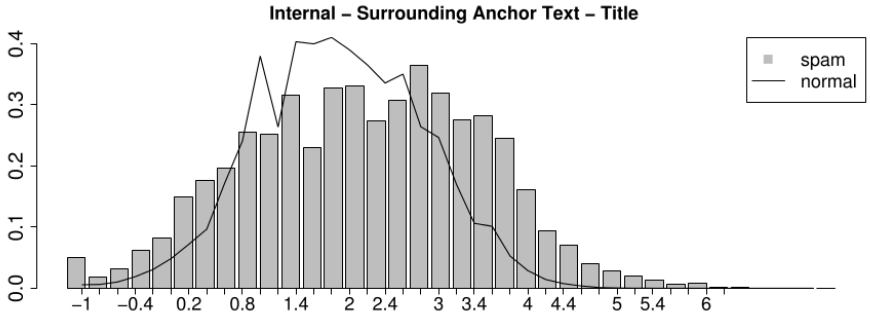

Figura 2.15: Istogramma della divergenza KL tra testo intorno alle ancore e titolo della pagina puntata basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

- termini nell'URL - titolo. Se bene nella precendente feature la sorgente di informazione della pagina sorgente poteva essere generata da un'altra persona in questo caso entrambe le sorgenti sono generate dal proprietario della pagina. Questo dovrebbe fornire coerenza tra le due sorgenti. In figura 2.16 è evidente che la curva delle pagine spam è più compatta della della curva dell'istogramma delle pagine normali.

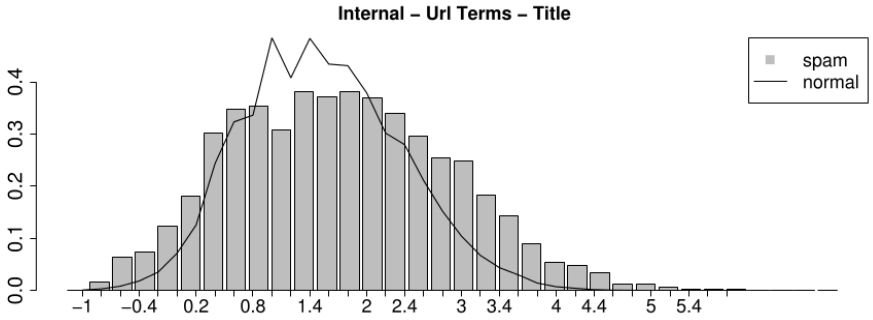

Figura 2.16: Istogramma della divergenza KL tra termini nell'URL e titolo della pagina puntata basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

- titolo - contenuto. I motori di ricerca danno un peso maggiora ai termini dell'URL di una pagina e del titolo. GLi spammer perfezionano i loro processi in modo tale da impostare termini chiave in queste sorgenti che vengono create. In figura 2.17 è rappresentata la divergenza tra le due distribuzioni. Questa misura consente di rilevare i casi di spam quando non vi è nessuna relazione

tra il titolo e il contenuro della pagina dello stesso sito.

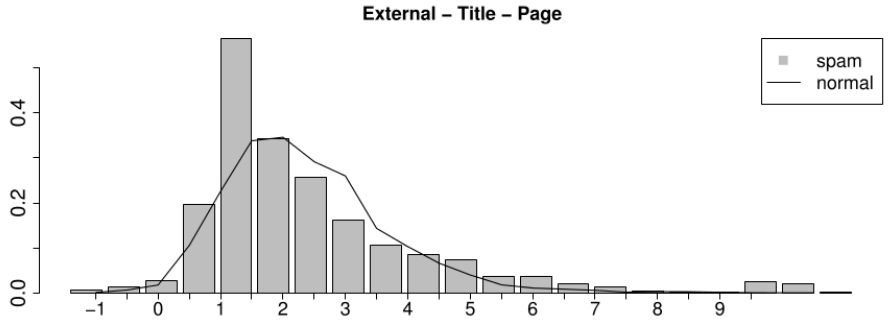

Figura 2.17: Istogramma della divergenza KL tra titolo e contenuto della pagina basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

- metatag. Sono stati usati per calcolare la divergenza con altre sorgenti di informazioni della pagina sorgente come il testo delle ancore e il testo intorno alle ancore e per calcolare la divergenza con sorgenti della pagina destinazione come il contenuto o i temini dell'URL. In figura 2.18 viene visualizzata la divergenza tra il testo delle ancore e i metatag.

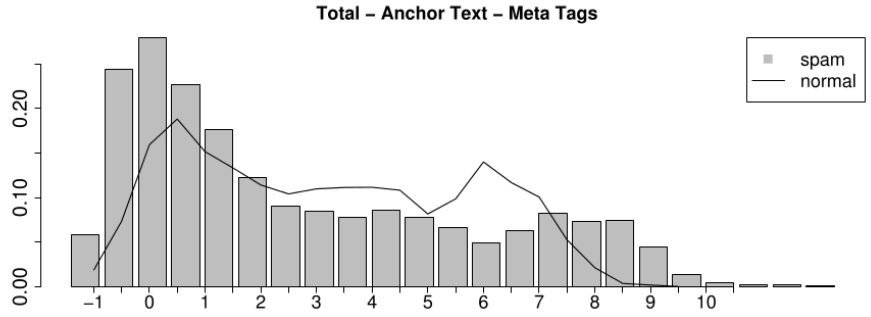

Figura 2.18: Istogramma della divergenza KL tra testo delle ancore e i meta tag della pagina basato sul dataset WEBSPAM-UK2006 utilizzato in [12]

Oltre alle feature descritte: testo delle ancore (A), termini nell'URL (U), testo intorno alle ancore (A), per le pagine sorgenti possono essere definite altre feature combinandole tra di loro: testo delle ancore e termini nell'URL (AU), testo intorno alle ancore e URL (SU) mentre per quanto riguarda le pagine destinazione possono

essere usate le stesse sorgenti cioè contenuto della pagina (P), titolo (T), metatag (M). Per maggiori dettagli [12].

Queste feature possono essere usate per istruire un classificatore e identificare le pagine spam.

2.1.4 Spam detection sulla base degli argomenti di una pagina web

In [2] viene fatta un'analisi sul contenuto spam usando dei modelli specifici e sono state proposte delle misure specifiche di diversità per identificare le pagine spam. Il metodo è basato su statistiche sulla base di argomenti delle pagine. Non utilizza statistiche basate sulle parole della pagina le quali ignorano la semantica tra le parole, ma si basa su statistiche degli argomenti in modo tale da catturare le feature linguistiche nascoste nel testo per capire se la pagina è spam o non spam. Le analisi vengono fatte usando dei modelli degli argomenti (topic model) come la *Latent Dirichlet Allocation (LDA)* che sono modelli statistici dei linguaggi per scoprire argomenti nascosti che compaiono in una collezione di documenti.

Un topic model è un modello statistico che scopre gli argomenti latenti presenti in una collezione di documenti. In generale LDA modella ogni argomento latente come una distribuzione probabilistica su un vocabolario e ogni documento come una distribuzione probabilistica sugli argomenti latenti.

Se analizziamo le feature nascoste di una pagina spam, possiamo notare che i contenuti di queste pagine, generati automaticamente, sono differenti nelle pagine che non sono spam. Da questa intuizione è stato proposto di analizzare il contenuto delle pagine web usanso i topic models e vengono definiti tre tipi di misure.

Una prima misura per determinare le pagine spam utilizzando LDA sfrutta il fatto che le pagine spam sono molto topic-centric, ovvero loro hanno uno specifico insieme di argomenti. In figura 2.19 vengono rappresentate quattro distribuioni degli argomenti di quattro pagine (due spam e due non spam) prese casualmente da un dataset. Si nota che le pagine spam hanno una distribuzione esponenziale, questo supporta la tesi degli autori che le pagine spam sono topic-centric in quanto queste pagine sono sviluppate per avere un alto ranking per un insieme di specifiche query di

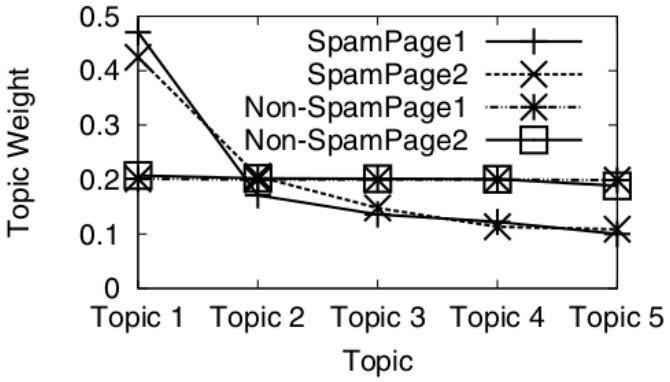

Figura 2.19: Distribuzione degli argomenti pesati per pagine spam e normali

ricerca per simili argomenti. Secondo le pagine normali tendono a coprire un insieme di argomenti aventi tutti lo stesso peso all'interno della pagina. Infatti gli autori sostengono che le pagine non spam come ad esmepio una homepage contengono vari argomenti come ad esempio i contatti, di che cosa si occupa la societa, altre informazioni. Per catturare questa caratteristica sulle distribuzioni dei pesi degli argomenti per le pagine spam e normali, è stato proposto una misura della diversità degli argomenti basata sulla varianza. Data una pagina web d, la sua distribuzione degli argomenti è $T(d) = \{t_1, t_2, ..., t_m\}$, dove ogni argomento $t_i (1 \leq i \leq m)$ è associato con un peso δ_{t_i}. La misura della diversità degli argomenti basata sulla varianza per d, denotata con $TopicVar(d)$, è calcolata come:

$$TopicVar(d) = \frac{\sum_{i=1}^{m} (\delta_{t_i} - u)^2}{m} \tag{2.2}$$

dove $u = \frac{\sum_{i=1}^{m} \delta t_i}{m} = \frac{1}{m}$. Questa misura è stata calcolata per ogni pagina all'interno del dataset. In figura 2.20 viene illustrata la distribuzione. Dalla distribuzione si nota che le pagine spam sono più topic-centric, la varianza dei pesi degli argomenti è relativamente larga rispetto alle pagine non spam. Tra queste la percentuale di spam è piccola. Quando la misura incrementa la percentuale di spam incrementa. Questa misura è un ottimo indicatore per rilevare lo spam.

La seconda misura per identificare le pagine spam utilizzando LDA permette di misurare la relazione semantica tra gli argomenti è la semantica tra le parole. Data una pagina web d, la sua distribuzione degli argomenti è $T(d) = \{t_1, t_2, ..., t_m\}$.

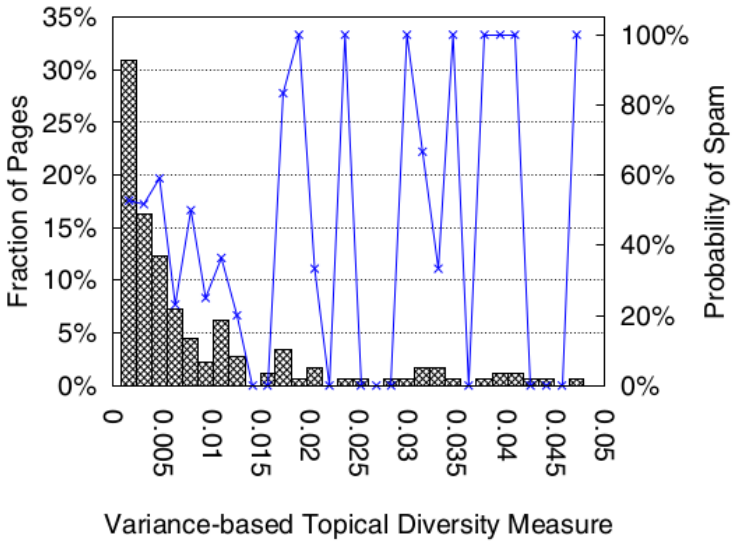

Figura 2.20: Prevalenza di spam relativa alla misura della diversità degli argomenti basata sulla varianza

La probabilità che un parola w appartiene a un argomento $t_i (1 \leq i \leq m)$ è definita come $\phi(w|t_i)$. Ongi argomento t_i è rappresentato come un insieme di parole denotate come $W(t_i)$. Intuitivamente due argomenti $t_i, t_j (1 \leq i, j \leq m)$ sono semanticamente relativi se le parole $W(t_i)$ e $W(t_j)$ sono semanticamente relative. Viene utilizzata una funzione di similarità $Sim(w_i, w_j)$ per ottenere le relazioni semantiche tra le due parole. La funzione di similarità usata è quella di Wordnet. Per misurare la relazione semantica tra due argomenti t_i, t_j, possiamo ottenere le similarità tra ogni coppia di parole dei due argomenti moltiplicate con le loro probabilità rispetto agli argomenti, per ottenere la media delle similiratà.

$$Sim(t_i, t_j) = \frac{\sum_{w_k \in W(t_i), w_l \in W(t_j)} Sim(w_k, w_l) X \phi(w_k|t_i) X \phi(w_l|t_j)}{\frac{|W(t_i)|X|W(t_j)|}{2}} \tag{2.3}$$

Usando un modello degli argomenti otteniamo m argomenti latenti. Ora basandoci sull'equazione 2.3 possiamo derivare una misura della diversità degli argomenti basata sulla semantica per gli m argomenti latenti di una collezione: data una pagina web d, la sua distribuzione degli argomenti è $T(d)$ allora la misura della diversità degli argomenti basata sulla semantica per un documento d è:

$$TopicSim(d) = \frac{\sum_{1 \leq i \leq j \leq m} Sim(t_i, t_j)}{\frac{1}{2} m(m-1)} \tag{2.4}$$

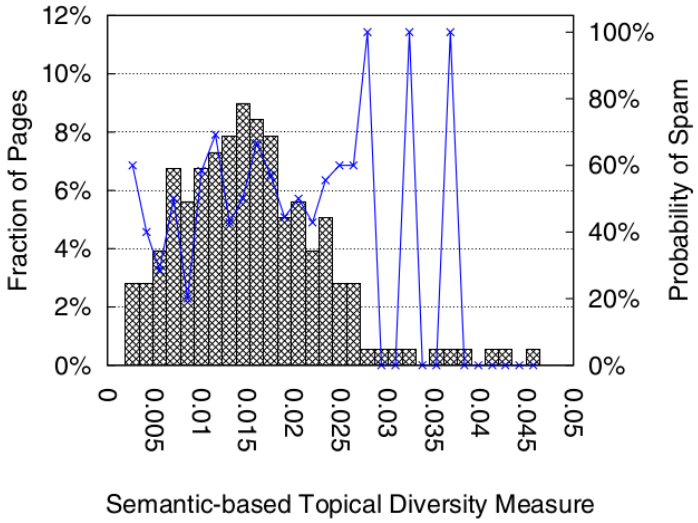

Figura 2.21: Prevalenza di spam relativa alla misura di diversità degli argomenti basata sulla semantica

In figura 2.21 mostra la distribuzione della misura di diversità degli argomenti basata sulla semantica. Dalla distribuzione si nota che quando la misura cresce la percentuale di spam aumenta, ovvero le pagine di spam hanno argomenti che sono molto in relazione semanticamente tra loro. Questo significa che sono icnentrate su pochi argomenti. Questo è un ottimo indicatore di spam.

L'ulitmi misura è basata sulla massima semantica ed è definita come:

$$TopicSimMax(d) = max\{Sim(t_i, t_j)|1 \leq i \leq j \leq m\} \tag{2.5}$$

La distribuzione della misura di diversità basata sulla massima semantica mostra che questa è più alta per le pagine che contengono spam. Quando le pagine spam sono create l'insieme degli argomenti è determinato. La distribuzione è rappresentata in figura 2.22. Possiamo utilizzare questa misura per rilevare lo spam del contentuto.

Per determinare se una pagina è spam oppure non spam, utilizzando questo metodo, vengono utilizzati algoritmi supervisionati di apprendimento per istruire un modello di classifizione di pagine spam usando misure di diversità di argomenti. In LDA noi possiamo impostare parametri come il numero di argomenti e il numero di parole per ogni argomento. Cambiando i parametri cambiano le prestazioni della classificazione.

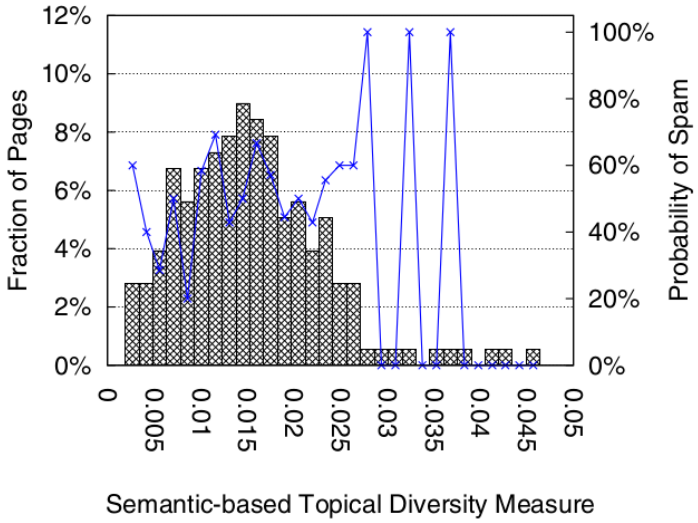

Figura 2.22: Prevalenza di spam relativa alla misura di diversità sulla massima semantica

2.1.5 Altre tecniche

In [1] viene presentato l'algoritmo WITCH (Web Spam Identification Through Content and Hyperlinks), un'algoritmo ibrido che utilizza sia il contenuto della pagina che la struttura dei link per identificare le pagine spam. Come descritto in precedenza nel sotto capitolo 2.1.1 e in particolare in [13], le pagine spam e non spam hanno differenti proprieta le quali possono essere sfruttate per costruire un classificatore. L'algoritmo WITCH oltre alle feature standard descritte in precedenza (vedi il sotto capitolo: 2.1.1)per identificare lo spam analizza la struttura dei collegameti tra le pagine. In particolare viene istruit un classificatore lineare nello spazio delle feature usando la SVM (Support Vector Machine) come funzione obbiettivo. I collegamenti tra le pagine sono utilizzati in modo da regolarizzare il grafo che produce una predizione che varia leggermente tra le pagine dei link. Il risultato è che il metodo SVM associato alla regolarizzazione del grafo è efficiente per il rilevamento di web spam.

Ci sono tanti tipi di tecniche che sono applicate per il rilevamento dello spam nel web in [17] viene proposto un metodo di spam detection denominato "Hidden style similaruìty measure" basato su feature extra testualei appartenenti alle pagina HTML. La tesi che sostengono gli autori è che le pagine spam che sono generato automaticamente non sono facili da rilevare utilizzando metodi classici di classi-

ficazione solamente basati sul conenuto. Il metodo invece di identificare lo spam attraverso l'uso di feature basate sul contenuto utilizza la struttura HTML di una pagina per classificare le pagine simili.

Bibliografia

[1] Jacob Abernethy, Olivier Chapelle, and Carlos Castillo. Web spam identification through content and hyperlinks. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '08, pages 41–44, New York, NY, USA, 2008. ACM.

[2] Cailing Dong and Bin Zhou. Effectively detecting content spam on the web using topical diversity measures. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 266–273, Washington, DC, USA, 2012. IEEE Computer Society.

[3] 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.

[4] Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004*, WebDB '04, pages 1–6, New York, NY, USA, 2004. ACM.

[5] Dennis Fetterly, Mark Manasse, and Marc Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 170–177, New York, NY, USA, 2005. ACM.

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

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

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

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

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

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

[12] Juan Martinez-Romo and Lourdes Araujo. Web spam identification through language model analysis. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '09, pages 21–28, New York, NY, USA, 2009. ACM.

[13] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 83–92, New York, NY, USA, 2006. ACM.

[14] 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.

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

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

[17] Tanguy Urvoy, Thomas Lavergne, and Pascal Filoche. Tracking web spam with hidden style similarity. In *AIRWeb*, pages 25–31, 2006.