



# TECNICHE ONLINE PER L'INDIVIDUAZIONE DI SPAM IN UN WEB CRAWLER

Relatore

*Paolo Boldi*

Correlatore

*Sebastiano Vigna*

Candidato

*Antonio Luca*

# Obbiettivo della tesi

---

Obbiettivo di questa tesi è stata l'analisi delle varie tecniche di spam detection descritte in letteratura al fine di valutarne il comportamento e vagliare la possibilità di utilizzo di tali tecniche online.

# Obbiettivo della tesi

---

Obbiettivo di questa tesi è stata l'analisi delle varie tecniche di spam detection descritte in letteratura al fine di valutarne il comportamento e vagliare la possibilità di utilizzo di tali tecniche online.

## Struttura della tesi

- ▶ Classificazione delle varie tecniche di spam detection sulla base dei segnali utilizzati.
- ▶ Analisi degli algoritmi offline durante la fase di crawling.

# Motivazioni

---

- ▶ I risultati di tale lavoro saranno utilizzati per essere integrati 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 Informatica, di un crawler chiamato BUBiNG, altamente configurabile ma privo al momento di qualunque forma di rilevazione di siti e contenuti malevoli.

# Il fenomeno del web spam

---

## Le cause

- ▶ Col crescere delle dimensioni del web, aumenta la difficoltà di una pagina di comparire tra i primi risultati di un motore di ricerca per una data query.

# Il fenomeno del web spam

---

## Le cause

- Col crescere delle dimensioni del web, aumenta la difficoltà di una pagina di comparire tra i primi risultati di un motore di ricerca per una data query.

## Conseguenze

- Sviluppo di meccanismi di spam per tentare di ingannare gli algoritmi dei motori di ricerca al fine di ottenere un rank maggiore per una data pagina web.

# Problemi

---

I problemi prodotti da web spam sono:

# Problemi

---

I problemi prodotti da web spam sono:

Le cause



# Problemi

---

I problemi prodotti da web spam sono:

## Le cause

- X Qualità delle ricerche compromessa penalizzando i siti web legittimi;

# Problemi

---

I problemi prodotti da web spam sono:

## Le cause

- X Qualità delle ricerche compromessa penalizzando i siti web legittimi;
- X Perdita della fiducia dell'utente nell'utilizzo di un motore di ricerca;

# Problemi

---

I problemi prodotti da web spam sono:

## Le cause

- X Qualità delle ricerche compromessa penalizzando i siti web legittimi;
- X Peridita della fiducia dell'utente nell'utilizzo di un motore di ricerca;
- X I siti spam possono essere usati come mezzo per malware, pubblicazione di contenuto per adulti e attacchi di tipo "fishing".

# Tecniche spam

---

- ▶ Tecniche di boost
  - ▶ Term spamming
  - ▶ Link spamming
  - ▶ Click spamming
- ▶ Tecniche di hiding
  - ▶ Content hiding;
  - ▶ Cloaking;
  - ▶ Redirection.

## Tecniche di boost: term spamming

---

Il term spamming sfrutta alcuni punti di una pagina web per manipolarne il ranking.

## Tecniche di boost: term spamming

---

Il term spamming sfrutta alcuni punti di una pagina web per manipolarne il ranking.

### Tassonomia

- ▶ Body spam;
- ▶ Title spam;
- ▶ Meta tag spam;
- ▶ Anchor text spam;
- ▶ URL spam.

## Tecniche di boost: 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  $t$ .

## Tecniche di boost: 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  $t$ .

### Tassonomia

- ▶ Outgoing link spam;
- ▶ Incoming link spam (Honeypot, Blog, Forum, Wiki, Scambio di link, Domini scaduti, Spam farm).



## Click spamming

---

I motori di ricerca utilizzano dati sul flusso dei click per regolare le funzioni di ranking, quindi gli spammer generano click fraudolenti per manipolare il comportamento di queste funzioni in modo tale da fare avere un rank migliore ai loro siti.

# Tecniche di hiding

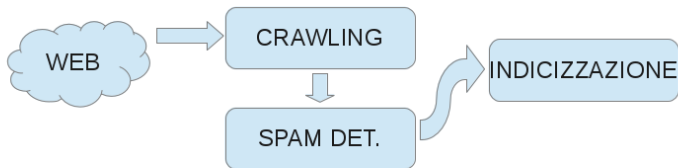
---

## Tassonomia

- ▶ Content hiding;
- ▶ Cloaking;
- ▶ Redirection.

# Spam Detection Offline

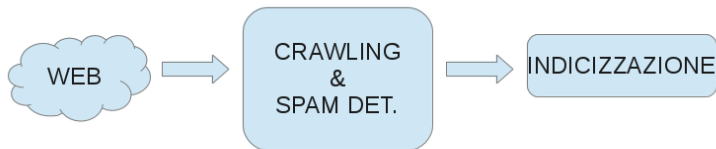
La maggior parte delle tecniche di spam detection operano in modalità offline.



# Spam Detection Online

---

Il rilevamento dello spam viene eseguito durante la fasi di crawling.



# Tecniche di spam detection

---

# Tecniche di spam detection

---

## Classificazione

# Tecniche di spam detection

---

## Classificazione

- ▶ Tecniche basate sul contenuto;

# Tecniche di spam detection

---

## Classificazione

- ▶ Tecniche basate sul contenuto;
- ▶ Tecniche basate sul grafo del web;



# Tecniche di spam detection

---

## Classificazione

- ▶ Tecniche basate sul contenuto;
- ▶ Tecniche basate sul grafo del web;
- ▶ Tecniche basate su segnali eterogenei.

# Tecniche basate sul contenuto - 1

---

Un metodo per identificare lo spam basandosi sul contenuto di una pagina web è quello di analizzare alcune feature delle pagine spam e confrontarle con le medesime feature delle pagine non spam al fine di ottenere dei valori con cui stimare la natura della pagina web.

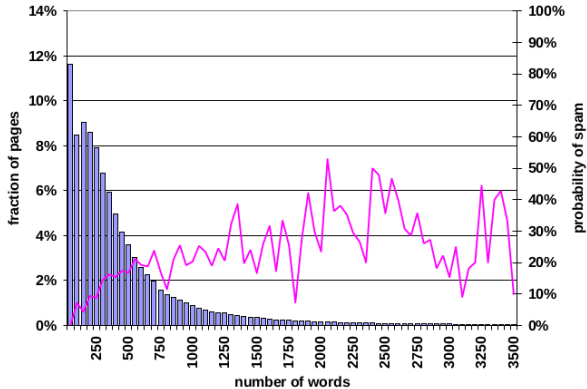
# Tecniche basate sul contenuto - 1

Un metodo per identificare lo spam basandosi sul contenuto di una pagina web è quello di analizzare alcune feature delle pagine spam e confrontarle con le medesime feature delle pagine non spam al fine di ottenere dei valori con cui stimare la natura della pagina web.

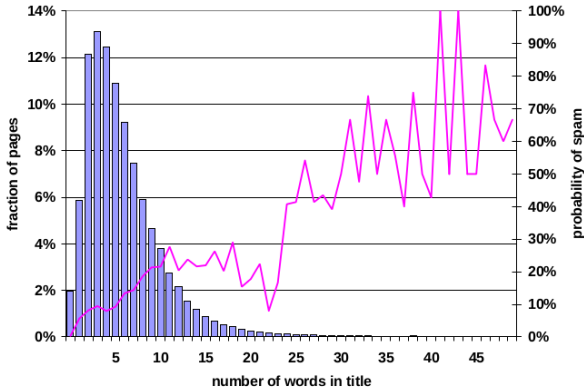
## Esempi di feature

- ▶ Numero di parole all'interno della pagina (Keyword Stuffing);
- ▶ Numero di parole all'interno dei titoli delle pagine;
- ▶ Lunghezza media delle parole all'interno delle pagine;
- ▶ Lunghezza testo all'interno del tag  $\langle a \rangle$ ;
- ▶ Frazione di contenuto visibile;

## Feature: Numero di parole nel body



## Feature: Numero di parole nel titolo



## Tecniche basate sul contenuto - 2

- ▶ Viene utilizzata la *Kullback-Leibler Divergence* (KLD) per misurare la divergenze tra le distribuzioni di probabilità dei termini di pagine web, applicandola a unità di testo della pagina di partenza e di quella linkata
- ▶ La Kullback-Leibler (KL) è una misura asimmetrica della divergenza che misura quanto male una distribuzione di probabilità  $M_q$  riesce a modellare  $M_d$

$$KLD(T_1 || T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)} \quad (1)$$

## KLD: Tipo di sorgenti

---

## KLD: Tipo di sorgenti

---

### Pagina di Partenza

- ▶ testo delle ancore
- ▶ testo intorno alle ancore
- ▶ termini nell'URL



# KLD: Tipo di sorgenti

---

## Pagina di Partenza

- ▶ testo delle ancore
- ▶ testo intorno alle ancore
- ▶ termini nell'URL

## Pagina di Arrivo

- ▶ titolo della pagina
- ▶ contenuto della pagina
- ▶ meta tag

## KLD: Tipo di sorgenti - 2

---

### Combinazione

- ▶ testo delle ancore - contenuto
- ▶ testo vicino alle ancore - contenuto
- ▶ termini nell'URL - contenuto
- ▶ testo delle ancore - titolo
- ▶ testo intorno alle ancore - titolo
- ▶ termini nell'URL - titolo
- ▶ titolo - contenuto
- ▶ metatag