



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

Relatore

*Paolo Boldi*

Correlatore

*Sebastiano Vigna*

Candidato

*Antonio Luca*

# Il fenomeno del web spam

---

## Le cause

- Col crescere delle dimensioni del web, aumenta la difficoltà per i webmaster di far comparire una pagina 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 per i webmaster di far comparire una pagina 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.
- ▶ Sviluppo di tecniche di spam detection.

# 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 online (durante la fase di crawling) degli algoritmi offline.

# 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

---

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

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 del testo all'interno dell'elemento  $\langle a \rangle$ ;
- ▶ Frazione di contenuto visibile;

# Tecniche basate sul grafo

---

## Tecniche base

- ▶ TrustRank
- ▶ Anti-trust Rank

## Tecniche basate sul grafo

### Tecniche base

- ▶ TrustRank
- ▶ Anti-trust Rank

Utilizzano una versione personalizzata di PageRank:

$$\alpha G + (1 - \alpha)\mathbf{1}v^t \quad (1)$$

# TrustRank

---

TrustRank tenta di assegnare un valore di rank maggiore alle pagine non spam rispetto alle pagine spam partendo da un insieme di pagine non spam.

# TrustRank

TrustRank tenta di assegnare un valore di rank maggiore alle pagine non spam rispetto alle pagine spam partendo da un insieme di pagine non spam.

## Assunzione

- ▶ Per determinare le pagine non spam, viene fatta un'assunzione empirica chiamata isolazione approssimata dell'insieme delle pagine buone, la quale afferma che le pagine non spam raramente punteranno a delle pagine spam.
- ▶ Gli sviluppatori di pagine web non spam non hanno interesse nel linkare pagine spam (a meno che tramite l'uso di tecniche come l'honeypot vengano "ingannati").



# TrustRank

---

- ▶ TrustRank quindi è una versione personalizzata di PageRank dove il vettore di preferenza  $v$  non rappresenta una distribuzione uniforme su tutte le pagine del grafo  $G$  ma una distribuzione personalizzata dalle pagine del seedset di partenza.

$$\alpha G + (1 - \alpha)\mathbf{1}v^t$$

# Anti-trust Rank

---

- ▶ Partendo dalla stessa intuizione dell'isolamento approssimato
- ▶ Anti-trust Rank utilizza un seedset set iniziale composto da pagine spam
- ▶ Quindi una pagina spam (conosciuta) è linkata da un'altra pagina spam in quanto una pagina non spam non ne avrebbe il motivo
- ▶ Come per TrustRank, Anti-trust Rank viene calcolato usando Pagerank personalizzato sul grafo trasposto (visto che adesso si prendono in considerazioni i link in entrata)
- ▶ Anti-trust Rank assegna un rank maggiore alle pagine spam

## Tecniche eterogenee: Header HTTP

---

- ▶ Questo metodo utilizza le informazioni racchiuse all'interno degli header HTTP per determinare le pagine spam
- ▶ Può essere usato come supporto ad altri metodi descritti in precedenza e può essere utilizzato in modo dinamico durante la fase di download delle pagine

## Tecniche eterogenee: Header HTTP

- ▶ Questo metodo utilizza le informazioni racchiuse all'interno degli header HTTP per determinare le pagine spam
- ▶ Può essere usato come supporto ad altri metodi descritti in precedenza e può essere utilizzato in modo dinamico durante la fase di download delle pagine

### Funzionamento

- ▶ Dopo aver effettuato la richiesta HTTP al server di una pagina vengono interpretati solo gli header della risposta HTTP
- ▶ Successivamente viene azionato un classificatore per valutare gli header come spam o non spam
- ▶ Se gli header vengono classificati come non spam allora si continua con la lettura del resto della pagina.

# Esperimenti

---

- ▶ Si è scelto, di valutare l'efficacia di due algoritmi link based offline (TrustRank e Anti-trust Rank) durante l'operazione di crawling ovvero in modo online
- ▶ Il razionale di tale analisi è stato la valutazione dell'operabilità di tali algoritmi durante l'esecuzione online e il confronto delle prestazioni rispetto all'utilizzo convenzionale offline

# Esperimenti

---

- ▶ Si è scelto, di valutare l'efficacia di due algoritmi link based offline (TrustRank e Anti-trust Rank) durante l'operazione di crawling ovvero in modo online
- ▶ Il rationale di tale analisi è stato la valutazione dell'operabilità di tali algoritmi durante l'esecuzione online e il confronto delle prestazioni rispetto all'utilizzo convenzionale offline

## Nota

La simulazione della fase di crawling è stata fatta tramite una visita in ampiezza su grafo.

## Esperimento 1 (Confronto online/offline)

---

- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_t$ , tra il vettore  $t$  di TrustRank ricavato sull'intero grafo  $G$  e il vettore  $t_i$  di TrustRank calcolato sul grafo temporaneo  $G_v$  ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza  $v$  con nodo sorgente  $s$ .

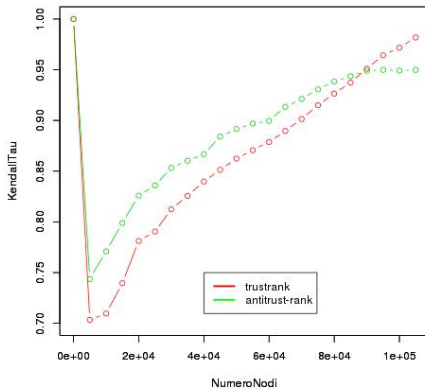
## Esperimento 1 (Confronto online/offline)

---

- ▶ Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_t$ , tra il vettore  $t$  di TrustRank ricavato sull'intero grafo  $G$  e il vettore  $t_i$  di TrustRank calcolato sul grafo temporaneo  $G_v$  ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza  $v$  con nodo sorgente  $s$ .
- ▶ Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_a$ , tra il vettore  $a$  di Anti-trust Rank ricavato sull'intero grafo  $G$  e il vettore  $a_i$  di Anti-trust Rank calcolato sul grafo temporaneo  $G_v$  ricavato ad ogni intervallo di nodi visitati lungo una visita in ampiezza  $v$  con nodo sorgente  $s$ .



# Esperimento 1 (Confronto online/offline): Grafici



## Esperimento 2 (Confronto online/offline parziale)

---

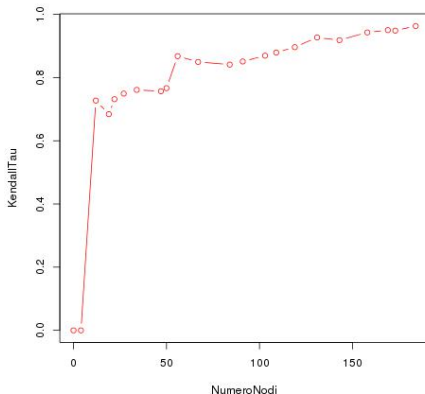
- Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_t$ , tra il vettore  $t$  di TrustRank ricavato sull'intero grafo  $G$  e il vettore  $t_i$  di TrustRank calcolato sul grafo temporaneo  $G_v$ , prendendo in considerazione i soli nodi spam.

## Esperimento 2 (Confronto online/offline parziale)

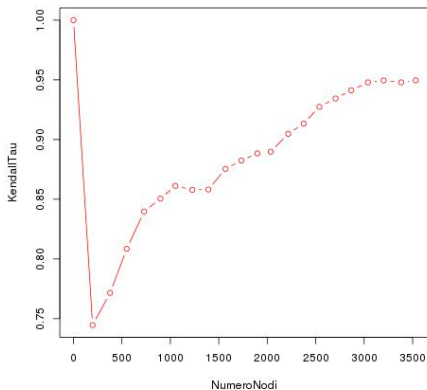
---

- ▶ Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_t$ , tra il vettore  $t$  di TrustRank ricavato sull'intero grafo  $G$  e il vettore  $t_i$  di TrustRank calcolato sul grafo temporaneo  $G_v$ , prendendo in considerazione i soli nodi spam.
- ▶ Calcolo della distanza, attraverso l'utilizzo della Tau di Kendall  $\tau_a$ , tra il vettore  $a$  di Anti-trust Rank ricavato sull'intero grafo  $G$  e il vettore  $a_i$  di Anti-trust Rank calcolato sul grafo temporaneo  $G_v$ , prendendo in considerazione i soli nodi non spam.

## Esperimento 2 (Confronto online/offline parziale): Grafico TrustRank

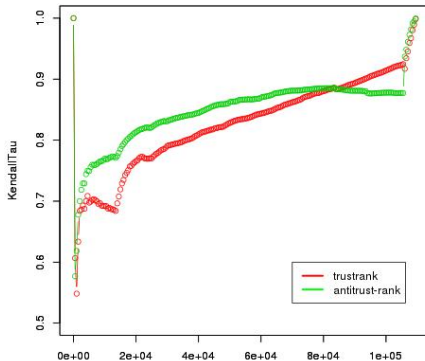


## Esperimento 2 (Confronto online/offline parziale): Grafico Anti-trust Rank



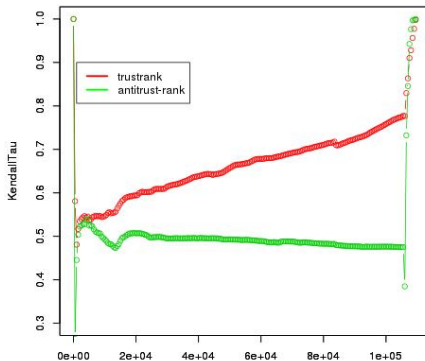
## Esperimento 3 (Avversariale)

Si è simulato una situazione avversariale dove il seedset utilizzato dagli algoritmi sia formato da nodi al limite del grafo  $G$ .



## Esperimento 4 (Avversariale con $\alpha=0.005$ )

Simile all'esperimento 3 ma il fattore di attenuazione  $\alpha$  è impostato a 0.005.



## Esperimento 5 (Separazione delle classi online)

- Dal vettore di TrustRank, calcolato sul grafo temporaneo, viene calcolata la differenza  $\Delta_t$  tra la media  $Mb_t$  dei valori di TrustRank dei nodi non spam e la media  $Ms_t$  dei valori di TrustRank dei nodi spam.

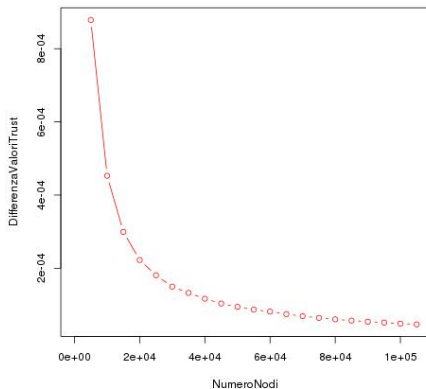
$$\Delta_t = Mb_t - Ms_t \quad (2)$$

- Dal vettore di Anti-trust Rank, calcolato sul grafo temporaneo, viene calcolata la differenza  $\Delta_a$  tra la media  $Ma_t$  dei valori di Anti-trust Rank dei nodi non spam e la media  $Ms_t$  dei valori di Anti-trust Rank dei nodi spam.

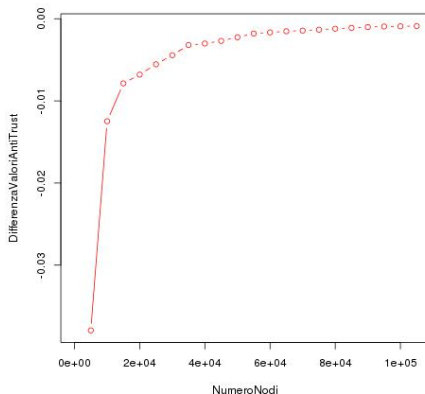
$$\Delta_a = Ma_t - Ms_a \quad (3)$$



## Esperimento 5 (Separazione delle classi online): TrustRank



## Esperimento 5 (Separazione delle classi online): Anti-trust Rank

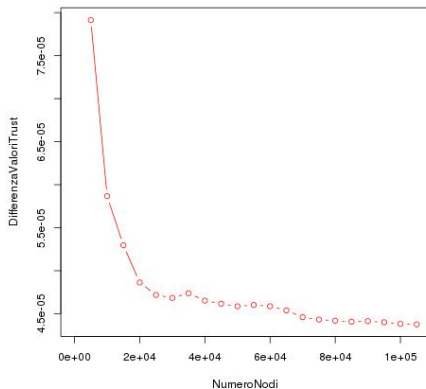


## Esperimento 6 (Separazione delle classi offline)

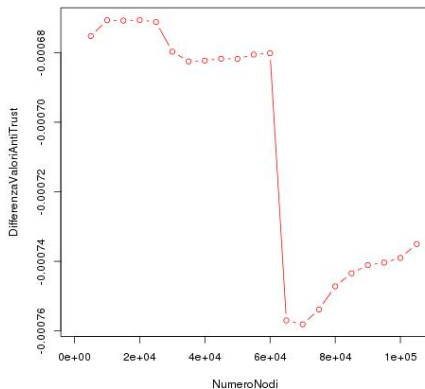
---

- ▶ Dal momento che l'esperimento numero 5 ha prodotto dei risultati che indicano un comportamento diverso da quello atteso, è stato implementato questo test per verificare la correttezza dei risultati ottenuti.
- ▶ Il test è simile all'esperimento 5 ma invece di usare i valori temporanei di TrustRank e Anti-trust Rank per calcolare  $\Delta_t$  e  $\Delta_a$ , durante la visita in ampiezza, si utilizzano i valori ricavati da TrustRank e Anti-trust Rank calcolati sul grafo completo.

## Esperimento 6 (Separazione delle classi offline): TrustRank



## Esperimento 6 (Separazione delle classi offline): Anti-trust Rank



# Conclusioni

---

- ▶ TrustRank e Anti-trust Rank possono essere usati in modo online in quanto approssimano abbastanza bene il loro comportamento offline
- ▶ Confrontando i due algoritmi si è evinto che Anti-trust Rank approssima, per quasi tutta la durata del crawling, meglio il comportamento offline e quindi è più indicato per essere usato durante la fase di crawling.
- ▶ Una volta identificate le pagine spam molte altre pagine potrebbero essere non scaricate.

## Sviluppi futuri

Sviluppo futuro di tale lavoro sarà la progettazione di un modulo di spam detection da inserire all'interno del web crawler BUBiNG.