

Il PageRank è obsoleto? Via libera al TrustRank

Prefazione:

Questo documento si basa sull'originale "Combating web spam with TrustRank" firmato da alcuni ricercatori del dipartimento di Computer Science della Stanford University e pubblicato nel 2004 dallo Stanford Daily Newspaper. Il documento è stato aggiornato il 27 febbraio 2005. Il 16 marzo 2005 la tecnologia TrustRank è stata brevettata da Google Inc.

Introduzione

Il termine web-spam identifica tutta una serie di tecniche per consentire a siti web di varia natura, inclusi siti dal contenuto promiscuo, di essere visualizzati tra i primi risultati nei motori di ricerca utilizzando determinate parole chiave pur non avendo contenuti esattamente inerenti i termini di ricerca utilizzati. (per una definizione più completa si rimanda a documentazione specifica).

Mentre per l'occhio umano non è molto difficile individuare questo tipo di siti o di pagine web, per un sistema automatico risulta più complesso definire il web-spam.

Il solo intervento umano sarebbe troppo oneroso, il sistema ideale si basa su un tipo di classificazione semi-automatica, che è una delle basi della recente tecnica dal TrustRank.

La tecnica dello web-spam può rientrare in due grandi famiglie:

- Link e parole chiave nascoste con lo scopo di ingannare i motori di ricerca. Questo perché i motori di ricerca indicizzano comunque le parole chiave extra e assegnano un valore al sito che le contiene.
- Creazione di un elevato numero di pagine che puntano ad una singola pagina designata in precedenza. In questa maniera i motori di ricerca ritengono rilevante la pagina linkata.

Queste tecniche potrebbero sfuggire ad un sistema automatico, mentre difficilmente potrebbero sfuggire all'occhio umano. Ma l'intervento umano su ogni pagina indicizzata da un motore di ricerca è praticamente impossibile, oltre che costosa. Resta comunque il fatto che senza una gestione certa della qualità dei siti indicizzati un motore di ricerca difficilmente può offrire risultati utili.

Lo scopo del TrustRank è quello di aiutare nella classificazione di una pagina o di un sito web.

Il processo per definire se una pagina è considerabile come web-spam, o comunque per la sua corretta classificazione qualitativa, può essere sintetizzato in tre passi:

- L'algoritmo prima seleziona un gruppo di pagine delle quali non è chiaro lo "spam status" (definite seed).
- Un esperto umano esamina le pagine e comunica all'algoritmo quali posso essere definite spam (bad pages) e quali no (good pages).
- L'algoritmo identifica le altre pagine sulla base della classificazione umana iniziale.

Assegnazione della fiducia - TrustRank

Definizione preliminare dell'algoritmo PageRank

Il PageRank è un diffuso algoritmo, sul quale si è basata per anni (e si basa tuttora) l'indicizzazione di Google, che assegna un punteggio ad ogni pagina, basandosi sul numero di link che puntano ad essa. Il fondamento del PageRank è che una pagina deve essere ovviamente importante se molte altre pagine puntano con un link ad essa. Il PageRank può essere considerato una versione più raffinata e complessa della "Link Popularity" (LP). Il PageRank di un sito aumenta in relazione alla qualità delle pagine web che linkano il sito (per pagine web di qualità si intendono quelle che a loro volta hanno un alto PageRank).

(per una definizione più completa si rimanda a documentazione specifica)

Basi del trustRank

La determinazione certa di una pagina può provenire esclusivamente dalla soggettiva valutazione umana, l'esperto in questione viene definito Oracolo.

Da questo è possibile comunque generare un semplice algoritmo che assegna un valore binario pari a 0 se la pagina contiene spam (bad page), oppure pari a 1 se la pagina è da considerarsi di buona qualità o senza spam (good page).

L'intervento dell'Oracolo come detto porta via molto tempo ed è anche costoso in altri termini, non è possibile un suo intervento costante.

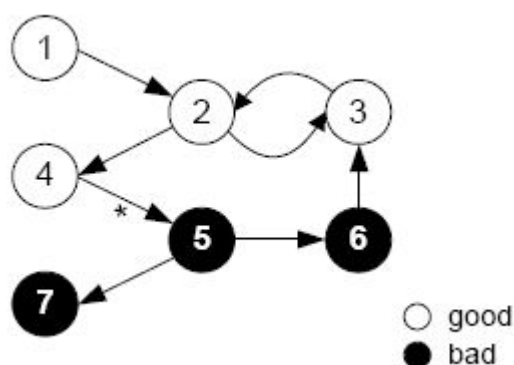
Per diminuire gli interventi dell'Oracolo è possibile adottare una tecnica basata sull'esperienza che ogni singolo navigatore del web può avere: una pagina buona difficilmente punta a una pagina cattiva. Questo perché l'autore di una pagina con una qualità elevata difficilmente ha interesse che venga collegata con pagine con una qualità minore o nulla. Per questo motivo le tecniche di spam hanno cominciato ad adottare dei sistemi per collegare pagine che a prima vista sembrano buone, a pagine contenenti web-spam, con varie tecniche:

- Introduzione nei siti che offrono la possibilità ai visitatori di inserire commenti, di un numero elevato di link verso le pagine cattive (questo problema viene risolto con la moderazione dei commenti o con la futura introduzione del tag link rel="nofollow" – NDR). Tecnica valida anche per forum o web chat.
- Creazione di pagine che offrono contenuti validi, ma che hanno un elevato numero di link nascosti verso siti equivoci. Questa tecnica

viene definita Honey Pot. Per avvalorare questa tecnica chi crea queste pagine inserisce anche numerosi link verso pagine buone.

- Creazione di directory basate sui risultati dei principali motori di ricerca, che utilizzano dei motori chiamati spam-engine, per piazzarsi nei primi posti delle ricerche.

In questi casi di esempio, adottando un algoritmo che funzioni da Oracolo, potrebbero crearsi situazioni equivocate. Se abbiamo un sito di 100 pagine, con 70 pagine buone (quindi con valore binario pari a 1) e 30 pagine cattive (con valore binario pari a 0), si avrà un indice di fiducia intermedio. In questo caso solo l'occhio umano riesce a stabilire una verità certa.



Inverse PageRank

Una tecnica valida per stabilire la qualità di un sito potrebbe venire dall'inversione della tecnica del PageRank. Ovvero dare una preferenza positiva a quella pagine dalla quale è possibile raggiungere molte altre pagine, basandosi quindi sul numero dei link uscenti (outbound links).

High PageRank

Per realizzare questa tecnica è stata effettuata una query su un motore di ricerca che utilizza PageRank come tecnica. Ogni risultato avrà vicino risultati con PageRank simile. Questo perché il PageRank si propaga tra i link. Pagine con un elevato PageRank saranno collegate a pagine con un PageRank simile.

Sperimentazione del TrustRank

Per valutare l'algoritmo TrustRank, il team ha usato l'indicizzazione del motore di ricerca Altavista. Per ridurre la complessità del sistema i test si sono svolti prendendo in considerazione la totalità di ogni sito web, e non ogni singola pagina che lo compone.

Quindi diversi miliardi di pagine sono stati raggruppati in 31.003.946 siti, usando uno degli algoritmi proprietari di Altavista.

Osservazione: un terzo dei siti selezionati non avevano alcun tipo di classificazione, questo perché l'algoritmo PageRank che propaga la fiducia si basa sulla presenza di link tra i siti. Ma questi siti comunque avevano un indicizzazione bassa, non è stato difficile quindi separare manualmente siti buoni e siti cattivi.

Questo tipo di controllo ha portato via delle settimane, ulteriore conferma dell'impossibilità dell'intervento completamente manuale.

Comparazione Inverse Page Rank / High Page Rank

Come prima azione è stata adottata la tecnica definita Inverse Page Rank per selezionare dei siti, dei quali sono stati esaminati i primi 25.000 risultati. Da questi è stato necessario eliminare una serie di siti per due motivi principali:

- alta presenza di siti-cloni della directory DMOZ, a scopo di web-spam.
- Alto numero di siti non indicizzati in nessuna directory principale e quindi reputati poco attendibili.

Dopo questo passaggio i siti attendibili si sono ridotti a 7.900. Di questi sono stati esaminati manualmente i primi 1.250 per selezionarne 178 da usare come gruppo (seed) di siti buoni.

Il numero relativamente ridotto della sezione ha consentito di adottare dei criteri molto rigidi di determinazione tra web-spam e pagine buone.

Nonostante questo è stato adottato un secondo filtro per selezionare i siti con una sicura e certa autorità (come siti istituzionali o di grandi compagnie). Questo secondo filtro si è reso necessario per garantire una buona longevità del gruppo della selezione (seed).

Valutazione dell'operatore del trustRank

Al fine di valutare la funzionalità del sistema TrustRank bisogna sottolineare nuovamente come il sistema Page Rank non garantisca in alcun modo la qualità dei siti indicizzati. Invece il sistema TrustRank effettua una netta separazione tra siti buoni e siti definiti come web-spam. Questi ultimi difficilmente possono avere un indice TrustRank molto elevato.

Dagli esempi è possibile capire l'efficacia dell'algoritmo TrustRank:

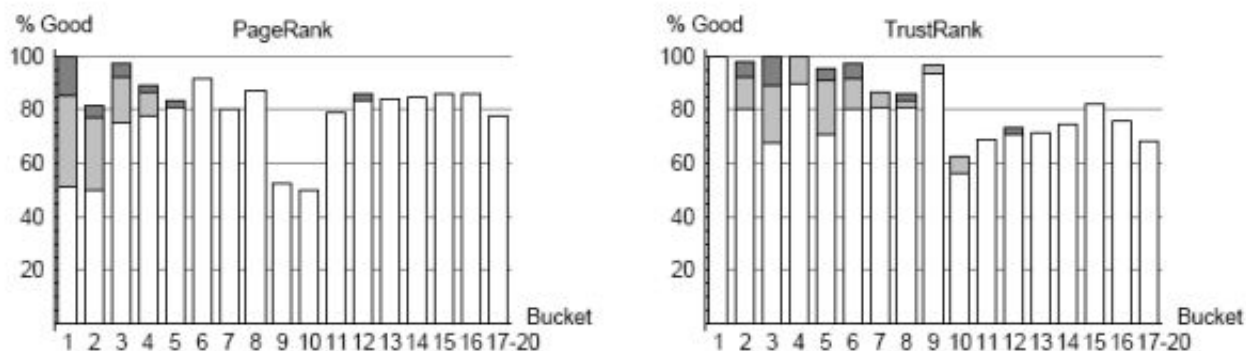
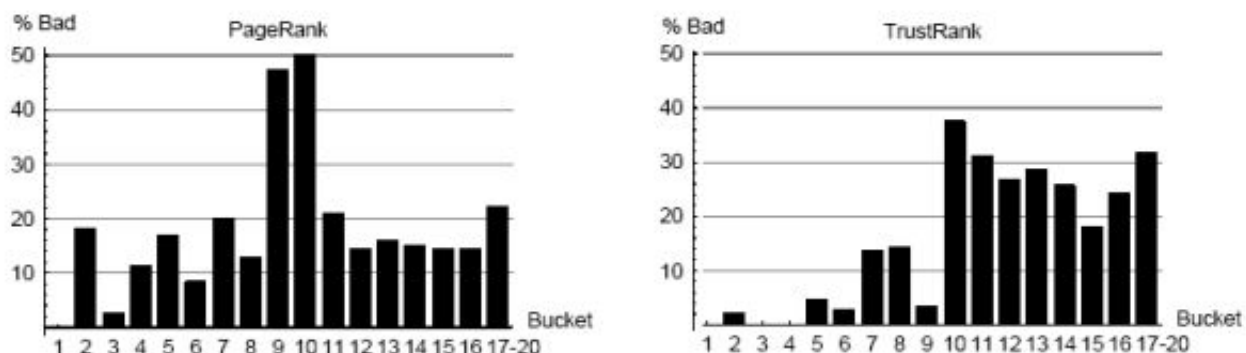


Figure 9: Good sites in PageRank and TrustRank buckets.



TrustRank quindi rimuove gran parte del web-spam dai risultati preminenti per una certa ricerca. Quindi garantisce come i siti più rilevanti siano buoni, ovvero di qualità elevata e senza web-spam.

E' importante specificare come guardando in basso, si evince che i siti meno rilevanti difficilmente siano distinguibili da quelli contenenti spam.

Conclusione

Con la crescita smisurata delle fonti e dei siti sul web i motori di ricerca giocano un ruolo fondamentale per la ricerca e soprattutto l'effettivo successo nella ricerca di informazioni.

Il web-spam demolisce questa capacità di successo nella ricerca di informazioni utili. I motori di ricerca quindi devono necessariamente evolvere.

Il sistema TrustRank, anche in combinazione al PageRank o altri algoritmi potrebbe contribuire a questa evoluzione.

Chiarimenti

Questo documento non garantisce nessuna attinenza con la ricerca "Combating web spam with TrustRank". L'autore non è responsabile per eventuali omissioni, rimaneggiamenti o errori. L'autore ha ommesso una lunga serie di formule e matrici matematiche presenti nel documento originario.