UNIVERSITÀ DEGLI STUDI DI SALERNO

WEB SPAMMING

(TECNICHE DI SPAM NEI MOTORI DI RICERCA)

SICUREZZA SU RETI 2

SCALA SANTOLO 0521000268

ANNO ACCADEMICO 2007/2008

Prof. Alfredo De Santis

Sommario Pag.

Introduzione

Il lavoro di questa tesina affronta il problema del **web spamming**. In relazione ai motori di ricerca, si usa il termine *spam* per indicare siti o pagine "preparati" con l'apposito scopo di arrivare alle prime posizioni dei risultati, senza offrire un reale contenuto utile al navigatore.
Per arrivare a discutere del problema del web spamming, quindi, dobbiamo prima capire e focalizzare l'attenzione sui motori di ricerca e sul loro funzionamento.
Questa tesina prende spunto da una serie di articoli e saggi riportati in bibliografia e scritti da esperti e professionisti del settore.
Questa tesina è sviluppata in tre capitoli. Si parte dai concetti di "motore di ricerca" e "web spamming", per arrivare a trattare, in maniera approfondita, alcune delle *tecniche* e gli *algoritmi* che vengono utilizzati dai motori di ricerca, ovvero l'algoritmo di Kleinberg (I capitolo) ed il PageRank (II capitolo). Infine vedremo come queste tecniche sono state sfruttate per alterare i risultati di ricerca, ovvero vedremo le varie le **tecniche di web spamming** (III capitolo).

Motori di ricerca e Web Information Retrieval

I motori di ricerca oggi rappresentano la frontiera della ricerca nel campo informatico. Parte del successo del web come lo conosciamo adesso è dovuto alla idea dei motori di ricerca.
Un **motore di ricerca** è un sistema automatico che analizza un insieme di dati spesso da lui stesso raccolti e restituisce un indice dei contenuti disponibili classificandoli in base a formule matematiche che ne indichino il grado di rilevanza data una determinata chiave di ricerca.
L'idea di indicizzare tutti i documenti esistenti per accedere alle informazioni di interesse è stata un'esigenza necessaria per far fronte alla grosse mole di dati presente sul web. All'inizio si cercava di far fronte alla crescente espansione e complessità del web attraverso l'uso di strumenti classici come i portali telematici**.** Si diffuse, così, l'idea che la navigazione nel web dovesse partire da qualche punto, da una radice ben precisa (*il concetto di directory*) e poi spostarsi tra le varie sezioni. I motori di ricerca si sono sviluppati – nella seconda metà degli anni Novanta – sulla base delle tecniche già esistenti di *information retrieval*, favorendo lo sviluppo di nuove tecnologie e metodologie che hanno portato alla realizzazione del cosiddetto *web information retrieval* e al successo di "Google" così come lo conosciamo oggi.

Le tecniche di *Information Retrieval* – sviluppatesi tra gli anni Sessanta e Ottanta – si riferiscono alla capacità dei calcolatori di processare testi, per esempio saper dire quante occorrenze di un parola o di una frase sono presenti all'interno di un testo (**string processing**); alla capacità di cercare documenti che si somigliano tra di loro; o alla capacità di sapere attribuire due documenti ad uno stesso autore. Queste caratteristiche sono molto utili per chi deve gestire biblioteche e dati testuali. Il Web per dimensione, complessità e caratteristiche di multimedialità è completamente diverso dall'insieme dei documenti omogenei soggetto alle classiche tecniche dei motori di *Information Retrieval*.

Il *Web information retrieval* si basa su due algoritmi principali:

- **HITS (Hypertext Induced Topic Search) - Ricerca Indotta dall'Ipertesto;**
- **PageRank (Il Rango di Page);**

Nei prossimi capitoli descriveremo, molto brevemente, il funzionamento ed i principi su cui si basano questi due algoritmi e vedremo, infine, come queste tecniche sono state *sfruttate* per falsificare il risultato delle query su pagine web.

Infatti, il problema maggiore è che il Web è *libero*. Con le tecniche di *Information Retrieval* non c'era modo di trarre in inganno il motore di ricerca associato; mentre nel Web moderno esiste il problema del **ranking** di una pagina, cioè la possibilità di apparire quanto più in alto possibile in un risultato di ricerca e spesso per far questo si usano pratiche scorrette nella strutturazione dei documenti (web *spamming*).

Web Spam

Il termine *web spam* si riferisce a tutte quelle azioni che hanno come scopo quello di elevare pagine web, riguardo alle interrogazioni dei motori di ricerca, allo scopo di attribuirgli un evidenza maggiore rispetto a quella che dovrebbero avere.[1]

Ci sono varie linee di pensiero su quale sia l'origine del termine *spam*. Una di queste linee sostiene che il termine *spam* trarrebbe origine da un divertente sketch del *Monty Python's Flying Circus* ambientato in un locale nel quale ogni pietanza proposta dalla cameriera era a base di Spam (*un tipo di carne in scatola*). Man mano che lo sketch avanza, l'insistenza della cameriera nel proporre piatti con "*spam*" ("*uova e spam, uova pancetta e spam, salsicce e*

[1] Definizione di ***web spam*** tratta dall'articolo "*Web Spam Taxonomy"* di Zoltàn Giongyi e Hector Garcia-Molina e riportato nella bibliografia.

spam" e così via) si contrappone alla riluttanza dell'avventore per lo "*spam*", il tutto in un crescendo di un coro inneggiante allo "spam" da parte di alcuni Vichinghi seduti nel locale.
Vista l'importanza che riveste la ricerca di informazioni nel web è ovvio che per una società, aumentare la propria esposizione può garantire un alto ritorno in termini economici. Visto che i motori di ricerca, in alcuni casi, sono le porte di ingresso del web, molti soggetti utilizzano tecniche specifiche per fuorviare i risultati dei motori.
In questa maniera aumentano la propria rilevanza, ma abbassano la qualità delle ricerche effettuate, inoltre aumentano il costo di ogni ricerca, da parte dei motori, per via delle pagine poco usabili che appesantiscono le query.
L'obiettivo di un motore di ricerca è quello di offrire alta qualità nei risultati identificando correttamente quali pagine sono rilevanti per una determinata ricerca. La rilevanza è misurata attraverso la somiglianza testuale della ricerca ed i contenuti della pagina. L'importanza globale di una pagina, indipendentemente dalla query, è giudicata dalla popolarità della pagina stessa, che nella maggior parte dei casi deriva dalla struttura dei link intorno ad essa. Lo spamming quindi è ogni azione umana mirata a garantire una rilevanza ingiustificata alla pagina.

Possiamo definire due macro-categorie per classificare le tecniche di spam:

1. La prima include le tecniche di amplificazione che garantisce e realizza l'alta rilevanza di una pagina (**boosting**).

2. La seconda racchiude quelle tecniche segrete che non alterano la query in se stessa, ma nasconde agli occhi umani le tecniche finalizzate a garantire l'alta rilevanza di una pagina (**hiding**).

Descriveremo queste tecniche e vedremo come vengono combattute. Essenzialmente arriveremo ad una conclusione che non esistono soluzioni finali, ma c'è il raffinarsi di alcune tecniche che vengono vanificate dal raffinarsi delle contromisure.

Capitolo 1

HITS (Hyperlinked Induced Topic Search)

(Sorgente autorevole in un ambiente ipertestuale)

Jon Kleinberg, docente associato alla *Cornell University*, è uno dei massimi esperti della teoria dei networks, in particolare del World Wide Web.
Kleinberg è stato un pioniere dell'analisi dei link per il potenziamento dei motori di ricerca; nel 1998 comincia a lavorare sull'algoritmo **HITS**.
La struttura di rete di un ambiente hyperlink può essere una ricca sorgente di informazioni sui contenuti dell'ambiente, ammesso che si abbiano a disposizione mezzi sufficienti per comprenderle.
In questa sezione verrà sviluppato un insieme di metodi per estrarre informazioni dalle strutture dei link di tali ambienti e verranno riportati degli esperimenti che dimostrano la loro effettività in una varietà di contesti nell'ambiente del World Wide Web. In particolare l'attenzione si focalizzerà sull'uso dei link per analizzare le collezioni di pagine rilevanti in una ricerca, relativa ad un determinato argomento, e per scoprire le pagine più autorevoli su tale argomento.
Il lavoro trae origine dal "problema della ricerca" sul World Wide Web; ricerca che può essere definita come il processo atto a scoprire le pagine rilevanti ad una data query.
Migliorare la qualità dei metodi di ricerca sul WWW è, al momento, un problema ricco e interessante e in molti modi ortogonale al concetto di efficienza dell'algoritmo e di memorizzazione. In particolare si consideri che i motori di ricerca correnti indicizzano una porzione piuttosto grande del WWW e rispondono in pochi secondi. Sebbene dovrebbe essere considerata l'utilità anche di uno strumento di ricerca con dei tempi di risposta più lunghi – che fornisca, però, dei risultati di valore più significativi per l'utente – è molto difficile dire in che modo tale strumento dovrebbe essere computato con questo tempo extra.
Chiaramente mancano delle funzioni obiettive che siano definite concretamente e, allo stesso tempo, corrispondano alla nozione umana di qualità.

Query e Sorgenti Autorevoli

Vediamo le differenze delle queries nei motori di ricerca di *Web Information Retrieval* e in quelli di *Information Retrieval*.

Non bisogna avere una concezione univoca della nozione di query, difatti esiste più di un tipo di query e la gestione di ciascuna di esse richiede una tecnica particolare. Si considerino per esempio i seguenti tipi di query:

- **Query specifica:** *Netscape supporta JDK 1.1?*
- **Query su argomenti più ampi:** *Trova informazioni sul linguaggio di programmazione Java;*
- **Query generiche:** *Pesca* (sono un esperto di botanica o un pescatore sportivo?);
- **Query per pagine similari (per pagine simili):** *Trova le pagine simili alla pagina java.sun.com.*

Concentriamoci per ora sui primi tre tipi di query che presentano diversi tipi di ostacoli nel Web:

1. La difficoltà nel trattare le query specifiche è dovuta innanzitutto a ciò che può essere definito il "*problema della scarsità*" o "SCARCITY PROBLEM"**.** Ci sono poche pagine che contengono informazioni specifiche e spesso è difficile determinarne l'identità. Il problema, dunque, è trovarle. Questa tipologia di query funziona bene nei motori di ricerca di *Information Retrieval*;
2. Per le query relative ad argomenti più estesi e generiche, ci si aspetta di trovare molte centinaia di pagine interessanti sulla rete; tuttavia, tale insieme di pagine potrebbe essere generato anche dal semplice *matching* con un termine, cioè dal fatto che ci siano, nei documenti, una o più parole che fanno matching con la stringa di ricerca, ma che in realtà non sono di interesse per quel particolare argomento. Di conseguenza, in questo caso, non è più un problema di scarsità, ma la difficoltà fondamentale può essere paradossalmente *"il problema dell'abbondanza"* o "ABBUNDANCE PROBLEM", ovvero il numero di pagine che possono essere ragionevolmente restituite dal motore di ricerca è troppo grande e abbassa i livello di "precisione".

Più ampio è l'indice e più ampio è il rumore. Dal punto di vista pratico restituire tanti risultati equivale a non restituire nessun risultato, nel senso che il numero di risultati restituiti è talmente ampio da essere umanamente ingestibile. Diventa, infatti, fondamentale il **Ranking.** Se l'informazione cercata non è presente nei primi 100 risultati (nelle prime 10 pagine nel caso di Google) allora è inutile continuare a cercare.

C'è stato un momento in cui le tecniche di *Information Retrieval* soffrivano di questi problemi. L'obiettivo dell'*Information Retrieval* è "riuscire in maniera efficiente a catalogare tutti i documenti, avere un indice quanto più ampio possibile e fornire risposte esclusivamente alle query specialistiche o specifiche". Purtroppo le query specifiche sono un insieme abbastanza piccolo rispetto all'insieme delle query prodotte dagli utenti nel Web.

La differenza principale tra *l'Information Retrieval* e il *Web Information Retrieval* è la natura **ipertestuale** del Web. Le informazioni ipertestuali devono essere usate per integrare le classiche informazioni usate nell'Information Retrieval per riuscire ad ottenere risultati di ricerca più efficienti per l'utente nel caso di abbondanza dei risultati.

Per fornire metodi di ricerca che funzionino in queste condizioni è, quindi, necessario un filtro che ci permetta di ottenere dalla collezione di pagine restituite un piccolo insieme di pagine "autorevoli " o "definitive".

La nozione di "autorevole", relativamente a query su argomenti estesi, è il punto focale dei motori di ricerca. Uno degli ostacoli fondamentali in cui si incorre è l'accuratezza del concetto di autorità in un particolare contesto di query: data una particolare pagina, come si fa a dire se è autorevole o meno?

E' utile discutere di alcune *complicazioni* che possono sorgere:

1. Innanzitutto si consideri l'obiettivo principale di restituire la home page dell'università di Harvard, www.harvard.edu, che è una delle pagine più autorevoli restituite se si esegue una query sulla stringa di ricerca "Harvard". Sfortunatamente esistono milioni di pagine Web, sulla rete, che contengono la stringa "harvard" e la pagina www.harvard.edu non è di certo quella in cui il termine compare più spesso o in un qualsiasi altro modo che possa aiutare nella ricerca. In realtà si sospetta che non ci siano delle misure "endogene" della pagina che ci consentano di definire la sua autorevolezza.
2. Si consideri, per esempio, il problema di trovare l'home page dei principali motori di ricerca: si potrebbe iniziare dalla query "motori di ricerca", ma c'è una difficoltà immediata consistente nel fatto che molti dei motori di ricerca autorevoli, come Yahoo o Altavista, non usano il termine nelle loro pagine. Questo è un fenomeno molto ricorrente, per esempio, non ci si deve aspettare che le home page di Honda o Toyota contengano il termine "industria automobilistica".

Analisi della struttura link

La struttura hyperlink tra le pagine della rete ci consente di aggirare molte delle difficoltà appena discusse.

Gli hyperlink codificano una considerevole quantità di giudizio umano latente e si ritiene che tale giudizio sia utile a definire la nozione di *autorità*. In particolare, la creazione di un link sul WWW rappresenta un'indicazione concreta del seguente tipo di giudizio: "*L'artefice della pagina P, nel momento in cui include un link alla pagina Q, ha in un qualche modo conferito autorevolezza alla pagina Q*".

Di conseguenza i link ci consentono di definire autorità potenziali semplicemente attraverso le pagine che puntano ad esse e ci permettono di aggirare il problema discusso precedentemente relativo alle pagine prominenti che non sono sufficientemente auto-descrittive.

Ci sono alcune caratteristiche fondamentali da tenere presenti per dare peso e importanza ad un link:

- Il link semplicemente per la sua esistenza in una pagina rappresenta un'informazione estremamente importante. In quanto autore della pagina sono a conoscenza dell'esistenza di un'altra pagina. Quindi c'è stato un processo mentale che mi ha portato a collegare la mia pagina ad un'altra pagina e di considerarla collegata, relata o semplicemente degna di essere linkata. Inoltre il testo usato per il link da me creato rappresenta un'informazione, una sintesi del testo nell'altra pagina che si sta collegando. All'inizio del Web c'era un modo estremamente sbagliato di scrivere i links, per esempio: *"Diego Armando Maradona è un giocatore.....per avere maggiori info clicca **qui**"*. Se invece di strutturare il link in questo modo, si creava un link con il testo ***"Diego Armando Maradona"***, si indicava al motore di ricerca che analizzava questa pagina un'importante informazione sul link, una sorta di etichetta del link.
- Altre caratteristiche fondamentali per il link sono la sua *età* e la *stabilità* del link, cioè se si tratta di un link rimasto fisso e accessibile per molto tempo. Questo significa che la pagina esiste e fa assumere peso al link.
- Molto importante, infine, è la relazione tra il *sito del link* e il *sito del target*. Ad esempio se il link si riferisce ad una pagina dello stesso sito (allo stesso dominio) - **Link Infra-Site**: in un contesto universitario un professore potrebbe semplicemente linkare una pagina di un altro professore semplicemente perché colleghi, questo indica che non c'è stato un processo mentale particolarmente importante nel creare il link. I *link infra-site* non sono particolarmente efficaci per dare autorità ad una pagina.

Dall'altra parte ci sono dei potenziali tranelli nell'applicazione dei link a questo scopo: prima di tutto, alcuni link sono creati per molte ragioni, molte delle quali non hanno nulla a che fare con il conferire autorità alle pagine; ad esempio un gran numero di link viene creato a scopi di navigazione (clicca qui per tornare al menu principale, home, next, back, up) mentre altri link rappresentano soltanto delle pubblicità a pagamento (Link di dubbia utilità).

Un altro problema è la difficoltà di trovare un bilanciamento appropriato fra i criteri di **"rilevanza"** e **"popolarità"**, ciascuno dei quali contribuisce alla nozione intuitiva di **"autorevolezza"**. Ci sono delle pagine che sono "**rilevanti"** e delle pagine che sono semplicemente **"popolari"** per la mia query, come ad esempio le pagine estremamente linkate dei siti generalisti come i portali, i siti di tecnologia, i blog. Questi siti sono linkati in modo estremamente naturale, ma non è detto che le pagine siano rilevanti semplicemente perché ci sono molti link che puntano lì.

La nozione di link è asimmetrica: La pagina A punta a B, quindi esiste un arco da A a B, ma non è detto che B punta ad A. Se anche B punta ad A, vuole dire che c'è un altro arco esplicito da B ad A.

Il lavoro qui svolto da Kleinberg punta ad usare questa nozione di link per estrapolare due concetti importanti*:* le **authority** e gli **hub.**

- **Il link tra un nodo ed un altro è come una dichiarazione di fiducia di un nodo verso il contenuto dell'altro nodo.**

E' importante considerare il problema inerente all'euristica da usare per localizzare le pagine autorevoli: fra tutte le pagine che contengono la stringa di ricerca, restituisci quelle che hanno il numero più alto di in-link (cioè di link in entrata).

- **Un authority è una pagina che ha molta autorità nel campo.**

Se si sta parlando del linguaggio Java, un motore di Information Retrieval restituisce tutte le pagine che contengono la parola Java. Una pagina con molta autorità su Java, ad esempio, è il sito della Sun, quindi una pagina a cui tutti i siti puntano.

Abbiamo già detto che per una gran quantità di query relative, ad esempio, ai motori di ricerca o alle industrie automobilistiche, alcune delle pagine più autorevoli non contengono nemmeno la stringa di ricerca associata. Allo stesso modo questa euristica potrebbe considerare una pagina universalmente popolare come netscape.com o yahoo.com come la pagina più autorevole rispetto ad una qualsiasi stringa di ricerca che la contenga.

In questa sezione viene proposto un modello basato sui link per il conferimento dell'autorevolezza e viene mostrato come questo porti ad un metodo che identifichi le pagine autorevoli del WWW per la ricerca di argomenti ampi (query ampie).

Il modello è basato sulla relazione che esiste fra le autorità per un argomento e quelle pagine che linkano a tali autorità: ci riferiamo a queste pagine con il termine **"hub".** L'hub può essere visto come **l'inverso di un'autorità.** L'hub punta a molte autorità ed *è in grado di indirizzare (smistare) verso queste autorità*, cioè queste parti autorevoli. La relazione tra hub e l'authority è molto forte. Una buona authority è puntata da buoni hub e un buon hub è puntato da buone authority.

Esiste un certo tipo di equilibrio naturale fra le pagine hub e le pagine autorevoli nel grafo definito dalla struttura link ed, in base a queste considerazioni, verrà sviluppato un algoritmo che permette di identificare entrambi i tipi di pagine simultaneamente. L'algoritmo opera su ***focused subgraphes*** del WWW che viene costruito a partire dall'output di un motore di ricerca testuale; la tecnica per costruire tali sottografi è progettata per produrre una piccola collezione di pagine che però contengono link alle pagine più autorevoli per un dato argomento.

Overview

L'approccio per scoprire nel WWW le fonti più autorevoli ha una natura globale, nel senso che si cerca di identificare la maggior parte delle pagine centrali per ricerche di argomenti vasti nel contesto della rete. Gli approcci globali comportano problemi di base che si presentano nella rappresentazione e nel filtraggio di molte informazioni, in quanto l'intero set di pagine rilevanti per una query relativa ad un argomento ampio può avere una dimensione molto ampia. Ciò è in contrasto con gli approcci locali che tentano di trovare la correlazione fra l'insieme di pagine web che appartengono ad un singolo sito o ad una intranet; in tal caso il volume di dati è molto più piccolo e spesso porta un insieme differente di considerazioni.

E' importante notare che i problemi principali che si incontrano sono fondamentalmente diversi dal problema del **"clustering":** il clustering indirizza la problematica di selezionare una popolazione eterogenea in sottopopolazioni che abbiano fattori comuni; nel contesto del WWW questo può includere la distinzione delle pagine relative a significati differenti del termine di ricerca. In tal modo il clustering è intrinsecamente differente dal problema di estrarre informazioni relative ad un vasto argomento, tramite la scoperta delle pagine autorevoli, sebbene la sezione successiva indicherà alcune connessioni fra le due problematiche.

Anche se siamo perfettamente in grado di scindere i molteplici significati di una stringa di ricerca ambigua (per esempio Windows o Gates), dovremmo comunque convivere con lo stesso problema di rappresentazione e filtraggio di un vasto numero di pagine che sono rilevanti per ciascuno dei significati del temine di ricerca.

Costruzione del Sottografo del WWW

Possiamo vedere qualsiasi collezione V di pagine iperlinkate (il WWW è un grafo) come un grafo diretto G=(V, E) dove i nodi V corrispondono alle pagine, mentre un arco diretto (p,q) in E, indica la presenza di un link da p a q. Diciamo che :

- l'**out-degree** di un nodo p è il numero di nodi a cui punta (i link in uscita);
- l'**in-degree** di un nodo p è il numero di nodi che hanno un link ad esso (i link in entrata).

Dato un grafo G, è possibile isolare piccole regioni o sottografi nel modo seguente:
se W<=V è un sottoinsieme delle pagine, useremo **G[W]** per indicare il **grafo indotto su W**: i suoi nodi sono le pagine in W e suoi archi corrispondono a tutti i link fra le pagine in W.
Supponiamo di eseguire una query su un argomento ampio, query specificata attraverso la stringa di ricerca **σ**: possiamo determinare le pagine autorevole mediante un'analisi della struttura link, ma prima è necessario determinare il sottografo del WWW sui cui dovrà lavorare l'algoritmo. L' obiettivo è quello di focalizzare lo sforzo computazionale sulle pagine rilevanti; così ad esempio potremmo restringere l'analisi all'insieme $\mathbf{Q_\sigma}$ di tutte le pagine che contengono la stringa di ricerca. Ciò però ha due svantaggi significativi:

1. innanzitutto questo set potrebbe contenere qualcosa come circa un milione di pagine e quindi portarci ad un considerevole costo computazionale;
2. in secondo luogo, abbiamo già detto che la maggior parte delle maggiori pagine autorevoli potrebbero non appartenere a questo insieme;

Idealmente, ci piacerebbe focalizzarci su una collezione $\mathbf{S_\sigma}$ di pagine con le seguenti proprietà:

1. ***S_σ è relativamente piccolo;***
2. ***S_σ è ricco di pagine rilevanti;***
3. ***S_σ contiene molte o la maggior parte delle pagine più autorevoli.***

Mantenendo S_σ piccolo, saremo capaci di affrontare il costo computazionale che deriva dall'applicazione di algoritmi non banali; assicurando, inoltre, che tale insieme sia ricco di pagine rilevanti, riusciremo a trovare più facilmente buone pagine autorevoli referenziate all'interno di S_σ.
Come possiamo trovare un tale insieme di pagine? Come facciamo a trovare S_σ?
Per un parametro **t** (tipicamente settato a 200) si selezionano dal motore di ricerca testuale le prime t pagine con il rango più alto per la query σ (questo insieme sarà detto "***root set R_σ***").
Questo insieme soddisfa le condizioni 1 e 2, ovvero S_σ relativamente piccolo ed S_σ ricco di pagine rilevanti; purtroppo, però non soddisfa la condizione 3, ovvero S_σ che contiene la maggior parte delle pagine più autorevoli.

Notiamo, infatti, che le prime t pagine restituite dai motori di ricerca testuali, conterranno tutte la stringa di ricerca σ, e dunque, R_σ è chiaramente un sottoinsieme della collezione Q_σ di tutte le pagine che contengono σ: abbiamo detto, però, che spesso anche Q_σ non riesce a soddisfare le 3 condizioni sopraelencate.

Inoltre è interessante notare che ci sono pochi link fra le pagine R_σ e dunque R_σ è privo di struttura. Per esempio, durante gli esperimenti svolti, si è osservato che il root set per la query "java", conteneva 15 link fra pagine appartenenti a diversi domini, e il root set per la query "chensorship" conteneva 28 link fra le pagine appartenenti a domini differenti. Questi numeri sono tipici per una varietà di query effettuate. Tali numeri dovrebbero essere confrontati con 200x199=39800 link potenziali che esistono fra le pagine del root set.

Possiamo usare il root set R_σ per produrre un insieme di pagine S_σ che soddisfi le condizioni che stiamo cercando: consideriamo un'autorità per l'argomento della query - sebbene tale pagina possa non apparire nell'insieme R_σ, è abbastanza probabile che però sia puntata da almeno una pagina in R_σ , per cui è possibile incrementare il numero di autorità nel sottografo espandendo R_σ con i link che entrano ed escono in tale sottografo.

In pratica si considerano le prime **t** pagine restituite da un motore di ricerca (per esempio le prime 200) e poi si aggiunge a tale insieme tutte le pagine più puntate da questo insieme e tutte le pagine che puntano maggiormente alle pagine di questo insieme. Alla fine si considerano le prime 10 come le fonti più autorevoli.

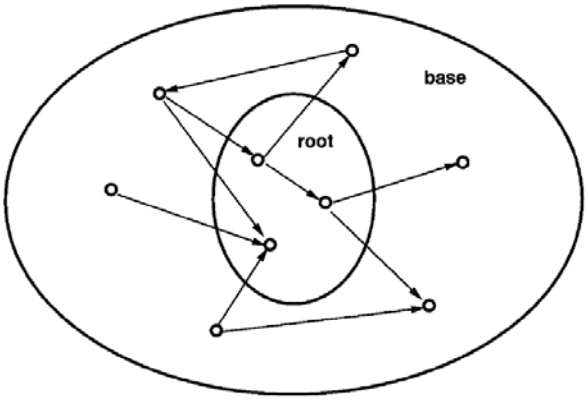

FIG. 1. Expanding the root set into a base set.

In termini concreti, definiamo la procedura seguente:

Sottografo (σ, E, t, d)

σ= stringa di ricerca;

E= motore di ricerca testuale;

t,d= numeri naturali;

Con R_σ denotiamo le prime t che risultano da E su σ (cioè le prime t pagine restituite dal motore di ricerca per la stringa σ)

Setta S_σ=R_σ //setta S_σ alle prime t pagine restituite dal motore di ricerca

For ciascuna pagina p in R_σ

Con Γ^+(p) denotiamo il set di tutte le pagine puntate da p;

Con Γ^- (p) denotiamo il set di tutte le pagine che puntano a p;

Aggiungi tutte le pagine in Γ^+(p) a S_σ

if $|\Gamma^-(p)|$<=d, then

Aggiungi tutte le pagine in Γ^+(p) a S_σ ;

else

Aggiungi un insieme arbitrario di d pagine in Γ^+(p) a S_σ ;

End

Return S_σ

In questo modo otterremo S_σ dalla crescita di R_σ, includendo in S_σ

- le pagine puntate da una pagina in R_σ
- le pagine che puntano ad una pagina in R_σ

Con la restrizione che consentiamo ad una pagina di R_σ di portare in S_σ al più **d** pagine che puntino ad essa. Quest'ultimo punto è cruciale in quanto alcune pagine web sono puntate da centinaia di migliaia di pagine ma noi vogliamo mantenere S_σ ragionevolmente piccolo.

Ci riferiremo ad S_σ come al ***Base Set*** di σ;

Negli esperimenti eseguiti verrà costruito tale Base Set invocando la procedura sopra elencata (Sottografo) con il motore di ricerca Altavista, t=200 e d=50.

Nei risultati di ricerca si nota che S_σ soddisfa, il più delle volte le 3 condizioni – la sua dimensione si aggira nel range 1000-5000 – e, come detto prima, l'autorità necessita solo di essere referenziata da una qualunque delle 200 pagine nel root set R_σ, per poter essere aggiunto ad S_σ.

Nella prossima sezione verrà descritto l'algoritmo per computare gli hub e le authority nel Base Set S_σ. Prima però, discutiamo un'euristica utile per considerare gli effetti dei link che servono semplicemente a scopi di navigazione.

Prima di tutto, con $G[S_\sigma]$ denotiamo il sottografo indotto sulle pagine in S_σ . Distinguiamo due tipi di link in $G[S_\sigma]$:

- **link trasverso:** è un link che si trova fra le pagine con differenti nomi di dominio;
- **link intrinseco:** è un link che si trova fra le pagine con lo stesso nome di dominio.

Per "nome di dominio" intendiamo il primo livello nella stringa URL associata ad una pagina.

I link intrinseci esistono solitamente per consentire la navigazione dell'infrastruttura di un sito; essi, rispetto ai link traversi, portano molte meno informazioni relative alle autorità delle pagine a cui puntano. Di conseguenza, i link intrinseci vengono eliminati dal grafo $G[S_\sigma]$ e vengono mantenuti solo gli archi corrispondenti ai link trasversi. Ciò darà luogo al grafo G_σ.

Questa è un' euristica veramente semplice, ma la troviamo funzionale per evitare molti dei problemi derivanti dal trattare i link utili alla navigazione allo stesso modo degli altri link.

Ci sono delle altre semplici euristiche che possono essere valutate per eliminare i link che non sembrano intuitivamente conferire autorevolezza. Un'euristica che ha senso menzionare è basata sulla seguente osservazione: "Supponiamo che un gran numero di pagine appartenenti ad uno stesso dominio, punti ad una singola pagina p; abbastanza spesso ciò corrisponde ad un avvertimento o ad un qualsiasi altro tipo di accordo fraudolenti tra le pagine che vi fanno riferimento, ad esempio, la frase "questo sito è stato progettato da..." e un link corrispondente alla fine della pagina in un dato dominio.

Per eliminare questo fenomeno è possibile fissare un parametro **m** (di solito compreso fra 4 e 8) e consentire solo ad m pagine, appartenenti allo stesso dominio, di puntare ad una qualsiasi pagina p. Di nuovo, si tratta di un'euristica molto semplice, ma potente in alcuni casi.

Computazione di Hub e Authority

Il metodo della sezione precedente ci fornisce un piccolo sottografo G_σ che si focalizza sull'argomento della query: esso contiene delle pagine molto rilevanti e delle pagine autorevoli. Ci preoccupiamo, adesso, di estrarre tali pagine autorevoli dalla collezione delle pagine, attraverso un'analisi della struttura link di G_σ.

L'approccio più semplice dovrebbe essere quello che sostiene di usare le pagine che si trovano nell'in-degree di G_σ - ossia le pagine che contengono il più ampio in-degree. Abbiamo rigettato questa idea precedentemente, quando la applicavamo alla collezione di tutte le pagine che contenevano la stringa di ricerca σ, ma adesso abbiamo costruito una piccola collezione di pagine rilevanti che contengono le autorità che vogliamo trovare. Tali autorità appartengono a G e sono referenziate dalle pagine contenute in G_σ.

Infatti, l'approccio del ranking tramite gli in-degree, tipicamente lavora molto meglio nel contesto di G_σ; in alcuni casi può produrre risultati di alta qualità in modo uniforme. Tuttavia, questo approccio ha ancora dei problemi significativi: per esempio, sulla query "Java" le pagine con il più ampio in-degree, sono quelle appartenenti a www.gamelan.com e www.java.sun.com , insieme alle pagine che danno consigli per le vacanze ai Carabi e la home page del libro Amazon.

Questo mix è rappresentativo del tipo di problemi che sorgono con uno schema semplice di ranking: mentre le prime due pagine www.gamelan.com e www.java.sun.com dovrebbero essere considerate una buona risposta, le altre due – pagine che danno consigli per le vacanze ai Carabi e la home page del libro Amazon –, invece, non sono significative per l'argomento della query; esse hanno un in-degree molto ampio, ma un'assenza di qualsiasi unità tematica.

La difficoltà di base è la tensione inerente che esiste all'interno del sottografo G_σ fra le pagine autorevoli e quelle che invece sono semplicemente molto popolari. Ci aspettiamo che quest'ultimo tipo di pagine abbiano un ampio in-degree senza però tenere conto dell'argomento di interesse per la query.

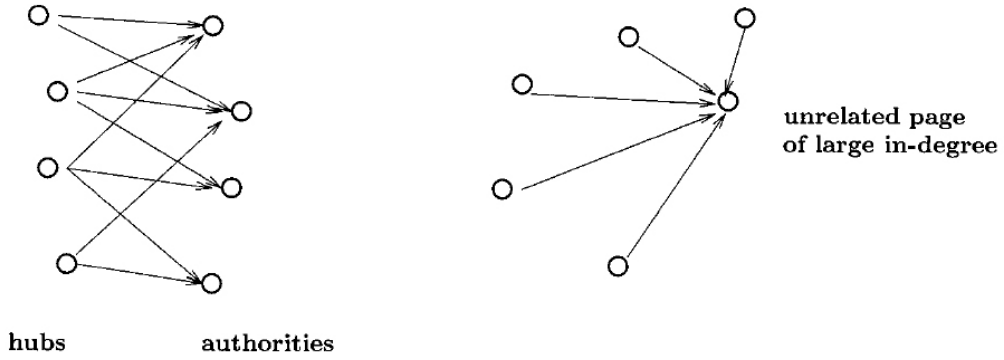

FIG. 2. A densely linked set of hubs and authorities.

Ci si potrebbe chiedere se per aggirare questo problema si può far uso del contenuto testuale delle pagine nella Base Set piuttosto che della struttura link di G_σ. Nella pratica è stato dimostrato che ciò è impossibile e che non è possibile ottenere risultati migliori di quelli che si riescono ad ottenere con l'approccio basato sui link, iniziando dalla seguente osservazione: le pagine autorevoli, rilevanti per la query iniziale, non dovrebbero avere soltanto un ampio in-degree; poiché sono tutte autorità relative ad uno specifico argomento, *tutte le pagine autorevoli dovrebbero condividere molte delle pagine che le puntano*.

In questo modo, in aggiunta alle pagine molto autorevoli, ci aspettiamo di trovare quelle che definiamo le *hub page*: si tratta di pagine che hanno link a molte pagine autorevoli e sono queste pagine che conferiscono autorità e ci consentono di eliminare quelle pagine non rilevanti per uno specifico argomento di query.

Le pagine Hub e le autorità esibiscono ciò che può essere considerata una "**relazione di mutuo rafforzamento**", nel senso che una buona pagina hub è una pagina che punta a molte buone pagine autorevole; una buona pagina autorevole è una pagina che è puntata da molte buone hub: chiaramente, se vogliamo identificare le pagine hub e le authority all'interno del grafo G_σ abbiamo bisogno di un metodo che interrompa tale circolarità.

Un algoritmo Iterativo

Si fa uso della correlazione fra le authority e gli hub attraverso un algoritmo iterativo che mantiene e aggiorna il peso numerico di ciascuna delle pagine; di conseguenza, a ciascuna pagina **p** associamo:

- un **peso di autorità** non negativo $\mathbf{x^{(p)}}$
- un **peso hub** non negativo $\mathbf{y^{(p)}}$

Manteniamo l'invariante che il peso di ogni tipo sia normalizzato così che la somma dei loro quadrati sia 1; cioè:

$$\sum_{p \in S_\sigma} (x^{(p)})^2 = 1$$

e

$$\sum_{p \in S_\sigma} (y^{(p)})^2 = 1$$

Vediamo le pagine con i valori di **x** e **y più grandi** come rispettivamente le **migliori authority** e i **migliori hub**.

Vediamo come HITS assegna valori (pesi) ad Authority ed Hub.

Numericamente è naturale esprimere la correlazione di mutuo rafforzamento fra hub e autorità nel modo seguente:

- Se p punta a molte pagine con grandi valori di x, allora gli dovrebbe essere assegnato un valore alto di y (**RICORDA: un buon hub punta a buone authority)**;
- se p è puntato da molte pagine con un grande valore di y, allora a p deve essere assegnato un valore alto di x (**RICORDA: una buona authority è puntata da buoni hub**).

Questo motiva la introduzione di 2 operazioni sui pesi, che possono essere denotate da ***I*** e **O.**

Dati i pesi $\{x^{(p)}\}$ e $\{y^{(p)}\}$, **<u>l'operazione I modifica il peso x</u>** (I sta per IN, cioè archi entranti nell'authority x) come segue:

$$x^{(p)} \leftarrow \sum_{q:(q,p) \in E} y^{(q)}$$

L'arco (q,p) indica che c'è un arco dall' hub y all' authority x.

Mentre **l'operazione O aggiorna il peso y** (O sta per OUT, cioè archi uscenti dall'authority x ed entranti nell'hub y) come segue:

$$y^{(p)} \leftarrow \sum_{q:(q,p)\in E} x^{(q)}$$

L'arco (q,p) indica che c'è un arco dall' authority x all' hub y.

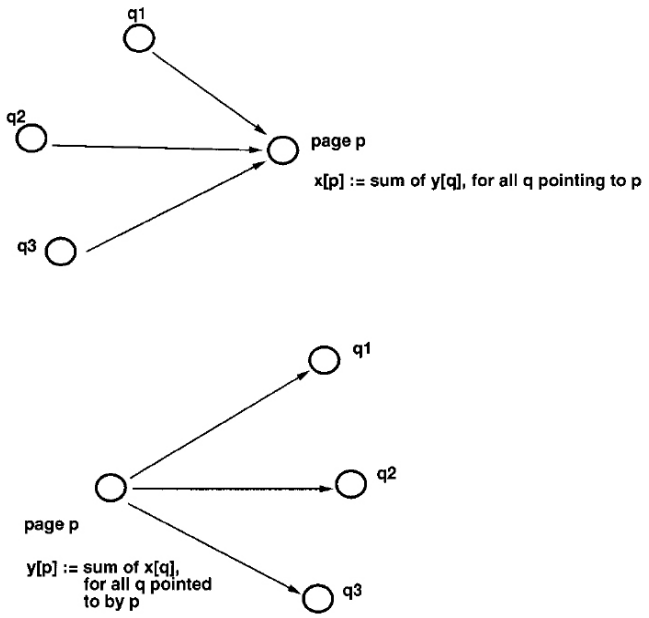

FIG. 3. The basic operations.

Grazie all'operazione I siamo in grado di stabilire quanto la mia pagina p è diventata autorevole.
Con l'operazione O siamo in grado di stabilire se la mia pagina p è un buon hub.

Di conseguenza I ed O sono gli strumenti di base mediante cui le hub e le autorità si rafforzano l'un l'altra. Per trovare i valori di equilibrio per i pesi è possibile applicare le operazioni I ed O in modo alternato e vedere se è raggiunto un punto fissato.
E' possibile ora dare una versione dell' algoritmo base: rappresentiamo l'insieme di pesi $\{x^{(p)}\}$ come un vettore x con una coordinata per ciascuna pagina in $G_{\sigma;}$ analogamente rappresentiamo l'insieme di pesi $\{y^{(p)}\}$ come un vettore y.

Iterate (*G*,*k*)

_G_= una collezione di _n_ pagine linkate

_k_= un numero naturale

Con _z_ denotiamo il vettore (1,1,1,....,1) in R^n

/*(il vettore di tutti 1 nello spazio n dimensionale)*/

Setta $x_0 = z$

Setta $y_0 = z$

For _i_=1,2,...,_k_

Applica l'operazione I a (x_{i-1}, y_{i-1}), ottenendo il nuovo peso x, ossia x'_i

Applica l'operazione O a (x'_i, y_{i-1}), ottenendo il nuovo peso y, ossia y'_i

Normalizza x'_i ottenendo x_i

Normalizza y'_i ottenendo y_i

End

Return (x_k, y_k)

Il passo di normalizzazione evita che il valore di authority e di hub cresca indefinitamente, in pratica si esprime il valore di authority e di hub di ogni pagina attraverso un valore compreso tra 0 e 1.

Normalizzare a 1 vuol dire che la norma (di solito la norma 2) delle componenti di un vettore (in questo caso per componenti di un vettore intendiamo i valori di authority e di hub associati alle pagine web) deve essere uguale a 1, ma le proporzioni devono rimanere invariate.

- **$\sum (x'_i)^2 = 1$ e $\sum (y'_i)^2 = 1$; $0 \leq x'_i, y'_i \leq 1$;**

Basta fare la radice quadrata della somma dei quadrati delle componenti del vettore e poi dividere ogni valore del vettore per il valore ottenuto (la norma).

Esempio:

- $\vec{v} = [a, b, c]$

Normalizzando:

- $\vec{n} = \frac{1}{\sqrt{a^2 + b^2 + c^2}} [a, b, c]$

Questa procedura può essere applicata per trovare le prime c autorità e i primi c hub nel seguente modo:

Filter (*G, k, c*)

***G*= una collezione di n pagine linkate**

***k, c*= un numero naturale**

(x_k, y_k)= Iterate (*G,k*)

Restituisci le pagine con le coordinate *c* più ampie in x_k come autorità

Restituisci le pagine con le coordinate *c* più ampie in y_k come hub

La procedura Filter con l'insieme G uguale a G_σ viene generalmente applicata con il valore di c tipicamente fra 5 e 10.

Limiti di HITS

Purtroppo non si può applicare HITS all'intero Web. Quello che si può fare è prendere un sottoinsieme di pagine che contengono la parola espressa attraverso la query, cioè un **CORE** (insieme significativo) di pagine. Di allargare questo **CORE** a distanza 1 attraverso tutte le pagine che puntano a questo **CORE** e tutte le pagine che sono puntate da questo **CORE**, e a questo punto applicare la procedura Iterate di Kleinberg. L'ampliamento serve soltanto per evitare problemi di sinonimia (come car, vehicle.....).

Si Limita la procedura Iterate ad un particolare numero di iterazioni ed inoltre si calcolano solo uno dei due vettori (authority o hub) per calcolare l'altro vettore al passo successivo come mostrato in precedenza. Non c'è bisogno di calcolare authority e hub contemporaneamente ad ogni passo. Si calcolano i valori di authority e al passo successivo usando questi valori si calcolano i valori di hub.

PRO

- **HITS** ha un **Ranking duale**, cioè per ogni query ottengo sia le authority e sia gli hub (per esempio nel caso di una ricerca su linguaggio Java ottengo sia i migliori siti che parlano di Java sia i migliori siti che linkano quest'ultimi);
- HITS trasforma un modello di Web Information Retrieval in un modello di Information Retrieval (infatti le tecniche iterative di calcolo dei pesi sulle pagine sono tecniche di Information Retrieval) su un **insieme di dati più limitato e quindi computazionalmente più gestibile;** viene usato il Web per selezionare la parte di Web su cui applicare gli algoritmi iterativi;

CONTRO

- Il problema principale che ha portato poi a sviluppare un Google PageRank piuttosto che un Google HITS, è che quest'ultimo algoritmo *si basa tutto sulla query;* ad ogni query si dovrebbe prima calcolare l'insieme n delle pagine, poi ampliare l'insieme ed infine applicare l'algoritmo per restituire le migliori authority ed hub; tutto questo non praticabile in pochi secondi;
- HITS è *molto suscettibile al Web Spamming*, in pratica attraverso il Web spamming si fa in modo che alcune pagine abbiano un certo ranking in una ricerca in maniera artificiale quando si effettua una particolare query)**;** Infatti *è molto facile diventare un buon Hub* perché è facile scrivere una pagina che contiene i migliori link (per esempio su linguaggio di programmazione Java) verso buone authority. Un volta diventato un buon Hub si può per esempio inserire anche alcuni link verso pagine che parlano di Java di scarso valore e in quel momento si sta facendo diventare anche queste pagine delle buone authority; quindi *è possibile far diventare artificialmente delle pagine delle buone authority*;
- Un altro problema è il cosiddetto *Topic Drift* ("drift" significa "deriva"), in pratica nell'espansione del **CORE** dell'algoritmo, si espande su pagine che magari hanno link su altri Topics (argomenti), quindi man mano che si itera il mio algoritmo si falsa il valore assoluto di ciascun authority;

Capitolo 2

Page Rank e l'ordine sul WEB

Kleinberg, con il suo algoritmo HITS, ha affermato che è possibile usare la struttura ipertestuale del Web per avere maggiori informazioni rispetto a tutte le tecniche standard di Information Retrieval.

Lawrence Page e Sergej Brin, due dottorandi all'università di Stanford, cominciano a lavorare a PageRank nel 1995. Il loro lavoro è influenzato dal lavoro di Kleinberg. Page e Brin a differenza di Kleinberg avevano un approccio pragmatico al problema: il loro obbiettivo era creare un motore di ricerca. Il loro scopo non era solo quello di risolvere il problema di un motore di ricerca dal punto di vista algoritmico, ma considerare tutta una serie di problemi che andavano dall'architettura, dal calcolo parallelo, programmazione funzionale, filesystem, reti, intelligenza artificiale...etc.

In questa capitolo verrà descritto il metodo di Page e Brin: **PageRank**, ossia un metodo per *qualificare le pagine Web in modo oggettivo e meccanico, misurando l'interesse umano e l'attenzione rivolta ad esse.*

Vedremo come PageRank computi in modo efficiente un gran numero di pagine e verrà mostrato come applicare il PageRank alla ricerca e alla navigazione dell'utente.

Introduzione e Motivazioni

Il World Wide Web ha introdotto dei cambiamenti significativi nel modo di ricercare le informazioni. E' molto ampio ed eterogeneo. Le pagine Web sono molto differenti fra loro, e i motori di ricerca devono anche considerare l'inesperienza degli utenti e la possibilità che le pagine siano ingegnerizzate per manipolare le funzioni di ranking dei motori di ricerca.
Tuttavia, diversamente dalle collezioni di documenti "flat" (piatte), il World Wide Web è ipertestuale e fornisce un considerevole ammontare di informazioni ausiliarie al top del testo delle pagine Web.
Innanzitutto forniamo una classificazione delle pagine Web che aiuta gli utenti a prendere atto velocemente della grande eterogeneità del mondo del Web

Diversità delle pagine WEB

PageRank parte dalle citazioni che sono espresse alla fine di un articolo scientifico (pubblicazione). Le citazioni di altri articoli in un articolo scientifico migliorano implicitamente il lavoro, la qualità e l'importanza degli articoli citati. Una citazione indica in qualche modo che i due articoli sono legati, similmente a ciò che avviene attraverso i links Web. Questo è il cosiddetto *impact factor*, cioè quante persone hanno letto quel particolare articolo e lo citano.
Sebbene ci sia già una grande mole di letteratura e di pubblicazioni accademiche (articoli scientifici), ci sono alcune differenze significative fra le pagine Web e le pubblicazioni. Diversamente dai documenti accademici, che sono sottoposti a revisioni scrupolose (i revisori sono di norma anonimi come l'autore dell'articolo scientifico per non influenzare in maniera artificiale la valutazione dell'articolo), le pagine Web sono libere da controlli di qualità e costi di pubblicazione. Con un semplice programma possono essere create molte pagine Web in modo semplice. Qualsiasi strategia di valutazione che conteggi aspetti replicabili di pagine Web può essere soggetta a manipolazione.
Inoltre, mentre gli articoli accademici sono unità di lavoro ben definite (auto-contenenti e contestualizzabili) e piuttosto similari in livello di qualità e in numero di citazioni, le pagine Web non hanno niente di tutto questo, variano in una scala molto più ampia per ciò che riguarda il livello di qualità, di lunghezza e obiettivi.
Per misurare l'importanza relativa di una pagina Web, si definisce il PageRank, che è un metodo per calcolare una classificazione di ciascuna pagina Web sulla base del grafo del Web.
Il PageRank trova applicazione nella ricerca e nella stima del traffico.

Un Ranking per le pagine sul WEB

Il Lavoro Relativo

Le citazioni sono importanti in questo contesto perché esistono tecniche note di Information Retrieval che sfruttano le citazioni.

Sono stati condotti molti studi sull'*analisi di citazioni accademiche*.

E' un naturale punto di partenza nel *Web Information Retrieval* perché questo approccio è molto simile a ciò che si potrebbe imparare nell'analizzare un link di una pagina Web.

Un link come una citazione: questo è il punto di partenza per il primo motore di ricerca di Lawrence Page e Sergej Brin che si chiamava BackLinks, la cui logica era:

- *"Conta quante pagine puntano a Tizio, se molte pagine puntano a Tizio allora forse quest'ultimo è rilevante"*. (**Una pagina molto citata è importante).**

Lawrence Page e Sergej Brin, però, già nel '98 sono consapevoli che questo approccio di contare tutte le pagine Web non è praticabile perché la dimensione del Web, già allora, era troppo grande: non si possono sapere con esattezza tutti i BackLinks di una pagina.

Inoltre, caratteristica principale, Page e Brin sono consapevoli che la sola citazione non basta, perché essa è automatizzabile (creare molte pagine che puntano a Tizio è facile), non è soggetta a revisione. In qualche modo bisogna pesare i BackLinks (le citazioni o link che puntano a Tizio devono essere pesati). C'è bisogno di un meccanismo che dia una misura della rilevanza di un link.

Nell'*Information Retrieval*, invece, le tecniche di BackLink possono essere applicate perché esiste un processo editoriale, quindi c'è revisione. C'è stata una gran quantità di attività recenti per esplorare la struttura link di un grande sistema ipertestuale quale è il WWW**.**

PageRank importa il meccanismo di "peso di un link" dall'algoritmo HITS di Kleinberg senza però importarne la complessità computazionale che rappresentava il vero scoglio all'implementazione di HITS. **PageRank è estremamente più semplice.**

La Struttura link del WEB

Anche se le stime variano, Il grafo corrente del Web ha circa 350 milioni di nodi (pagine) e 2.3 miliardi di archi (link). Ogni pagina ha molti link in avanti e indietro e non possiamo sapere se siamo riusciti a navigare tutti i link all'indietro di quella pagina, ma nel caso in cui l'abbiamo scaricata, allora siamo sicuri di tutti i links in avanti (forward link). Le pagine web variano molto nel numero di backlinks (links all'indietro).

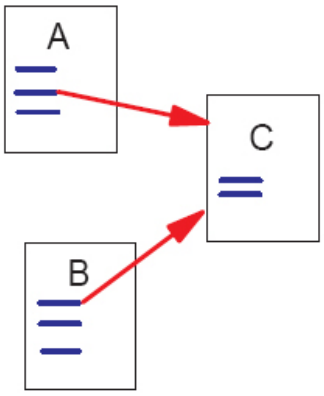

Figure 1: A and B are Backlinks of C

Generalmente le pagine linkate maggiormente sono più importanti di quelle che sono puntate di meno.

Propagazione del Ranking attraverso i link

Diciamo che una pagina ha un rank alto se la somma del rank dei suoi backlinks è alta. Ciò copre sia il caso in cui una pagina ha molti backlinks, sia il caso ha in cui una pagina ha pochi backlinks ma con un rank alto.

Definizione di PageRank

Sia **u** una pagina Web.

Poniamo:

- $\mathbf{F_u}$ = l'insieme di pagine puntate da **u** ("F" sta per Forward);
- $\mathbf{B_u}$ = l'insieme di pagine che puntano a **u** ("B" sta per Backward);

sia $\mathbf{N_u}$ in valore assoluto uguale alla cardinalità di $|\mathbf{F_u}|$, cioè il numero di link di **u**, e sia **c** un fattore usato per la normalizzazione cosi che la classifica totale di tutte le pagine web sia costante.

Una definizione iniziale di un **ranking R** di una pagina **u** che è una **versione semplificata di PageRank**:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Questa formalizza l'intuizione della precedente sezione**.**

La formula ha il seguente significato. Il rank della pagina **u** è uguale al prodotto di **c** per la sommatoria per tutte le pagine **v** che puntano ad **u** (cioè per ogni pagina in $\mathbf{B_u}$), del rank di **v** diviso il numero di links che escono da **v**.

Questa definizione è **ricorsiva**, ma si presta ad un algoritmo iterativo come quello di Kleinberg, in pratica si può partire con un qualsiasi insieme di rank e si può iterare la computazione e il calcolo finché converge. In essa c'è il concetto di **rango**, cioè l'importanza di una pagina.

L'interpretazione di questa definizione porta ad un *problema di flusso in un grafo*. Infatti, il rank di una pagina è suddiviso tra gli archi uscenti della pagina è **equidistribuito** tra tutti i links di **v** che puntano ad altre pagine, questo per contribuire anche ai ranks di queste pagine a cui essi puntano.

La costante "**c**" è minore di 1 perché c'è un numero di pagine che non ha link in avanti e il loro peso è perso dal sistema.

La **figura 2** dimostra la propagazione del rank da una coppia di pagine ad un'altra:

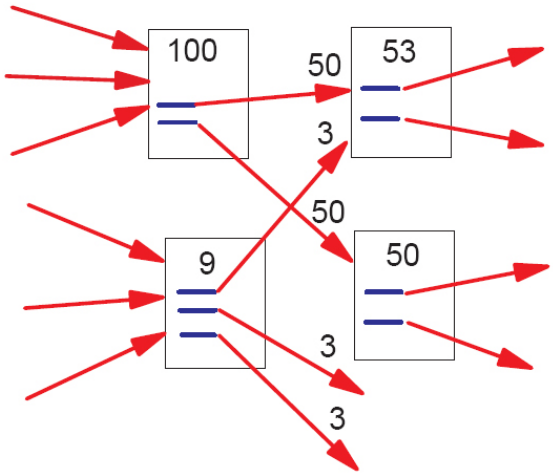

Figure 2: Simplified PageRank Calculation

Se c'è una pagina con rank 9 e con 3 links, allora ciascuno di questi links trasporta un rank uguale a 3.

La **figura 3** mostra invece una progressiva soluzione di stato per un insieme di pagine:

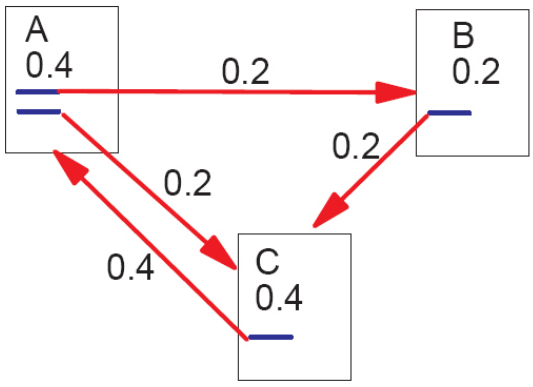

Figure 3: Simplified PageRank Calculation

Diamo adesso una definizione di PageRank attraverso l'algebra lineare trasformandolo in un problema matriciale.

Sia **A** una matrice quadrata **n x n** con righe e colonne che corrispondono alle **n** pagine web, sia l'elemento in posizione $\mathbf{A_{u,v}} = \mathbf{1/N_u}$ se c'è un arco da **u** a **v**, in un certo senso l'elemento esprime il peso dell'arco **(u,v)**, e $\mathbf{A_{u,v}} = \mathbf{0}$ se non c'è un arco. Se trattiamo **R** (il rank di tutte le pagine sul Web) come un vettore sulle pagine web, allora avremo che **R = cAR**.

Così **R** è un **autovettore** di **A** con **autovalore c** corrispondente. Infatti vogliamo l'autovettore dominante di A. Può essere computato applicando ad A ripetutamente un qualsiasi vettore di partenza non degenerato.

C'è un piccolo problema con questa funzione semplificata di classificazione: si considerino due pagine web che puntino l'un l'altra ma a nessun altra pagina e supponiamo che ci sia qualche pagina web che punti ad una di esse. Durante l'iterazione, il loop accumulerà rank ma non distribuirà mai nessun rank perché non ci sono outedge (archi uscenti dal loop). Il loop forma una specie di trappola che può essere chiamata **ranksink**.

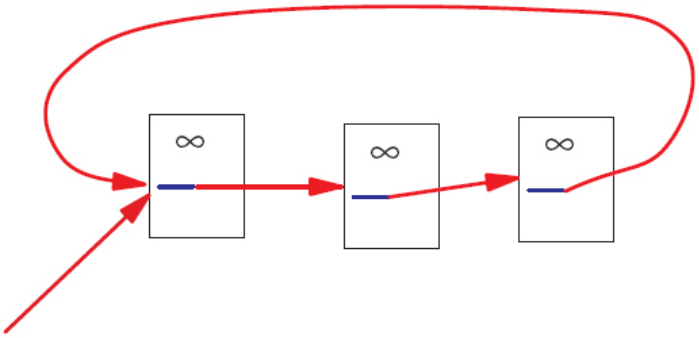

Figure 4: Loop Which Acts as a Rank Sink

Esiste inoltre la possibilità che si verifichi del **rankleak** (perdita di rango), cioè che un grafo perda **rank** a causa di nodi che accumulano soltanto e non ridistribuiscono.

Per far fronte a questi problemi si deve modificare leggermente la funzione di **PageRank**. La definizione di PageRank arriva a definire un modello della navigazione che un utente normalmente compie.

Attraverso il modello semplificato di PageRank visto prima, la probabilità che un utente visiti una nuova pagina dipende dal rank della pagina in cui si trova e dal suo numero di links uscenti. Nel modello semplificato, quindi, si presuppone che le pagine che hanno rank 0 (peso 0) non vengano mai visitate e che l'utente navighi esclusivamente attraverso links ipertestuali. Ogni utente naviga usando i links ipertestuali, ma spesso usa anche la barra degli indirizzi o i bookmarks. Il navigatore, quindi, spesso smette di percorrere la struttura a grafo e si sposta direttamente in qualche altra parte del Web.

L'idea quindi è la seguente:

> *"da una pagina A si può andare ad una pagina B, anche se non esistono links diretti dalla pagina A alla pagina B".*

La nuova funzione di ranking ha una componente che si riferisce alla classica navigazione Web, questa volta pesata da fattore **d** chiamato **dumping factor**, compreso tra 0 e 1, normalmente uguale a 0,87.

Nella nuova formula quindi viene moltiplicata la stessa formula vista in precedenza per un altro termine che indica la probabilità che si sceglie una pagina a caso tra le **n** pagine e cioè **(1 - d)**.

$$R(u) = d \cdot \sum_{v \in B_u} \frac{R(v)}{N_v} + (1-d) \cdot \frac{1}{n}$$

In altre parole la probabilità (il rank) che l'utente arrivi attraverso la navigazione alla pagina **u,** dipende si dal rank delle pagine che puntano verso **u**, ma dipende anche dalla probabilità che l'utente ad un certo punto decida di saltare direttamente (scrivendo l'indirizzo nella barra degli indirizzi) alla pagina **u**, questa probabilità è uguale a **(1 - d)**.

L'effetto che si ha con questa nuova formula di PageRank è che le pagine che avrebbero avuto un rank 0 di essere selezionate, ricevono una piccola probabilità di essere selezionate.

Grazie a questo metodo si evita che un insieme di pagine che non hanno links uscenti assorbano del ranking, visto che cmq da queste pagine possiamo spostarci in una delle **n** pagine attraverso metodi di navigazione diversi dai links ipertestuali.

Questo modello viene chiamato *random surfer* e dà l'idea di un utente che abbandona il normale flusso fornito da PageRank e salta ad una pagina in un'altra posizione con una certa probabilità.

Personalizzazione del Ranking

Il PageRank potrebbe essere personalizzato. Infatti la probabilità che un utente ad un certo punto abbandoni il normale flusso di navigazione ipertestuale e si sposti direttamente ad un'altra pagina (per esempio un giornale on-line, un particolare sito, etc) varia da utente ad utente a seconda dei gusti, passioni e necessità di quest'ultimo.

Già nel '98, dunque, Page e Brin avevano dato una prima anticipazione verso il modello cui oggi si sta evolvendo Google. Tutti i servizi aggiuntivi che Google offre oggi, come GMail, GoogleCalendar, GoogleDocs, etc.. sono tutti modi per acquisire informazioni sul profilo dell'utente.

Abbiamo introdotto PageRank, a questo punto possiamo fare un'osservazione e chiederci:

"Ma la query dov'è?"

La risposta è che la query non c'è, è tutto basato sulla struttura del grafo.

PRO PageRank

- È un algoritmo **query independend**, una pagina ha alto ranking o basso ranking in maniera indipendente dalla query;
- Il fatto di essere query independend permette a PageRank di **essere calcolabile offline** e memorizzare il rank delle pagine da qualche parte. Questo è un grandissimo vantaggio, perché uno dei problemi principali dei motori di ricerca che effettuavano computazione online era il tempo di risposta;
- È efficiente anche dal punto di vista computazionale (non richiede calcoli complessi come l'algoritmo di kleinberg);
- E' più difficile fare spamming

Capitolo 3

Motori di Ricerca e Web Spamming

Tutti i motori di ricerca sul Web, oggi, devono affrontare problemi che i sistemi di *Information Retrieval* non dovevano affrontare: utilizzare i motori di ricerca per particolari interessi e piegarli ai propri scopi, cioè fare in modo che i motori di ricerca rispondano artificialmente a determinate query in modo che determinate pagine abbiano un rank più alto nei risultati di ricerca, così da catturare l'attenzione dell'utente.

I motori di ricerca oggi rappresentano la porta d'ingresso al Web. Il *Web Spamming* ha l'obiettivo di *falsare il risultato restituito da un motore di ricerca* per catturare l'attenzione dell'utente. Recentemente l'ammontare di web spam è cresciuto vertiginosamente producendo un degrado dei risultati dei motori di ricerca. Il web spam inquina gli indici dei motori di ricerca con pagine non pertinenti, aumentando il costo per processare una query. Per fornire servizi di qualità a basso costo, un motore di ricerca deve individuare il web spam. Per combattere lo spam bisogna capire in profondità le tecniche adottate dagli spammer.

Presentiamo una tassonomia di tecniche di Web spamming che ci può aiutare a sviluppare appropriate contromisure.

L'obiettivo di un motore di ricerca è quello di fornire risultati di alta qualità individuando correttamente tutte le pagine web rilevanti per una specifica query, presentando all'utente alcuni tra i risultati più importanti di queste pagine rilevanti. La rilevanza di una pagina è misurata attraverso la somiglianza testuale tra la query effettuata e la pagina cercata. Alle pagine può essere dato un punteggio di rilevanza numerica specifico per la query. Più alto è il numero, più rilevante è la pagina per quella query. L'importanza si riferisce alla *popolarità globale* (query-independent) di una pagina, come può essere dedotto dalla struttura dei links (pagine con molti links in entrata sono molto importanti). In pratica i motori di ricerca usualmente uniscono *rilevanza* ed *importanza* per calcolare un ranking (una classifica) combinato di punteggio che è usato per ordinare i risultati della query presentati all'utente.

Il termine *spamming* (o *spamdexing*) si riferisce ad ogni deliberata azione umana intenta ad innescare una ingiustificata e favorevole rilevanza o importanza per alcune pagine web, considerando il valore reale della pagina. Il termine *spam* identifica, quindi, tutti questi oggetti (elementi di contenuto di pagina o links) che sono il risultato di alcune tecniche di spamming. Gli utenti che praticano spammig sono chiamati *spammers***.**

Esistono due categorie di tecniche si spamming**:**

- **Tecniche di Boosting (Tecniche di aumento del ranking):** metodi attraverso i quali si cerca di realizzare una alta rilevanza o importanza per alcune pagine;
- **Tecniche di Hiding (Tecniche di offuscamento):** metodi che non influenzano gli algoritmi di ranking (algoritmi di classificazione) dei motori di ricerca, ma che sono usati per nascondere (hide) le tecniche di Boosting adottate agli occhi degli utenti e a quelli del motore di ricerca, il quale non riesce più a controllare in automatico le tecniche di boosting.

Tecniche di Boosting

Ci sono diverse tipologie di tecniche di boosting. La **figura 1** ci mostra la tassonomia delle tecniche di Boosting:

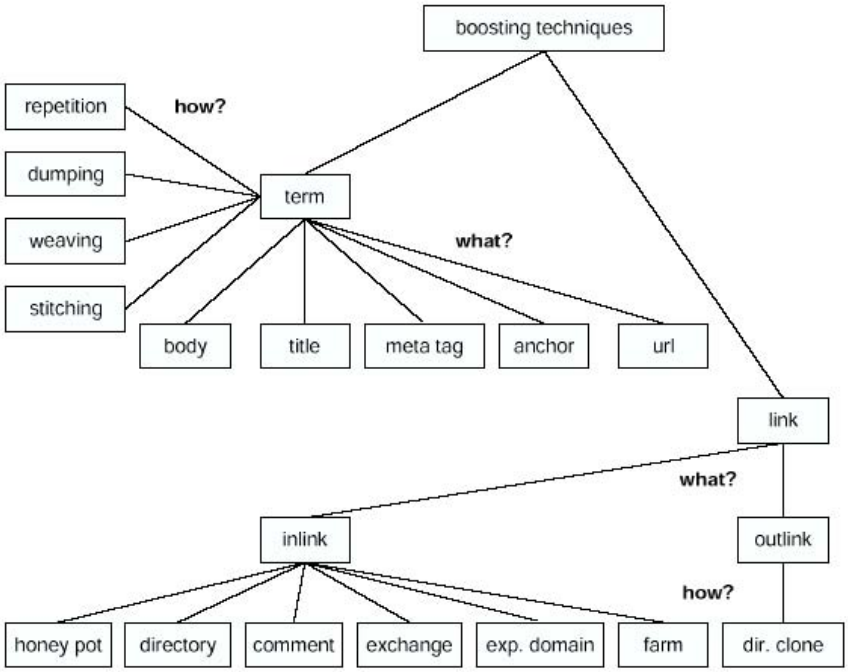

Figure 1: Boosting techniques.

Le principali tecniche di boosting sono due:

1. **Term Spamming**
2. **Link Spamming**

Si discuterà prima della tecnica *term spamming*, di come questa sfrutta le tecniche dei motori di ricerca (*target algorithms*) per alterare i risultati di ricerca, e di come la tecnica del *term spamming* viene vanificata dal raffinarsi delle contromisure. Successivamente, l'attenzione verrà focalizzata sulla tecnica del *link spamming.*

Term Spamming

Le tecniche di *term spamming* si riferiscono essenzialmente allo spamming rispetto alle tecniche tradizionali di *information retrieval*[2] che si basavano esclusivamente sull'analisi testuale dei termini. Nella valutazione della rilevanza testuale, in ogni motore di ricerca viene valutata la presenza dei termini della query nelle pagine web. Ogni tipo di locazione è chiamato *field* (*campo*). I campi di testo (*text fields)* comuni per una pagina *p* sono il *document body*, *il titolo*, *il tag meta* nell'header HTML e la *URL della pagina p*. In più, il testo ancora associato con le URLs che puntano a *p* stesso sono considerate anche appartenenti alla pagina *p* (campi di testo ancora). I termini nei campi di testo di *p* sono usati per determinare la rilevanza di *p* rispetto ad una specifica query (che rappresenta un gruppo di termini di query), spesso con differenti pesi dati ai diversi campi. *Term spamming* sono tecniche che confezionano il contenuto di questi campi di testo (*text fields*) con lo scopo di produrre pagine di *spam* rilevanti per alcune queries.

Target Algorithms

Gli algoritmi usati dai motori di ricerca per fare il *rank* delle pagine web, basato sui loro *text field*, usano la metrica **TFIDF** *(Term Frequency, Inverse of Document Frequency)* utilizzata nell'*Information Retrieval*.

La definizione della metrica **TFIDF** è la seguente:

Dato uno specifico *text field*, per ogni termine *t* che è comune al *text field* e alla query, **TF(*t*)** è la frequenza (*Term Frequency*) del termine *t* nel *text field*.

- *Esempio: Per una particolare istanza se il termine "mela" compare 6 volte nel body del documento su un totale di 30 termini, allora TF("mela") è 6/30=0.2.*

La frequenza inversa del documento (*Inverse Document Frequency*) **IDF(*t*)** di un termine *t* è relativa al numero di documenti nell'intera collezione che contengono il termine *t*.

- *Esempio: Per una particolare istanza se il termine "mela" appare su 4 dei 40 documenti della collezione, il suo IDF ("mela") sarà uguale a 10.*

[2] Vedi introduzione

Il risultato totale **TFIDF** di una pagina *p* rispetto ad una query *q* è calcolato su tutti i termini *t* in comune tra *p* e *q*:

$$TFIDF(p,q) = \sum_{t \in p \&\& t \in q} TF(t) \cdot IDF(t)$$

Il senso di questa formula è "*Se un certo termine è molto frequente in un certo documento (TF(t)) e non è presente o è poco presente nel resto dei documenti (IDF(t) - frequenza inversa del documento) allora vuol dire che quel documento è molto rilevante per quel termine di query, mentre se è presente anche negli altri documenti allora è poco rilevante*".
Attraverso il risultato di **TFIDF** gli spammers possono raggiungere due risultati importanti, entrambi rivolti a produrre una pagina rilevante:

- per un largo numero di queries (per esempio ricevere un risultato di TFIDF non zero);
- oppure produrre pagine molto rilevanti per una specifica query (esempio ricevere un TFIDF molto alto).

Il primo risultato può essere realizzato includendo un largo numero di termini distinti in un documento. Il secondo risultato può essere ottenuto ripetendo alcuni termini "target".
Assumiamo che gli spammers non hanno il reale controllo sui risultati IDF dei termini, ma solo del TF. Inoltre molti motori di ricerca ignorano completamente i risultati IDF. Così il primo modo per accrescere il TFIDF è aumentare la frequenza del termine all'interno di specifici campi di testo di una pagina.

Tecniche di Term Spam

Spamming di termini testuali relativo alle tecniche classiche di *Information Retrieval*. Le tecniche base di *term spamming* possono essere raggruppate in base al campo di testo in cui lo spamming occorre. Quindi distinguiamo:

- ***Body spam***: I termini di spam sono inseriti nel body di un documento. E' il più popolare e il più vecchio quanto i motori di ricerca stessi;
- ***Title spam*:** Oggi giorno i motori di ricerca danno un alto peso ai termini che compaiono nel titolo di un documento. Quindi ha senso inserire termini di spam nel titolo;

- ***Meta Tag spam***: I meta tags che compaiono nell'*header* di un documento HTML sono da sempre l'obbiettivo dello spamming. A causa del pesante spamming i motori di ricerca correntemente danno una bassa priorità a questi *tags*, oppure li ignorano completamente. Ecco un semplice esempio di **spammed keywords meta tag**:

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

- ***Anchor text spam***: Proprio come il titolo del documento, i motori di ricerca assegnano un alto peso ai termini testuali che rappresentano *ancore* (*link*), perché si suppone che offrano un sommario del documento puntato. Quindi i termini di spam sono spesso inclusi nel testo *ancora* degli hyperlink HTML verso una pagina. E' da notare che questa tecnica di spamming è differente dalla tecnica prima vista, nel senso che i termini di spam sono aggiunti non alla pagina target stessa, ma alle altre pagine che puntano alla pagina target. Così come il testo *ancora* riceve indicizzazione per entrambe le pagine, così lo spamming ha impatto sul ranking di entrambe le pagine sorgente e destinazione. Un altro esempio di testo-spam è:

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

- ***URL spam***. Alcuni motori di ricerca *spezzano* l'URL di una pagina nell'insieme di termini che sono usati per determinare la rilevanza di una pagina. Per sfruttare questo, gli spammers spesso creano lunghe URLs che includono sequenze di termini di spam. Per esempio ci si potrebbe imbattere in:

```
buy-canon-rebel-20d-lens-case.camerasx.com,
buy-nikon-d100-d70-lens-case.camerasx.com,
...
```

Spesso le tecniche di spamming sono combinate.

E' possibile raggruppare le tecniche di term spamming anche sulla base del tipo di termini che sono aggiunti ai campi di testo. Così possiamo avere:

- ***Repetition (Ripetizione)*** di uno o più termini specifici. In questo modo gli spammers realizzano una rilevanza maggiore per un documento rispetto ad un piccolo numero di termini di query; questa tecnica è facile da rilevare perché non è normale una ripetizione di termini molto vicini tra loro;
- ***Dumping (Riversamento - Scaricamento)*** di un largo numero di termini non collegati, spesso riguardanti l'intero dizionario. In questo modo gli spammers ottengono delle pagine rilevanti rispetto a molte queries differenti. Il Dumping è efficace contro quelle queries che includono termini relativamente rari e poco chiari: per queste queries è probabile che solo un paio di pagine sono rilevanti, così anche una pagina spam con basso livello di rilevanza/importanza potrà apparire tra i risultati al top;
- ***Weaving (Tessitura)*** di termini di spam in contenuti copiati. Molto spesso gli spammers duplicano il corpo del testo di nuovi articoli disponibili sul Web e inseriscono in essi termini di spam in posizioni random. Questa tecnica è efficace se l'argomento del testo originale è così raro che solo un piccolo numero di pagine esitono. Il Weaving è anche usato per disilludere e nascondere alcuni termini spam all'interno del testo, così che gli algoritmi dei motori di ricerca che filtrano le ripetizioni evidenti sono fuorviati. Un piccolo esempio di spam weaving è:

  *Remember not only **airfare** to say the right **plane** **tickets** thing in the right place, but **far cheap travel** more difficult still, to leave **hotel rooms** unsaid the wrong thing at **vacation** the tempting moment.*

Un sistema di riconoscimento del linguaggio naturale valuterebbe correttamente questa frase che darebbe punteggio alto per qualcosa che vende biglietti aerei, stanza di albergo...etc;

- ***Phrase stitching (cucire)*** è usato dagli spammers per creare contenuto velocemente. L'idea è quella di incollare insieme frasi o proposizioni possibilmente da diverse sorgenti; la pagina spam apparirà al top per le queries in ognuno degli argomenti delle proposizioni originali. Se uno spammer usasse questo documento come sorgente potrebbe creare il seguente collage:

  The objective of a search engine is to provide high-quality results by correctly identifying. Unjustifiably favorable boosting techniques, i.e., methods through which one seeks relies on the identification of some common features of spam pages.

I Limiti del term Spamming

Le tecniche di term spamming sono ben note ai motori di ricerca, nel senso che i motori di ricerca hanno imparato ad apprezzare già da *altavista*, che schermava i termini che venivano ripetuti in maniera innaturale. Essenzialmente, passata la n-esima ripetizione non c'era più fattore cumulativo per cui si assumeva che quella pagina fosse spamming e quindi si andava al contrario e si decrementava il contatore fino al rischio di essere *bannati,*cioè eliminati dai database dei motori perdendo ogni visibilità.

Nella maggior parte dei casi, quindi, l'individuazione del term spamming è *molto facile* da ottenere. Una frase reale, con una grammatica e una sintassi, è visualmente molto diversa da un'asettica lista di keyword (*parole chiavi*). Basta osservare semplicemente le differenze visuali tra frasi reali e testo rimpinzato di keyword e si è verificato quanto un paragrafo di testo sembra "*naturale*". Inoltre, molti webmaster spesso scrivono paragrafi di keyword con *segnali evidenti* di **keyword stuffing** (*parole chiavi ripetute in maniera eccessiva ed inutile*); È quasi come se mettessero un'insegna gigante "*Hei, Google! Qui si fa keyword stuffing*!" nelle loro pagine web. Questo rende il rilevamento delle keyword perfino più facile.

Visto che l'individuazione del *keyword stuffing* non può essere perfetta al 100% e siccome potrebbe generare alcuni "falsi positivi", un buon motore di ricerca non dovrebbe penalizzare una pagina web per aver usato il *keyword stuffing* ma potrebbe solo calcolare l'importanza/peso di una parola prendendo in considerazione quanto testo naturale sta intorno alla parola. Questo minimizzerebbe gli effetti negativi di un'errata interpretazione del testo.

Quindi, quando uno spammer ottiene una buona posizione con una pagina strapiena di keyword, tende a pensare che il *keyword stuffing* abbia funzionato, quando in realtà potrebbe essere possibile per la pagina ottenere posizioni persino migliori con testo chiaro e genuino invece che paragrafi di keyword.

Link Spamming

Accanto alle metriche basate sui termini, i motori di ricerca fanno affidamento sulle informazioni dei link per determinare l'importanza delle pagine web. Perciò gli spammers spesso creano strutture link sperando di accrescere l'importanza di una o più pagine. In questa sezione entrano in azione le conoscenze delle tecniche HITS e PageRank che vengono utilizzate per derivare informazioni sulla rilevanza e, quindi, sulla base della struttura ipertestuale delle pagine che compongono la nostra collezione di documenti.

Target Algorithms

Per discutere gli algoritmi target dei *links spam*, descriviamo il seguente modello. Per uno spammers ci sono tre tipi di pagine sul Web:

- ***Inaccessible page (pagine inaccessibili)*** sono quelle che uno spammer non può modificare e sono quelle fuori dalla portata dello spammer; lo spammer non può influenzare i links in uscita, potrebbe però ancora puntare alle pagine inaccessibili.
- ***Accessible page (pagine accessibili)*** mantenute da altre persone (presumibilmente non affiliate con lo spammer), ma che possono essere modificate in limitati modi da uno spammer. Per esempio uno spammer potrebbe postare un commento in blog e il commento potrebbe contenere un link verso un sito spam. Siccome uno spammer che si infiltra in pagine accessibile non è un utente onesto, assumiamo che avrà a disposizione un limitato budget *m* di pagine accessibili. Per semplicità si assume inoltre che lo spammer potrà aggiungere al massimo un link in uscita ad ogni pagina accessibile.
- ***Own page (pagine proprie)*** mantenute dallo spammer, che ha il completo controllo del loro contenuto. Chiameremo il gruppo di pagine proprie come ***spam farm***. Un obiettivo degli spammers è incrementare (*gonfiare*) l'importanza di una o più pagine proprie. Per semplicità diremo che c'è una singola pagina target *t*. C' è un certo costo di mantenimento associato con le pagine proprie di uno spammer (registrazione del dominio, web hosting), così si può assumere che uno spammer ha un limitato budget *n* di tali pagine, che non includono la pagina target.

Con questo modello in mente, ci sono due algoritmi **Hits** e **PageRank**, usati per calcolare il punteggio di importanza basato sulla informazione del link.

Link Spamming su HITS

L'algoritmo originale di HITS fu introdotto per classificare (rank) pagine riguardanti un determinato argomento (topic). E' molto comune, comunque, usare l'algoritmo su tutte le pagine nel Web per assegnare punteggi (score) globali di *hub* e di *authority* ad ogni pagina. In accordo alla definizione circolare di HITS, importanti pagine *hub* sono quelle che puntano a molte importanti *authority*, mentre importanti pagine *authority* sono quelle puntate da molti *hubs*. Un motore di ricerca che usa l'algoritmo HITS per classificare (rank) pagine ritorna come risultato di una query una mescolanza di pagine con il più alto punteggio (score) di hub e authority.

Il punteggio di hub può essere facilmente spammed aggiungendo links in uscita verso un grande numero di pagine ben conosciute e rilevanti, come per esempio ***www.cnn.com***, ***www.mit.edu***. Così uno spammer aggiungendo molti links in uscita (outgoing links) alla pagina target *t* potrebbe accrescere il suo punteggio di *hub*.

Ottenere, invece, un alto **punteggio di authority** è più complicato, perché esso implica avere molti links in entrata (*incoming links*) provenienti da presunti importanti *hubs*. Uno spammer può gonfiare il punteggio di hub delle sue *n* pagine (ancora una volta aggiungendo molti outgoing links ad esse) e successivamente far puntare queste *n* pagine verso la pagina target. I links provenienti da importanti hubs accessibili possono accrescere ulteriormente il punteggio di authority della pagina target *t*. Perciò, il ruolo qui è "***the more the better***": senza la limitazione del budget di *n* pagine lo spammer potrebbe avere tutte le pagine proprie e accessibili che vuole, che puntano alla pagina target. Pagine proprie non-target dovrebbero anche puntare ad altre molte importanti (conosciute) authority il più possibile***. In HITS quindi si sfrutta la simmetria tra hub e authority per ottenere facilmente dello spam.***

Link Spamming su PageRank

PageRank usa l'informazione del link in entrata (*incoming link*) per assegnare un punteggio globale a tutte le pagine nel Web. Esso assume che il numero di links in entrata verso una pagina è collegato con la popolarità media tra gli utenti web della pagina (le persone punterebbero verso pagine che trovano importanti). L'intuizione che sta sotto l'algoritmo è che una pagina web è importante se molte altre pagine web puntano ad essa. Corrispondentemente, PageRank è basato su un mutuo rafforzamento tra le pagine: l'importanza di certe pagine influenza ed è influenzata dall'importanza di alcune altre pagine. Un' analisi recente dell'algoritmo mostra che il **punteggio PageRank totale *PR(Γ)*** di un gruppo ***Γ*** di pagine (oppure una singola pagina) dipende da **quattro fattori (due che vanno a sommarsi e due che vanno a sottrarsi)**:

$$PR(\Gamma)=PR_{static}(\Gamma)+PR_{in}(\Gamma)-PR_{out}(\Gamma)-PR_{sink}(\Gamma)$$

Dove $\mathbf{PR_{static}}$ è la componente di punteggio dovuta dalla distribuzione di punteggio statico (salti random tra le pagine); $\mathbf{PR_{in}}$ è il punteggio ricevuto attraverso i links in entrata provenienti da pagine esterne; $\mathbf{PR_{out}}$ è il punteggio che lascia il gruppo di pagine Γ attraverso i links in uscita verso le pagine esterne; e $\mathbf{PR_{sink}}$ il punteggio perso (dissipato) dovuto a quelle pagine nel gruppo Γ che non hanno outgoing links.

Per un modello di spam farm, la formula precedente ci conduce verso una classe di link ottimali e strutturati che provvede a massimizzare il punteggio di PageRank della pagina target *t*. Una simile struttura ottimale è presentata nella **figura 2:**

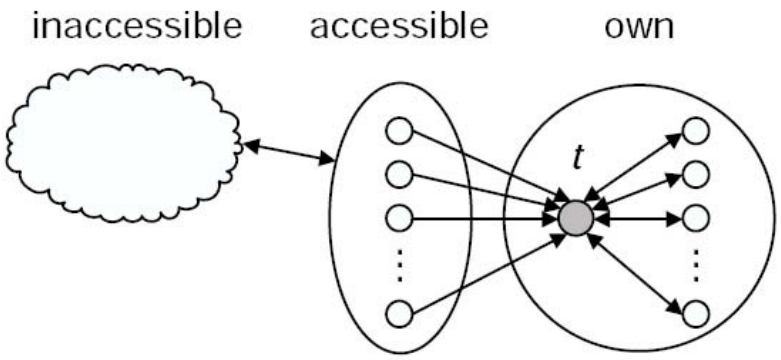

Figure 2: An optimal link structure for PageRank.

Questa struttura ha le desiderabili e discutibili proprietà che:

- tutte le pagine *proprie* di uno spammer devono essere raggiungibili da una pagina *accessibile* (così esse possono essere *crawled*[3] da un motore di ricerca per essere catalogate);
- ottenere questo risultato usando un minimo numero di links.

Possiamo osservare come la struttura presentata massimizzi il punteggio di PageRank totale della spam farm e in particolare della pagina *t*:

1. Tutte le *n proprie pagine* disponibili sono parte della spam farm, massimizzando il punteggio statico $PR_{static}(\sum)$;
2. Tutte le *m pagine accessibili* puntano alla spam farm, massimizzando il punteggio in entrata $PR_{in}(\sum)$;
3. I links che puntano fuori la spam farm sono soppressi creando un $PR_{out}(\sum)$ uguale a zero;
4. Tutte le pagine all'interno della farm hanno alcuni links in uscita rendendo zero il punteggio della componente $PR_{sink}(\sum)$.

All'interno della spam farm, il punteggio della pagina target *t* è massimizzato perché:

1. Tutte le *pagine accessibili* e le *pagine proprie* puntano direttamente al target, massimizzando il suo punteggio in entrata $PR_{in}(t)$;
2. La pagina target *t* punta alle altre *pagine proprie*. Senza tali links *t* avrebbe perso una significante parte del suo punteggio ($PR_{sink}(t) > 0$), e le *pagine proprie* sarebbero non raggiungibili al di fuori della spam farm. Inoltre i link uscenti dalla pagina target *t* sono finalizzati a dare rank e a riprenderlo dalle pagine proprie. Nota che non sarebbe stato saggio aggiungere links dalla pagina target alle pagine al di fuori della farm, perché questi links farebbero decrescere il PageRank totale della spam farm.

[3]Un **crawler** (detto anche **spider** o **robot**), è una delle componenti dei motori di ricerca che analizza i contenuti delle pagine web della rete intera in un modo metodico e automatizzato e le cataloga (**crawling**).

Tecniche di Link Spam

Possiamo raggruppare le tecniche basate su link spamming in:

- tecniche che aggiungono numerosi *outgoing links* verso pagine popolari;
- tecniche che raccolgono molti *incoming links* ad una singola pagina o ad un gruppo di pagine;

Outgoing links. Uno spammer può aggiungere manualmente un numero di outgoing links verso pagine popolari molto conosciute, sperando di accrescere il punteggio di hub della pagina. Allo stesso tempo il metodo più usato per creare un grande numero di outgoing links è il ***directory cloning***: si possono trovare sul Web un numero di siti a directory, alcuni dei più grandi e conosciuti sono **dmoz.org** oppure **dir.yahoo.com**.

Queste directories organizzano il contenuto web intorno ad argomenti (topics) e sottoargomenti (subtopics), e listano siti rilevanti per ognuno di questi argomenti. Uno spammers allora potrebbe semplicemente replicare alcune o tutte le pagine in una directory, e così creare una massiccia struttura di outgoing links velocemente.

Incoming links. Con lo scopo di accumulare un numero di incoming links verso una singola pagina target *t* o un insieme di pagine, uno spammer può adottare alcune delle seguenti strategie:

- Creare un ***honey pot (pentola di miele).*** Un insieme di pagine che fornisce alcune utili risorse (per esempio una copia della documentazione di Unix), ma che hanno anche links nascosti verso la pagina target spam *t*. Gli honey pot attirano gli utenti che sono tentati a linkare queste pagine e gonfiano (boosting) indirettamente il ranking della pagina target(s).
- **Infiltrare una web directory.** Diverse web directories permettono ai webmasters di postare links ai loro siti sotto alcuni argomenti (topic) nella directory. Può accadere che l'editor di tali directories non controlla e verifica strettamente l'aggiunta del link, o può essere raggirato da uno spammer esperto. In queste situazioni, gli spammers possono essere abilitati ad aggiungere nella directory links di pagine che puntano alle sue pagine target. Queste directories tendono ad avere sia un alto PageRank che un alto punteggio di hub, quindi questa tecniche di spamming va ad accrescere (boosting) sia il PageRank che il punteggio di authority delle pagine target.

- **Postare links sui blogs, sui message board non moderati, nei guest books o sui wikis.** Come già menzionato uno spammer può includere URLs alle sue pagine spam attraverso dei semplici ed innocenti commenti/messaggi che posta. Senza un editor o un moderatore che sorvegli tutti i commenti/messaggi sottomessi, sulle pagine di blog, dei message board, o guest book è facile collegarli a pagine di spam. Anche con la presenza di un editor o un moderatore, può essere non banale determinare commenti/messaggi di spam perché esso potrebbe impiegare alcune tecniche di hiding presentate nella prossima sezione. Ecco un esempio semplice di un commento spam in un blog che ha le caratteristiche sia di link che di text spamming:

```
Nice story.  Read about my <a href="http://
bestcasinoonlinever.com">Las Vegas casino</a>
trip.
```

E' importante menzionare che i commenti di spamming nei blog stanno guadagnando popolarità e questo non è solo un problema per i motori di ricerca, ma ha forti influenze direttamente sulla comunità dei milioni di bloggers: per gli utenti web che hanno un loro blog, i commenti di spamming rappresentano un fastidio simile all'e-mail spamming. Molti bloggers mantengono una lista di nomi di dominio che appaiono nelle spam URLs.

- **Partecipare allo scambio di link.** Spesso un gruppo di spammers configurano un struttura di scambio dei links così che i loro siti puntano ad ogni altro;
- **Acquistare domini scaduti.** Quando un nome di dominio scade, le URLs nei vari altri siti puntano alle pagine all'interno del dominio scaduto per lungo tempo. Alcuni spammers comprano domini scaduti e li popolano con spam che comporta vantaggio di falsa rilevanza/importanza portata dall'insieme dei vecchi links; *le keyworld all'interno di un nome di dominio hanno un grande peso per i motori di ricerca;*
- **Creare una propria spam farm**. Al giorno d'oggi gli spammers possono controllare un grande numero di siti e creare arbitrarie strutture di links con lo scopo di aumentare il ranking di alcune pagine target. Mentre questo approccio era proibitivo e costoso in passato, oggi è molto comune grazie all'abbattimento dei costi di registrazione di un dominio e di web hosting.

Link Spam basato su Stime di Massa

Ogni motore ha fissato una serie di regole e consigli utili (linee guida) da seguire per non correre il rischio di essere *bannati.*

I motori di ricerca **Inktomi**, **Google** e **Fast Search** considerano *spam* le pagine con un grande numero di link ma senza contenuto. Queste pagine vengono chiamate anche **FFA** (*Free-For-All*) perché permettono di generare falsa *Link Popularity*.

Google, in particolare, combatte contro tutte quelle tecniche che cercano di influenzare il PageRank. Ultimamente Google sta testando nuovi filtri per il rilevamento del link spam basato su **stime di massa**.

Le *link farm* (*spam farm*) sono dei *web host* che raccolgono link a pagine *web spam*, in modo da aumentare il PageRank delle pagine collegate. Dato che una pagina *spam* non viene linkata dal "popolo" di Internet, in quanto ritenuta inutile; ci sono due modi per aumentare il PageRank di una pagina *spam*:

1. avere un elevato numero di link entranti da pagine con basso PR(page rank)
2. avere pochi link segnalati da pagine con elevato punteggio.

Le *link farm* preferiscono il primo metodo perché è il più economico. Quindi creare molte farm collegate fra loro spesso su hosting gratuiti che spingano un migliaio di siti è moderatamente economico. Acquistare un link su un sito con un alto punteggio, innanzitutto, costa e poi non ne serve certamente uno, ma una bel gruppetto. I motori di ricerca hanno applicato delle contromisure per azzerare e penalizzare gli abusi e le farm.

Un metodo abbastanza spartano per risolvere il problema sta nel creare due gruppi disgiunti di pagine: uno racchiude quelle fidate (senza spam), l'altro con le pagine ritenute spam. Per determinare se una pagina non inclusa nell'insieme contiene spam si contano i link entranti. Se sono maggiori quelli dalle pagine rispettabili che dalle spam, la pagina è considerata come fidata, altrimenti viene collocata nel mucchio delle pagine spam. La quantità di PageRank accumulato da una pagina attraverso i link inseriti nelle pagine spam è detta *spam mass*.

Un alternativa al precedente metodo di classificazione consiste nel calcolare il *PR* apportato dalle pagine nel gruppo fidato e in quello spam. Se il punteggio fornito dalle pagine rispettabile è maggiore dello *spam mass*, la pagina viene messa assieme alle pagine fidate.

Il crescere delle farm e della diffusione capillare dello spamming hanno reso inefficaci le contromisure sopraelencate.

Oltre ai link inseriti nelle link farm, gli spammer riescono a introdurre link in pagine fidate. In alcuni casi l'autore di queste pagine fidate può non essere volontariamente consapevole di favorire lo spamming.
Il *linking spam* inserito nei *blog, forum o guestbook* è, in parte, risolvibile dall'amministratore dei siti fidati inserendo tag *no follow* nei commenti e vietando l'inserimento di link nelle firme dei post.
Gli spammer che creano *honey pot* usano tecniche che si basano sulla creazione (o generazione) di contenuti. Utenti inconsapevoli possono inserire un collegamento verso l'*honey pot*, senza capire che il loro link favorisce gli spammer. Questa tecnica è molto difficile da rilevare.
Gli autori delle pagine fidate non sono intenzionalmente promotori di spam, dunque è controproducente penalizzare questi siti usando i due metodi antecedenti.

Come fa una pagina ad trovarsi nell'insieme fidato?

L'insieme viene inizialmente riempito manualmente. Vengono scelti da persone in carne ed ossa una serie di siti fidati. Nei laboratori di *Yahoo* sono state fatte alcune sperimentazioni sulla determinazione dei due gruppi.
Molti siti fidati scelti sono stati prelevati da una directory gestita da utenti. Altri sono: siti di università (parecchi domini **.edu**), siti governativi statunitensi (**.gov**).
La lista dei siti spazzatura di partenza è quella che ogni motore di ricerca ha già raccolto nelle rilevazioni delle pagine spam. La *relative spam mass* di una pagina è una parte del suo PageRank ottenuto dalle pagine spam che la linkano.
Il metodo si basa sul modello "spartano" già enunciato, sebbene vi siano alcune differenze e miglioramenti.
Una pagina con pochissimo PageRank può essere esclusa dal calcolo siccome una piccola quantità di PR non dà benefici agli spammer. Inoltre le pagine spam hanno generalmente un PR maggiore di 2. Quindi è possibile escludere le pagine con basso punteggio dal calcolo per il rilevamento dello spam di massa.
Se un sito fidato viene linkato da una pagina spam, la sua *relative spam mass* è molto minore di quando potrebbe essere se il sito fosse non fidato (non necessariamente deve far parte di quelli spazzatura).
Naturalmente il fattore di *relative spam mass* deve essere mitigato, poiché si potrebbero creare delle *spam farm* per sfavorire dei siti bersaglio.

Anche con questo sistema tra le pagine fidate sono state riscontrate, manualmente, delle pagine che contengono spam nonostante siano veramente poche. Questo sta a significare che pure pagine che verrebbero considerate rispettabili sono finite nella spazzatura.
Per esempio nei risultati di spam sia il dominio *adobe.com*, questo è dovuto a un grandissimo numero di link al programma **Adobe Acrobat Reader**, sia *macromedia.com*

Tecniche di Hiding

E' usuale per gli spammers nascondere i segni eloquenti delle loro attività (termini ripetuti, una lunga lista di links). ***Le tecniche di boosting sono evidenti sia ai motori di ricerca che all'utente***. Gli spammers usano varie tecniche per nascondere i loro abusi agli utenti web che visitano le pagine spam, o agli editors delle compagnie dei motori di ricerca che cercano di identificare le istanze di spam. Le tecniche di hiding possono essere riassunte nella **figura 3:**

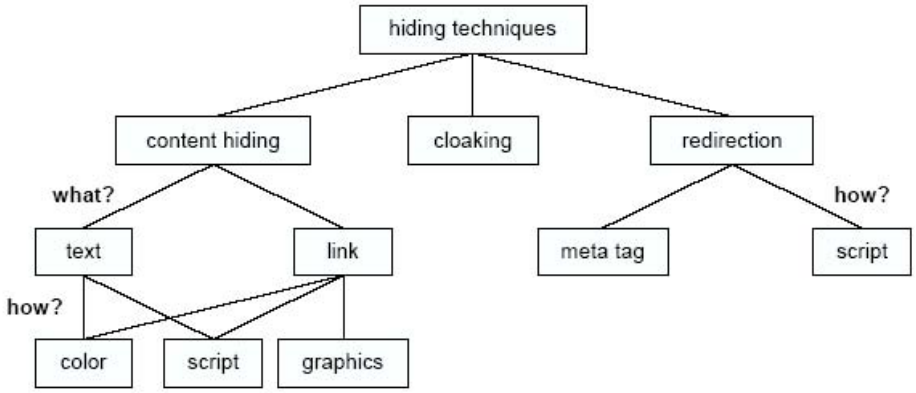

Figure 3: Spam hiding techniques.

Definiamo prima le tecniche di **Content Hiding**, poi le tecniche di **Cloaking**, infine le tecniche di **Redirection**.

Content Hiding

Termini di spam o links spam possono essere resi invisibili quando il browser visualizza la pagina. Una tecnica comune usa un appropriato schema di colori: termini nel body di un documento HTML non sono visibili se sono visualizzati sul display con lo **stesso colore del background**. Lo schema di colore può essere definito o nel documento HTML oppure nel file CSS allegato. Mostriamo un semplice esempio HTML:

```
<body background="white">
   <font color="white">hidden text</font>
   ...
</body>
```

In un modo simile, spam links possono essere nascosti evitando testo *ancora*. Spesso gli spammers creano minuscole **immagini *ancora* di 1x1 pixel** anche loro **trasparenti** o dello **stesso colore del background**:

```
<a href="target.html"><img src="tinyimg.gif"></a>
```

Uno spammer può anche usare *scripts* per nascondere alcuni elementi visuali nella pagina settando l'attributo di stile HTML ***visible=false.***
Riconoscere il background o il layout di una pagina è difficile da parte dei motori di ricerca. Lo stesso discorso per il significato delle immagini. Capire inoltre che cosa fa uno script è impossibile.

Rilevamento di Content Hiding

Non sempre è possibile riconoscere del contenuto nascosto all'interno di una pagina web. Nella pratica viene effettuato il *parsing*[4] del codice CSS e HTML estrapolando i valori dei colori del testo (e dello sfondo), salvandoli in alcune strutture dati. Successivamente, quando l'algoritmo effettua il *parsing* dell'HTML in cerca del testo, esso conosce in quali colori il testo e lo sfondo verrebbero disegnati da un browser. Se i due colori sono identici, il testo è considerato invisibile.
È possibile individuare anche i colori simili. Se il contrasto tra il colore di sfondo e del testo è troppo basso, il testo è considerato difficilmente percettibile dall'occhio umano e viene riportato un avviso.
Proprio dopo la fase di *parsing* del CSS e prima di quella dell'HTML, vengono scaricate anche tutte le immagini usate nella pagina web per i fondali. Siccome i metodi per nascondere il testo possono usare sfondi monocromatici, è necessario *pre-elaborare* tutti i fondali per capire se sono immagini monocromatiche o multicolori e per ricordare i valori dei colori necessari alla fase di *parsing* del codice HTML.
La fase di elaborazione dell'immagine è piuttosto veloce perché non è strettamente necessario analizzare tutti i pixel di un'immagine per capire se è monocromatica.

[4] In informatica, il **parsing** o **analisi sintattica** è il processo atto ad analizzare uno stream continuo in input (letto per esempio da un file o una tastiera) in modo da determinare la sua struttura grammaticale grazie ad una data grammatica formale (*wikipedia*)

Cloaking

Se gli spammers possono chiaramente identificare i *web clients crawler*, essi possono adottare la strategia del ***cloaking***: *dato un URL, degli spam web servers ritornano uno specifico documento HTML ad un regolare web browser, mentre essi ritornano un differente documento ad un web crawler*. In questo modo, gli spammers possono presentare l'ultimo contenuto inteso ai web users (senza traccia di spam nella pagina web), e allo stesso tempo inviare documenti spammed ai motori di ricerca per l'indicizzazione.
L'identificazione dei *web crawler* può avvenire in due modi. Da un lato gli spammers possono mantenere una lista di indirizzi IP usati dai motori di ricerca, ed identificare i web crawlers basandosi sul matching degli indirizzi IPs. Dall'altro lato, un web server può identificare la richiesta di un documento da parte di un'applicazione basandosi sul campo ***user-agent*** nel messaggio di richiesta HTTP. Per esempio il nome di user-agent usato dal browser Microsoft Internet Explorer 6 è:

```
GET /db_pages/members.html HTTP/1.0
Host: www-db.stanford.edu
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Win-
dows NT 5.1)
```

I nomi di *user-agent* non sono strettamente standardizzati. Le applicazioni includono questo nome nel messaggio di richiesta HTTP. Tuttavia, i *crawlers* dei motori di ricerca identificano essi stessi attraverso un nome diverso da quelli usati dalle tradizionali applicazioni di web browser. Questo è fatto con lo scopo di permettere ai webmasters di bloccare l'accesso a parte del contenuto, a parametri di controllo del traffico di rete, o in caso si compiano legittime ottimizzazioni. Per esempio alcuni siti forniscono ai motori di ricerca versioni delle loro pagine. La versione non riguarda i links di navigazione, immagini di presentazione, pubblicità, ma il contenuto della pagina. Questa specie di attività sarebbe anche accolta dagli stessi motori di ricerca, perché aiuta l'indicizzazione dei contenuti utili.

Redirection

Un altro modo di nascondere il contenuto spam in una pagina è attraverso il redirezionamento (redirection) automatico del browser ad una altra URL appena la pagina è caricata. In questo modo la pagina otterrà l'indicizzazione da parte del motore di ricerca perché contiene altre informazioni utili oltre alla redirezione, ma l'utente non vedrà mai la pagina; pagine con azioni di redirezione come intermediari (proxies o vie d'accesso) per i targets finali, in cui gli spammers tentano di servire all'utente il raggiungimento del loro sito attraverso i motori di ricerca.

La redirection può essere realizzata in diversi modi. Un approccio semplice è di trarre vantaggio dal meta tag ***refresh*** nell'header di un documento HTML. Settando il tempo di refresh a zero la refresh URL alla pagina target, spammers possono realizzare la redirezione prima che la pagina venga visualizzata nel browser:

<meta http-equiv="refresh" content="0";url=target.html>

Mentre il precedente approccio non è difficile da implementare, i motori di ricerca possono comunque facilmente identificare i tentativi di redirezione facendo il parsing dei meta tags. Alcuni spammers sofisticati realizzano la redirezione come parte di qualche *script* in una pagina. Essendo script essi sono eseguiti dai crawlers:

```
<script language="javascript"><!--
   location.replace("target.html")
--></script>
```

Soluzioni al Cloaking e Redirection

Attualmente non sono state ancora trovate delle contromisure adatte ed efficaci per combattere queste tecniche. Tuttavia è possibile segnalare *siti spamming* ai motori di ricerca in modo da aiutare a mantenere la qualità dei risultati di ricerca più pertinenti per ciascuna ricerca effettuata. Questi esaminano attentamente ogni segnalazione di pratiche ingannevoli e adottano le misure appropriate qualora si individuano veri e propri illeciti. In alcuni casi molto gravi, procedono alla rimozione immediata degli spammer dal loro indice onde evitare che compaiano nei risultati di ricerca.

Cosa Evitare

Volendo riassumere, quali sono le cose da evitare nel posizionamento di un sito?

- Non usare trucchi. Sono troppo rischiosi e hanno la necessità di un continuo controllo e manutenzione del sistema.
- Evitare l'uso di tecniche dichiaratamente non gradite dai motori, pena l'esclusione definitiva dagli indici (cloaking, redirect).
- Evitare l'eccessiva presenza di codice Javascript nelle pagine.
- Realizzare siti con troppi temi non attinenti tra loro, soprattutto in settori a forte competizione.
- Puntare con link siti spam.

Se un motore di ricerca valuta una pagina come spam, le conseguenze che ne derivano possono essere una penalizzazione del posizionamento delle pagine del sito; l’esclusione della pagina dal database del motore di ricerca, dell'intero dominio, dell'IP; segnalazione nelle "liste nere" dei motori di ricerca.

Il comportamento dei navigatori

Per meglio accontentare e fidelizzare i propri navigatori, i motori di ricerca studiano i comportamenti e la psicologia degli utenti mediante la registrazione e l'analisi di fenomeni che potrebbero in qualche modo evidenziare preferenze e necessità.

Un'alta durata media delle visite alle pagine di un sito, per esempio, potrebbe far ritenere che lo stesso sia interessante. Il concetto di durata di una visita è alquanto evanescente e molto ci sarebbe da dire a proposito, ma ciò esula da questa trattazione. In termini statistici sui grandi numeri, comunque, si accetta il significato e quanto rappresenta. Quindi se ci si ferma molto su una pagina, probabilmente è perché la stessa è interessante (anche se ovviamente si potrebbe essere semplicemente occupati in altro, pur avendo la pagina a video).

La richiesta di inserimento di una pagina tra i "preferiti" del browser, quando intercettata dal motore mediante un toolbar, sicuramente attesta l'interesse dell'utente per la stessa. L'apposizione sulla pagina di una funzione che faciliti il visitatore a fare ciò, potrebbe aumentare il numero di inserimenti e quindi di feedback ai motori, con evidenti benefici.

Il numero di click effettuati sulla *URL* della pagina, a fronte delle ricerche condotte, è un'ulteriore spinta al posizionamento. Scrivere buone descrizioni che invoglino il click, oltre a portare traffico, aumentano il rank avviando un loop in crescendo sicuramente vantaggioso.

L'inserimento di **Feed RSS[5] e Trackback[6]**, favorendo lo scambio dei link ed aumentando i feedback inviati ai motori, aumentano la sensazione di interesse per la pagina.

Studiare il comportamento dei navigatori, comprenderne gusti, esigenze e gestualità, portando tale conoscenza nelle pagine, conduce ad enormi benefici.

[5] Il **feed rss** è un'unità di informazioni formattata secondo specifiche (di genesi XML) stabilite precedentemente. Ciò per rendere interoperabile ed interscambiabile il contenuto fra le diverse applicazioni o piattaforme (*wikipedia*).

[6] Il ***Trackback*** è un meccanismo per la comunicazione e la notifica tra due risorse. Si è molto diffuso nei blog. In questo caso, un blog riceve una serie di ping (che contengono link) da altri blog e, solitamente, mostra, sotto ad ogni post, l'elenco dei ping ricevuti e riferiti a quello specifico post. (*wikipedia*)

Materiale bibliografico:

- Kleinberg "Authoritative Sources in a Hyperlinked Environment", Journal of ACM Vol. 46 No. 5, Sept. 1999 (primi 4 sezioni, escluso la dimostrazione del teorema 3.1) http://citeseer.ist.psu.edu/cache/papers/cs/18533/http:zSzzSzwww-dbs.cs.uni-sb.dezSzpublic_htmlzSzlehrezSzprosem00paperszSz1zSzkleinberg-jacm99.pdf/kleinberg99authoritative.pdf
- "The PageRank Citation Ranking: Bringing Order to the Web", di L. Page, S. Brin, R. Motwani, T. Winograd, a http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf&compression=&name=1999-66.pdf (par. 1 e 2 (ma solo per la definizione del funzionamento base di PageRank, fino alla Definition 1 esclusa))
- La definizione (corretta) di Pagerank viene presa da "Deeper inside PageRank", di Amy N. Langville and Carl D. Meyer, ACM Internet Mathematics, vol.1, n.3, a http://www.internetmathematics.org/volumes/1/3/Langville.pdf (solo paragrafi 1, 2, 3)
- Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005. http://airweb.cse.lehigh.edu/2005/gyongyi.pdf
- "The anatomy of a large-scale Hypertextual Web Search Engine" di Sergey Brin e Lawrence Page a http://www-db.stanford.edu/~backrub/google.html
- Dany Sullivan, "Search Engines Size" da Search Engine Watch (28/1/2005) a http://searchenginewatch.com/reports/article.php/2156481 con alcuni aggiornamenti a http://blog.searchenginewatch.com/blog/041111-084221
- Danny Sullivan, "Nielsen NetRatings Search Engine Ratings" (24/1/2006) a http://searchenginewatch.com/reports/article.php/2156451: i risultati piu' recenti di maggio: Google rappresenta il 50% delle query negli US a http://www.nielsen-netratings.com/pr/pr_060525.pdf
- First International Workshop on Adversarial Information Retrieval on the Web (Airweb 2005) a http://airweb.cse.lehigh.edu/2005/
- La pagina di Wikipedia su PageRank http://en.wikipedia.org/wiki/PageRank
- Google History a http://www.google.com/corporate/history.html
- Le ultime mosse di Google nel supercomputing a "Google's not-so-very-secret weapon" su International Herald Tribune del 31/6/2006 di J. Markoff e S. Hansell a http://www.iht.com/articles/2006/06/14/technology/web.0614search.php
- HITS e PageRank sono molto simili e appartengono ad una unica categoria di algoritmi: "PageRank, HITS and a Unified Framework for Link Analysis" di C.Ding, X. He, P. Husbands, H.Zha, H.Simon a http://www.siam.org/meetings/sdm03/proceedings/sdm03_24.pdf
- Lista delle pagine con un alto PageRank http://en.wikipedia.org/wiki/List_of_websites_with_a_high_PageRank
- http://en.wikipedia.org/wiki/Search_engine_optimization con molti link
- Un sito con alcune simulazioni (in Excel) ed alcune discussioni http://www.pagerank.dk/
- Alcune tecniche di redirection in Javascript: A Taxonomy of JavaScript Redirection Spam Kumar Chellapilla and Alexey Maykov, Third Workshop on Adversarial Information Retrieval on the Web (AIRWeb) 2007.
- Dany Sullivan, "Search Engines Size" da Search Engine Watch (28/1/2005) a http://searchenginewatch.com/reports/article.php/2156481 con alcuni aggiornamenti a http://blog.searchenginewatch.com/blog/041111-084221
- Danny Sullivan, "Nielsen NetRatings Search Engine Ratings" (24/1/2006) a http://searchenginewatch.com/reports/article.php/2156451: i risultati piu' recenti di maggio: Google rappresenta il 50% delle query negli US a http://www.nielsen-netratings.com/pr/pr_060525.pdf
- First International Workshop on Adversarial Information Retrieval on the Web (Airweb 2005) a http://airweb.cse.lehigh.edu/2005/
- La lista delle pubblicazioni dei "googlers" a http://research.google.com/pubs/papers.html
- http://www.princeton.edu/~kung/ele571/571-MatLab/571BP_Chad/kmeans.m Procedura di normalizzazione del vettore riga authority && hub