Motori di ricerca e Web Spamming

SICUREZZA SU RETI II (A.A. 2007-2008)

Prof. Alfredo De Santis

UNIVERSITÀ DEGLI STUDI DI SALERNO

A cura di:
Scala Santolo 0521/000268

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

Introduzione

- Quante volte avete provato a cercare su Internet un argomento di vostro interesse senza conoscere quale fosse la fonte da cui attingere le informazioni?

- Quante volte avete cercato di rintracciare un sito che vi aveva interessato la settimana precedente e non avete usato l'accortezza di registrarvelo subito tra i "Preferiti”?

- Quando avete avuto almeno una di queste necessità, sicuramente vi è venuto in aiuto un **Motore di Ricerca**.

La definizione di motore di ricerca

- *I **motori di ricerca** sono programmi (o software) che passano tutto il tempo a setacciare la grande quantità di informazioni presenti sul Web, creando immensi **elenchi** di tutte le informazioni.(encarta)*

Una definizione alternativa..

- *Un **Motore di Ricerca** è un sistema di accesso alle informazioni del Web che contiene un vastissimo archivio di siti per ciascuno dei quali in base alla richiesta dell'utente, il sistema riporta in un elenco, definito meglio come risultati della ricerca, un **Titolo** ed una **Descrizione** sintetica di quanto è trattato da ognuno dei siti individuati..(wikipedia)*

L'origine dei Motori di Ricerca

- Nascono nel 1995 per scopi accademici
- Prendono spunto dall'**Information Retrieval**
  - Tecniche usate
    - Costruzione indici, strutture dati
- Non solamente ricerca di un pattern
  - Il grado di somiglianza tra questi due documenti
    - Esempio questo documento è abbastanza simile a quest'altro?

L'origine dei Motori di Ricerca (2)

- Anno 96-98 c'è una rivoluzione nell'information retrieval dovuta alla nascita del web
  - Algoritmo di Kleinberg
  - Algoritmo Page Rank

- Il web rappresenta un ideale piattaforma per adoperare queste tecniche.

Come si utilizza un motore di ricerca?

- E' sufficiente digitare una frase che riguarda l'argomento di interesse, cliccare sul pulsante **"Cerca"** e il Motore *"gira"* riportando dopo pochi secondi migliaia di risultati suddivisi (a meno che non si voglia diversamente) in numero *n* alla volta per pagina.

- L'utente può fornire diversi tipi di query
- Ogni query richiede un approccio diverso
- Esistono 3 tipi principali di query:
  - Specific queries: "Qual'è il quarto pin di una porta nand 74l600?"
  - Broad topic queries: "Trova informazioni sui web browser?"
  - Similar-page queries: "Trova pagine simili a www.toyota.it "

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

I Problemi relativi alle varie query

- **Specific queries** -> problema della povertà: esistono pochissime pagine che contengono le informazioni volute ed è difficile identificarle
- **Broad topic queries** -> problema dell'abbondanza: il numero di pagine che potrebbero essere ragionevolmente rilevanti è troppo grande per un utente umano

I Limiti dell'Information Retrieval

- Le tecniche di "Information Retrieval" funzionano particolarmente male con le broad topic queries
  - Più ampio l'indice maggiore è il rumore
  - Informazioni presentate in maniera piatta, derivate "solo" dalla frequenza della parola all'interno del documento

- L'utente nelle query ad ampio raggio (broad topic queries) va cercando le informazioni più importanti, più **autorevoli**..

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

La nozione di authority

- In base a quale criterio stabiliamo l'autorità di una pagina?
- L'approccio classico:
  - Text-based (Within document ranking): una pagina è autorevole se contiene tante volte le parole della query.

Esempio

- Problema del contenuto: quante volte il sito della FIAT contiene il termine "produttore di automobili" ?
  - 0 termini
- Problema dello spam: un utente che voleva aumentare il rank della sua pagina inseriva nel tag metadati le parole volute

Esempio Query (ALTAVISTA)

- **(java) Altavista**
  - http://www.gamelan.com/
    - *Gamelan*
  - http://java.sun.com/
    - *JavaSoft Home Page*
  - http://www.digitalfocus.com/digitalfocus/faq/howdoi.html
    - *The Java Developer: How Do I . . .*
  - http://lightyear.ncsa.uiuc.edu/;srp/java/javabooks.html
    - *The Java Book Pages*
  - http://sunsite.unc.edu/javafaq/javafaq.html
    - *comp.lang.java FAQ*
  - http://www.isolajava.com/
    - *The island of java*
  - www.javacoffeebreak.com/
    - *Java coffee*

Analisi della struttura dei link

- Un approccio migliore è quello di usare i link che implicitamente codificano un giudizio umano
  - Rivoluzione da "Information Retrieval" in "**Web** Information Retrieval"

- Il concetto cardine: l'autore della pagina p, includendo un link alla pagina q (sua successore) ha in qualche modo conferito autorità alla pagina q
- Approccio classico-> una semplice euristica:

  Fra tutte le pagine che contengono la query string, seleziona quelle con un elevato numero di link entranti

  Difetti:
  - Si confonde attinenza con popolarità
  - Soggetta a spam: creazione di pagine finte che puntano alla pagina di cui si vuole far aumentare il rank

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

Definizione di Authority && Hub

- Prima definizione
  - Un Autorità è una pagina puntata da molte altre pagine
- Definizione di Kleinberg
  - **L'autorità non è solo chi è puntato da molte pagine, ma è chi è puntato da "buone pagine" (hub)**

HITS (Kleinberg)

Relazione di mutuo rinforzo

- Una buona *authority* è puntata da buoni *hub*
- Un buon *hub* punta a buone *authority*

Overview algoritmo di Hits

- L'algoritmo di Hits usa una combinazione di ricerca: text-based + link-based.
- L'algoritmo si divide in 2 passi:
  - Costruzione del focused subgraph Sσ
  - Calcolo degli Hub e Authorities da Sσ
- L'algoritmo opera sul web graph:
  - Le pagine del web sono i nodi del grafo
  - I link entranti o uscenti sono gli archi del grafo

Costruzione del focused subgraph Sσ

L'obiettivo: ottenere un insieme di pagine Sσ con le seguenti proprietà:

(a) È relativamente piccolo: l'algoritmo richiede poche risorse computazionali

(b) È ricco di pagine rilevanti: è facile trovare buone authorities

(c) Contiene molte delle authorities più forti

- Gli oggetti necessari:
  - ε: un motore di ricerca text-based: AltaVista oppure HotBot
  - σ: una query string di tipo broad topic
- Come troviamo Sσ ?
  - Root set Rσ: composto dalle t (≈200) pagine con più alto ranking risultanti da una query σ al motore ε. Esso rispetta (a) e (b) ma non (c).
  - Base set Sσ: espandiamo Rσ con i successori e con i predecessori (al max d ≈ 50) di ogni pagina in Rq

Costruzione del focused subgraph Sσ

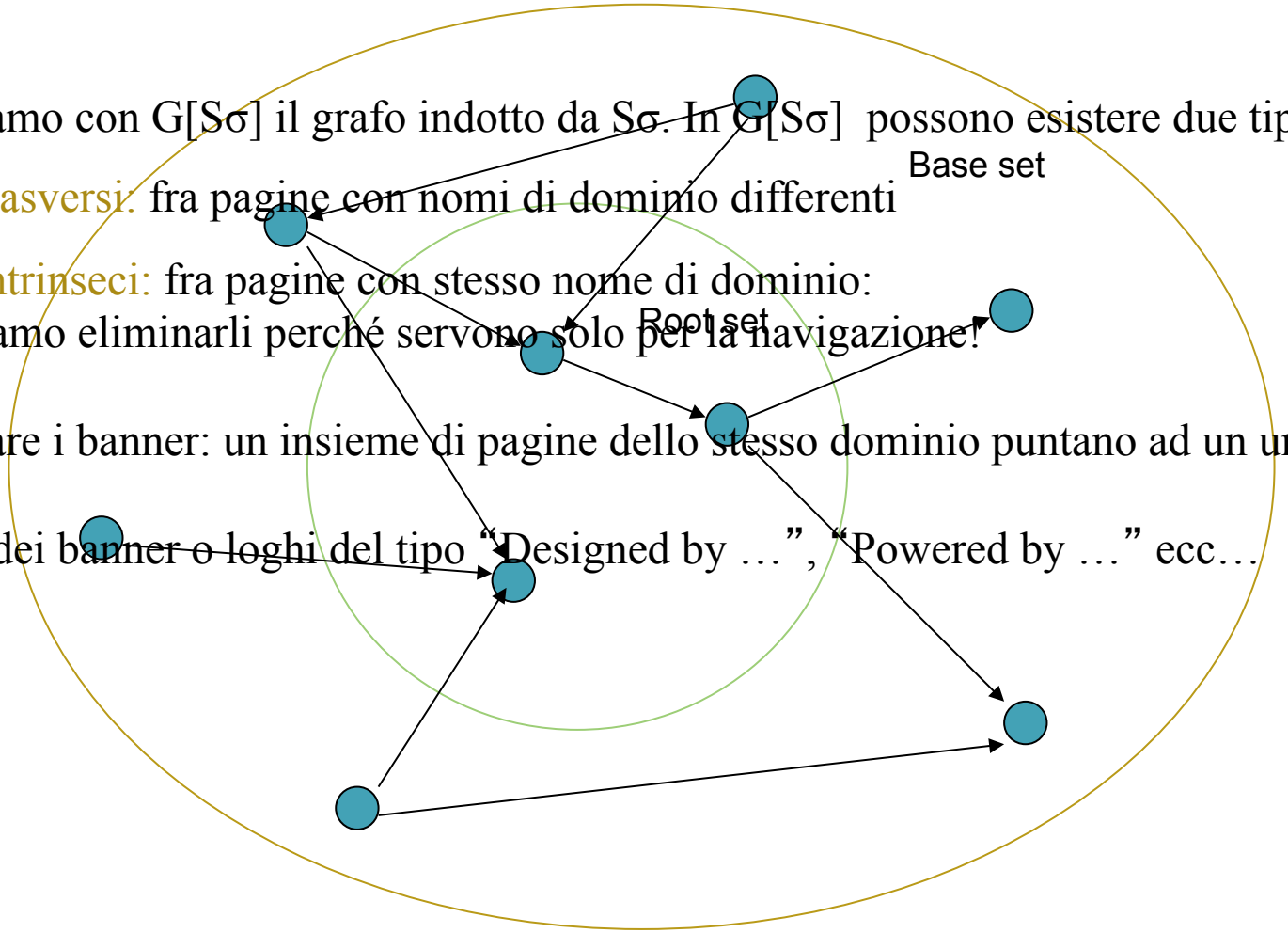
- Lo pseudo-algoritmo, come funzione:
- Subgraph(ε,σ,d)

```
Sσ = Rσ
For each page p Є Rσ
   Out(p)= i successori di p
   In(p)= i predecessori di p
   Aggiungi tutte le pagine Out(p) a Sσ
   If |In(p)|≤d then
        Aggiungi tutte le pagine in In(p) a Sσ
   else
        Aggiungi un insieme arbitrario di d pagine da In(p) in Sσ
  end if
end for
return Sσ
```

Costruzione del focused subgraph Sσ

Denotiamo con G[Sσ] il grafo indotto da Sσ. In G[Sσ] possono esistere due tipi di link:

- link trasversi: fra pagine con nomi di dominio differenti
- link intrinseci: fra pagine con stesso nome di dominio: dobbiamo eliminarli perché servono solo per la navigazione!



Eliminare i banner: un insieme di pagine dello stesso dominio puntano ad un unica pagina p.
Tipico dei banner o loghi del tipo "Designed by …", "Powered by …" ecc…

Calcolare Hub e Authorities

- L'obiettivo:
  - estrarre le authorities dalla collezione di pagine Sσ, attraverso un analisi della struttura dei link di G[Sσ].
  - Distinguere fra pagine "universalmente popolari" (alto grado di ingresso) da quelle autorevoli
- Osservazione:
  le pagine autorevoli, non solo devono avere un alto grado di ingresso, ma ci dovrebbe essere una sovrapposizione considerevole dei loro predecessori.
- Riassumendo:
  - Authorities: pagine buoni sorgenti di contenuto
  - Hubs: pagine buoni sorgenti di links verso molte Authorities
- Una relazione mutuamente rinforzante:
  - Buon hub: pagina che punta a molte buone authorities
  - Buon authority: pagina che è puntata da molti buoni hubs
- Individuare hubs e authorities: dobbiamo rompere la circolarità

  ci serve un algoritmo iterativo !

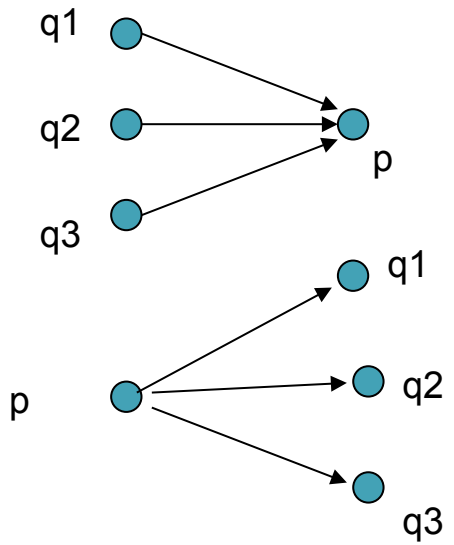
L'algoritmo iterativo (1)

- I vettori dei pesi:
  - ogni pagina ha associato un peso di authority a[i] ed un peso di hub h[i]
  - Quindi abbiamo due vettori n-dimensionali (n= |Sσ|)
  - I vettori a ed h sono normalizzati:
    $\sum a[i]^2=1$ e $\sum h[i]^2=1$ $0 \le a[i],h[i] \le 1$

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \end{bmatrix} \qquad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$$

- L'operazione di IO:
  - esprime la relazione di mutuo rinforzo fra h ed a
  - Operazione O: se p punta a molte pagine con alti valori di a, allora riceverà un alto valore di h
  - Operazione I: se p è puntato da molte pagine con alti valori di h, allora riceverà un alto valore di a

$$a(p) = \sum_{q \to p} h(q) \quad , \quad h(p) = \sum_{q \to p} a(q)$$



RISULTATI HITS

(Gates) Authorities

- http://www.roadahead.com *Bill Gates: The Road Ahead*
- http://www.microsoft.com/ *Welcome to Microsoft*
- http://www.microsoft.com/corpinfo/bill-g.htm

(Search Engines) Authorities

- http://www.yahoo.com Yahoo
- http://www.excite.com/ Excite
- http://www.mckinley.com/ Welcome to Magellan!
- http://www.lycos.com/ lycos home page

(Censorship) Authorities

- http://www.eff.org *EFFweb—The Electronic Frontier Foundation*
- http://www.eff.org blueribbon.html *The Blue Ribbon Campaign for Online Free Speech*
- http://www.cdt.org *The Center for Democracy and Technology*
- http://www.vtw.org *Voters Telecommunications Watch*
- http://www.aclu.org *CLU: American Civil Liberties Union*

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

La genesi

- **1998: John Kleinberg, professore associato a Cornell University, lavora su HITS**
  - Tecnologia per motore di ricerca
  - Basato sulla struttura ipertestuale del Web
- **Presenta il suo lavoro alla conferenza ACM Symp, on Discrete Algorithms (SODA) Gennaio 1998**
- **A Stanford, Larry Page e Sergey Brin stanno lavorando dal 1995 su un motore di ricerca**
  - Per agosto 1998, iniziano a lavorare (nella loro dorm) alla commercializzazione del loro prodotto
  - Presentano alla WWW8 Conference il loro prodotto

La genesi

- 1998: pubblicano PageRank (che significa il rank di larry Page) e apre Google
- 1999: si ingrandisce (passa a 8 impiegati)
- 2000: 60 impiegati, team con Yahoo
  - 100 milioni di query al giorno
- 2001: Google image, 3 miliardi di documenti indicizzati

Applicazioni Google

- 2002: Google box, Adwords, Google news
- 2003: Google Deskbar, Blogger
- 2004: 4.28 miliardi di pagine, ricerche localizzate e personalizzate, Gmail, Picasa, Desktop search
- 2005: Google Maps, Google store, Google Video, Google Talk, Google Analytics
- 2006: Googlea Earth, acquista DoubleClick

Indice

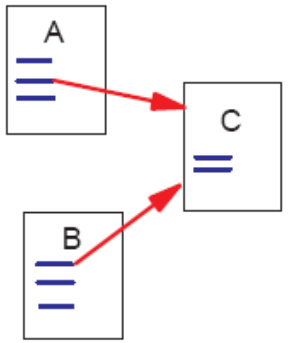
- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

L'idea del Backlink

- **Citazioni accademiche: riferimento a risultati precedenti**
  - Utilizzato per stabilire la "rilevanza" di un risultato *(impact factor)*
- **Alcune differenze:**
  - Nessuna *peer review* per pubblicare un sito
  - Possibile automatizzare la creazione di centinaia di pagine che "citano" altre pagine
  - Articoli scientifici: unitari, auto-contenentesi, dimensioni standard (poche decide di pagine)

L'idea del Backlink

- **Tecniche note ed utilizzate in Information Retrieval per analizzare e valutare le citazioni**
  - Naturale punto di partenza
- **Si può pensare al link come una citazione:**
  - Una pagina con molte citazioni è "importante"
  - Per esempio: Yahoo!



- **Alcuni problemi nell'usare solo i backlinks come misura della importanza**
  - La dimensione del web non permette (1998!☺) la certezza di avere tutti i backlink di una pagina
  - È importante che i backlink siano pesati:
    - Essere linkati da Yahoo! ha peso maggiore che essere linkati da PincoPallino

Page Rank -1

- **Sia *u* una pagina web**
- **Sia F_u l'insieme di pagine a cui *u* punta**
  - **Forward links**
  - **e sia N_u la cardinalità di F_u**
- **Sia B_u l'insieme di pagine che puntano a *u***
  - **backlinks**
- **Sia *c* un fattore per la normalizzazione**
  - **Il rango totale di tutte le pagine deve essere costante**
- **Allora una versione (semplificata) del ranking è**

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

Un esempio del ranking semplificato

- **Il rank di una pagina viene diviso tra gli archi uscenti dalla pagina, contribuendo alle pagine verso cui sono diretti**
- **Equazione ricorsiva: il rank di *u* dipende dal rank delle pagine in B_u**

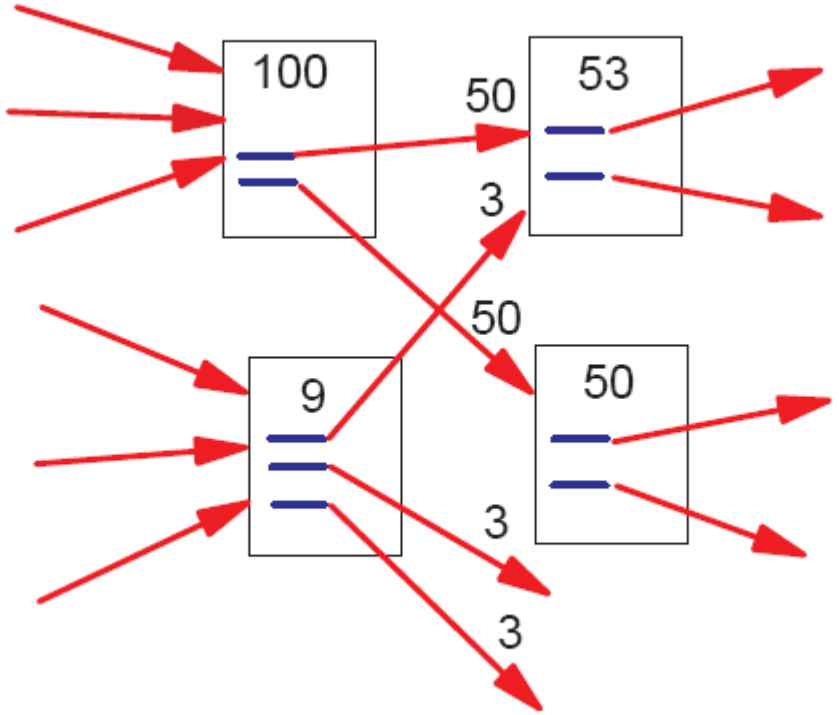


Figure 2: Simplified PageRank Calculation

Interpretazione con la algebra lineare

- **Se A è una matrice quadrata *n* x *n* (*n* pagine)**
  - **tale che l'elemento u,v di A vale: $1/N_u$ se esiste l'arco (u,v) e 0 altrimenti**
- **Se R viene considerato come un vettore R sulle pagine web**
- **Allora, da:**

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

- Deriva che **R=cAR**
- Quindi **R** è l'autovettore di **A** e **c** è l'autovalore corrispondente

Alcune critiche al ranking

- **A *rank sink*: prende tutto il ranking del resto dei nodi**

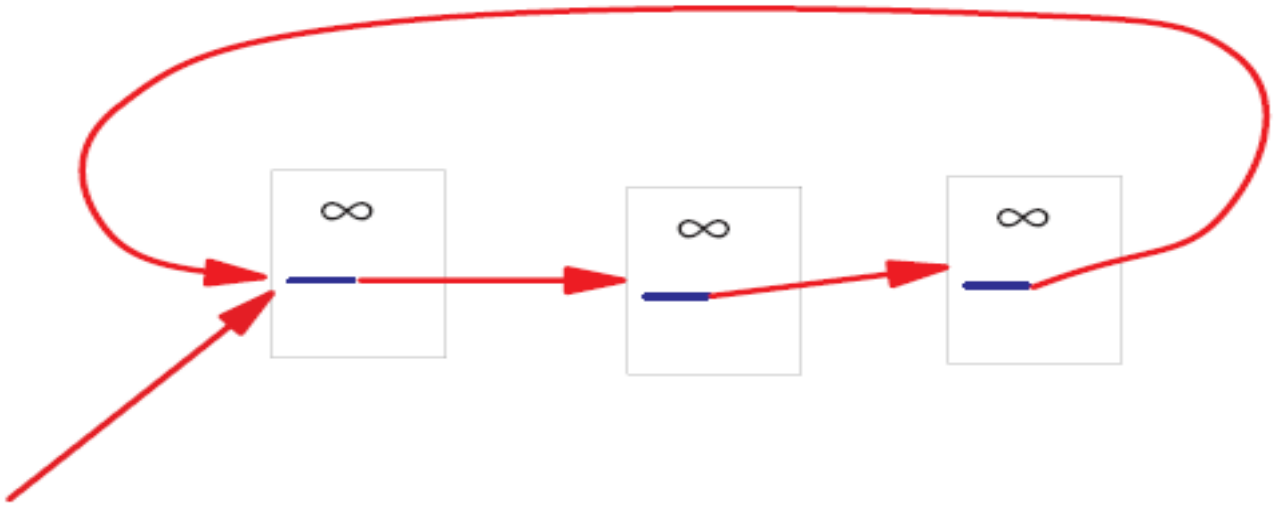


Figure 4: Loop Which Acts as a Rank Sink

- **Un *rank leak*: un nodo senza archi uscenti che prende tutto il ranking del grafo**

La definizione di PageRank

- **L'idea: evitare che una transizione abbia peso 0**
  - Infatti, in ogni momento, seppur con bassa probabilità, un utente sulla pagina A può passare anche ad una pagina B che non ha link entranti in A

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

  - Il fattore ***d*** viene detto **damping factor**
  - L'effetto è che pagine che avrebbero probabilità zero di essere selezionate, hanno una (seppur bassa) probabilità di accesso

Il random surfer

- Segue il flusso, secondo la importanza che fornisce il PageRank
- .. ma ogni tanto si "annoia" e salta a caso (con distribuzione equiprobabile) su un'altra pagina
- Alcuni possibili miglioramenti
  - Personalizzazione della pagina (delle pagine) a cui si salta in maniera periodica

Indice

- Introduzione
  - La definizione di motore di ricerca
  - L'origine del motore di ricerca
  - Come si utilizza un motore di ricerca?
- I problemi relativi alle varie query
- La nozione di authority
- HITS: Hypertext Induced Topics Search (Kleinberg)
- Google
  - PageRank
- Web spamming

WEB SPAMMING

- Azione compiute per “obbligare” i motori di ricerca a dare un ranking maggiore ad una pagina rispetto a quanto meritato
- Anche chiamato *spamdexing*
- Due principali tecniche:
  - Tecniche di boosting
  - Tecniche di *hiding* (cercano di nascondere le tecniche di *boosting*)

Boosting: term spamming

- In ogni motore di ricerca, viene valutata la presenza dei termini della query nelle pagine
- A seconda del *text field* diverso peso
  - Text field: titolo (TITLE), metadati, URL, body, etc.
- Una metrica utilizzata nell'information retrieval
  - Term Frequency Inverse Document Frequency **(TFIDF)**
- Dato un termine t, **TF(t)** è la frequenza con cui il termine appare nel text field
- Dato un termine t, **IDF(t)** è l'inverso della frequenza del termine in tutti i documenti che compongono la collezione

TF IDF e lo spam

- Data una pagina p e una query q,

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

- La logica: un termine che è frequente in un documento (TF alto) ma poco frequente in tutti gli altri documenti (IDF alto) è importante …
  - altrimenti no
- Gli spammer hanno accesso solo al TF (non avendo controllo sulla intera collezione)
  - quindi la loro tecnica è di ripetere il termine all'interno dei text field della pagina

Diversi tipi di term spamming

- Body spam
- Title spam
- Meta tag spam (a volte ignorato dai motori di ricerca)
- Anchor text spam: assume peso maggiore solo da parte dei motori di ricerca
- URL spam

Una altra caratterizzazione di term spamming

- Basata sul tipo di termini aggiunti
  - **Repetition**: ripetizione di termini, in modo da aumentare la rilevanza per un ristretto numero di query term
  - **Dumping**: si scaricano interi dizionari, in modo da rendere la pagina rilevante per diversi term query
  - **Weaving**: duplicazione di articoli interi, con termini di spam inseriti in posizioni a caso
    - Esempio: 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
  - **Phrase stitching**: attaccare frasi di diversi documenti insieme in modo da aumentare la possibilità che sia rilevante per diverse query term

Link spamming

- Definizioni**:**
  - **Inaccessible Page**: Pagine inaccessibili agli spammer
  - **Accessible Page**: Pagine accessibili agli spammer (modificabili in maniera limitata dagli spammer: es. blog, comments, etc.)
    - **Limitate in numero m in quanto costoso l'accesso e manuale l'intervento dello spammer**
  - **Own Page**: Pagine proprie degli spammer (completamente sotto controllo), dette anche spam farm
    - Limitate in numero n
- **Obiettivo dello spammer: aumentare il ranking di una pagina target t**

Link spamming su HITS

- **Hub e authorities**
- **Hub: facile da "spammare"**
  - Basta diventare un buon hub
    - **Inserendo link a buone authority, come CNN, MIT, etc.**
- **A questo punto il gioco è fatto:**
  - Ho un buon hub
  - Lo faccio puntare a molte pagine alle pagine dello spammer
  - E poi da queste pagine aggiungere link alla pagina t

Link spamming su PageRank

- **Una interpretazione del PR di una pagina (o un insieme di pagine) S:**

$$PR(S)=PR_{static}(S)+PR_{in}(S)-PR_{out}(S)-PR_{\sin k}(S)$$

- **Le 4 componenti**
- **PRstatic(S):** la parte relativa al contributo statico (jump random)
- **PRin(S)**: il ranking ricevuto da link che entrano in S
- **PRout(S)**: il ranking donato verso l'esterno
- **PRsink(S)**: il ranking dissipato all'interno dalle pagine in S che non hanno archi uscenti

Una strategia per spamming su PR

- **Si usa un architettura che rende massima PR(t)**

$$PR(t)=PR_{static}(t)+PR_{in}(t)-PR_{out}(t)-PR_{\sin k}(t)$$

- **Proprietà**
  - **Tutte le pagine proprie sono raggiungibili da quelle accessibili**
    - Un crawler le trova e le cataloga
  - **Usa un numero minimale di link**

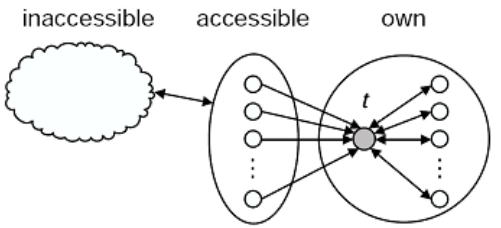


Figure 2: An optimal link structure for PageRank.

- **In particolare**
  - Tutte le n pagine proprie sono parte della farm, massimizzando static
  - Tutte le m pagine accessibili puntano alla farm, massimizzando in
  - Nessun link in uscita, out a zero
  - Nessuna pagina senza link uscenti, sink a 0

Tecniche per manipolare link

- **Link in uscita**: Utilizzare intere sezioni delle directory presenti su Web (DMOZ, Yahoo!, etc.) in modo da avere massicce strutture di link validi in uscita
- **Link in entrata:**
  - creare una *honey pot* pagine che hanno risorse utili ma anche alcuni link nascosti alle pagine target (spesso anche utilizzato con le directory "copiate")
  - infiltrare una web directory (inserendo propri link)
  - post link, blog, guestbook, wiki etc.
  - scambiare link con altri spammer
  - comprare domini scaduti

Spam Hiding

- **Serve per nascondere i segni evidenti del boosting**
- **Alcune tecniche comuni:**

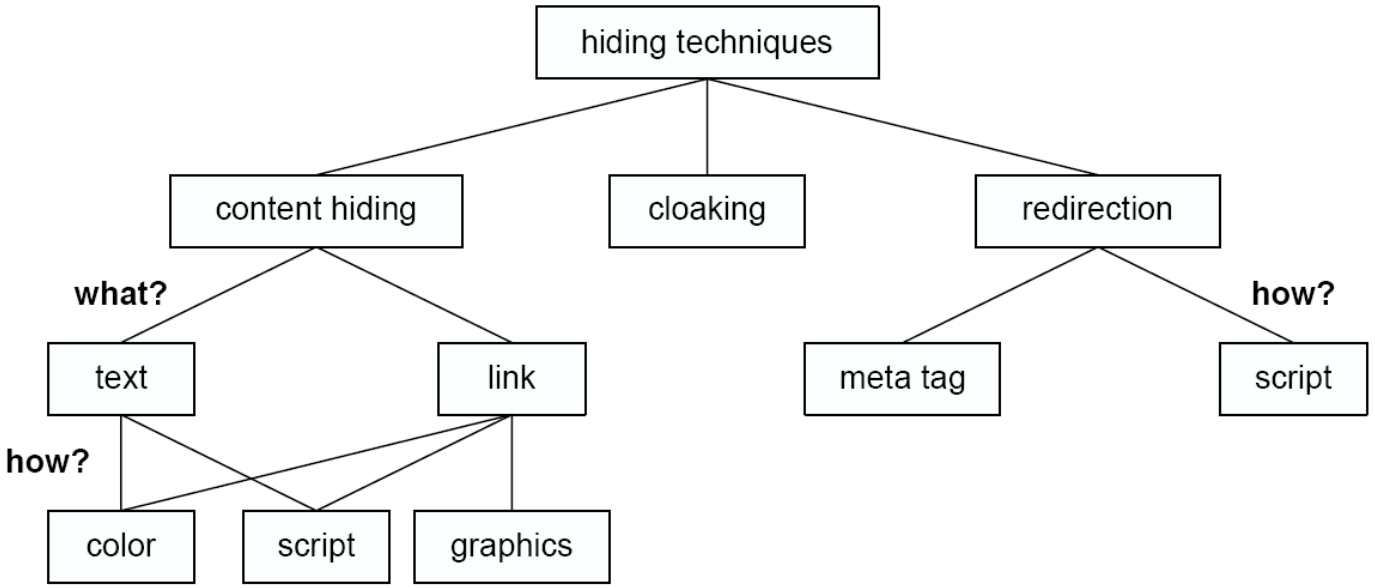
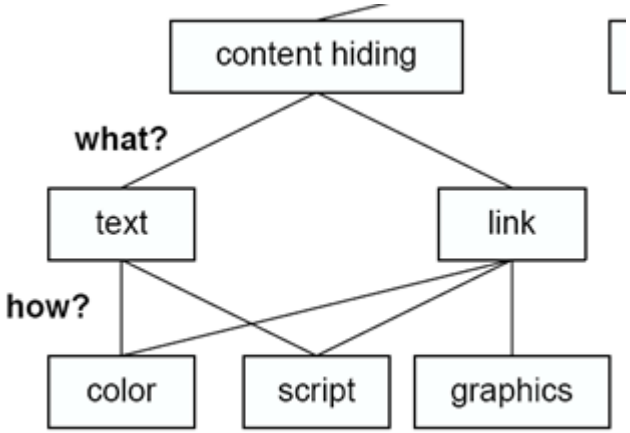


Figure 3: Spam hiding techniques.

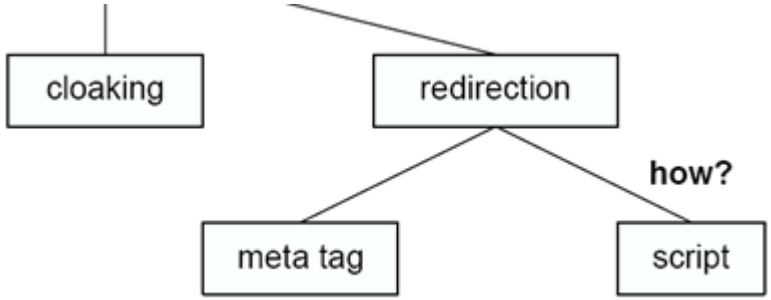
Content Hiding

- **Scrivere testo con colore uguale al background**
- **Immagini di 1 pixel**
  - che sono trasparenti o dello stesso colore
  - che contengono URL spam
- **Script (settando visible a false)**



Cloaking e Redirection

- **Cloaking**: Server che restituiscono pagine diverse a seconda se è un "regolare" browser oppure un crawler
  - attraverso il campo User-agent
- **Redirection**: una URL, viene rediretta ad un'altra pagina, ma il contenuto sulla prima pagina viene comunque usato dal search engine (*doorways*)
  - refresh tag in meta field del documento HTML
  - scripting Javascript



FINE

UNIVERSITÀ DEGLI STUDI DI SALERNO

Scala Santolo0521/000268