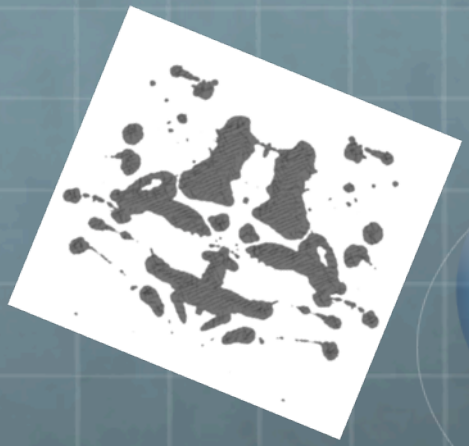
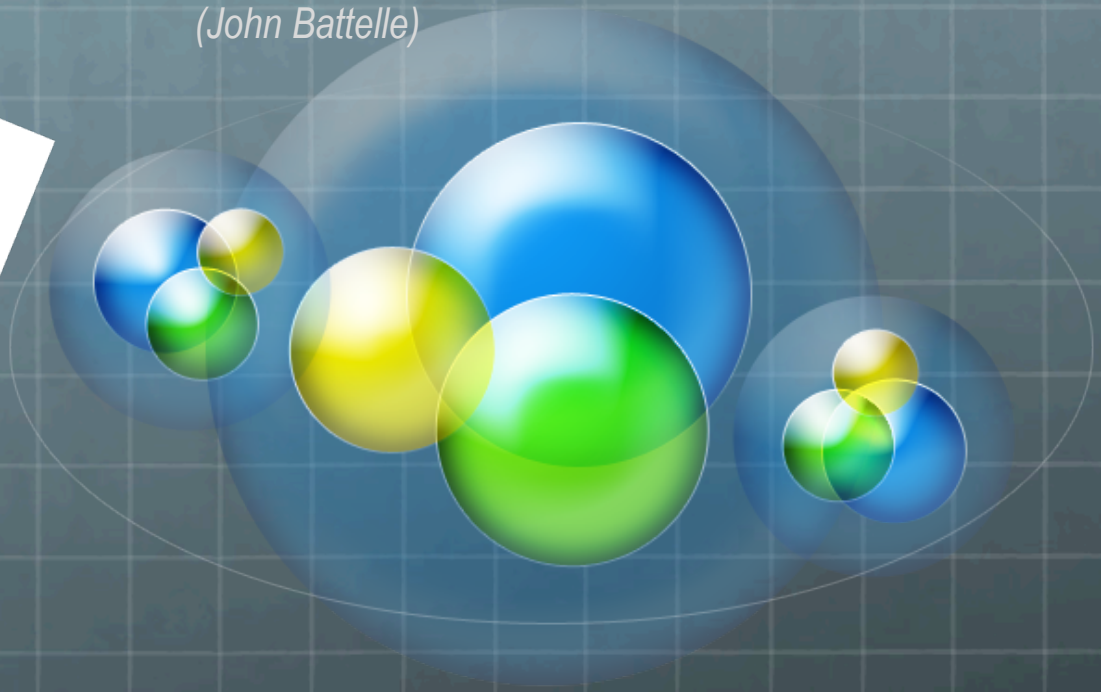
«Google è un globale test di Rorschach. Vediamo in esso ciò che vogliamo vedere.»

(John Battelle)





TECNOLOGIE INFORMATICHE MULTIMEDIALI

Corso di Laurea "*Scienze e Tecnologie della Comunicazione*"

Prof. Giorgio Poletti (giorgio.poletti@unife.it)

a.a. 2013-2014

Sviluppo della lezione

Contenuti

- Algoritmi di ricerca delle informazioni

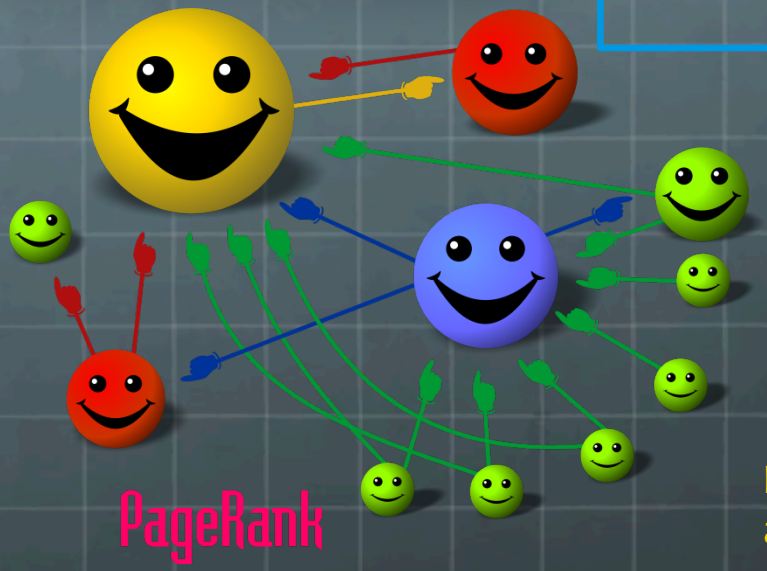
Attività

- Analisi dell'algoritmo di page rank

Algoritmi di Ricerca delle Informazioni

ALGORITMO DI ANALISI

Assegnazione di un peso ad ogni elemento di collegamento ipertestuale di un insieme di documenti



PageRank è un marchio Google

Algoritmo brevettato dall'università di Stanford

PR(E)= PageRank di E è il valore, peso numerico, associato ad un elemento E

- Una pagina è tanto più importante quanto più numerose sono le pagine che la puntano

Nomen omen…

Algoritmo ideato, tra gli altri da LAWRENCE EDWARD "LARRY" PAGE fondatore, con SERGEY BRIN, di Google.

Algoritmo PageRank

RETE WEB

struttura

Grafo
DIANMICO

componenti

Pagine e link

NODI: PAGINE

ARCHI: LINK IPERTESTUALI

componenti

DOCUMENTI : oltre **$4*10^9$** *(4 miliardi, senza contare i nodi isolati, pagine senza link)*

LINK : oltre **$6*10^9$** *(6 miliardi)*

SERVER DI RETE: *oltre* **$2*10^6$** *(200 milioni)*

OLTRE 8*106 (800 MILIONI DI UTENTI)

interagiscono

Algoritmo PageRank

GRAFO SEMPLIFICATO PER IL WEB

finito

Grafo **G(P,L)**, insieme di pagine e link

Se orientato

Relazione ↔ $p_1 \leftrightarrow p_2$ se

- ESISTE UN CAMMINO DA p_1 A p_2
- e
- ESISTE UN CAMMINO DA p_2 A p_1

è

RELAZIONE DI EQUIVALENZA

composto da

Classi

dette

COMPONENTI (FORTEMENTE) CONNESSE DEL GRAFO

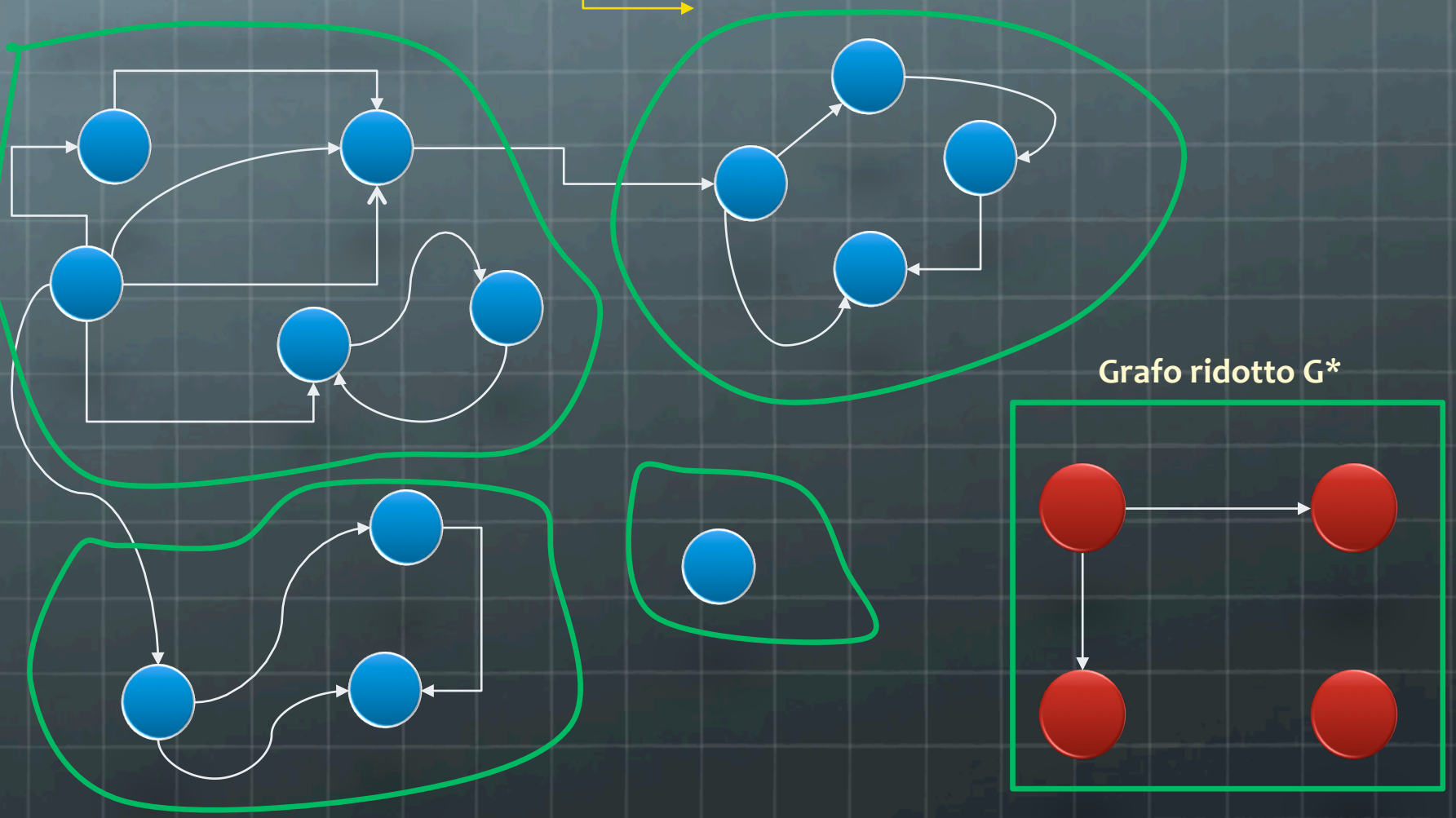
DEFINIRE UN GRAFO RIDOTTO G*, *i nodi sono le classi e due classi C_1 e C_2 sono connesse se esiste un nodo in C1 collegato a un nodo in C2, esiste un arco da C_1 a C_2*

Permette di

Algoritmo PageRank

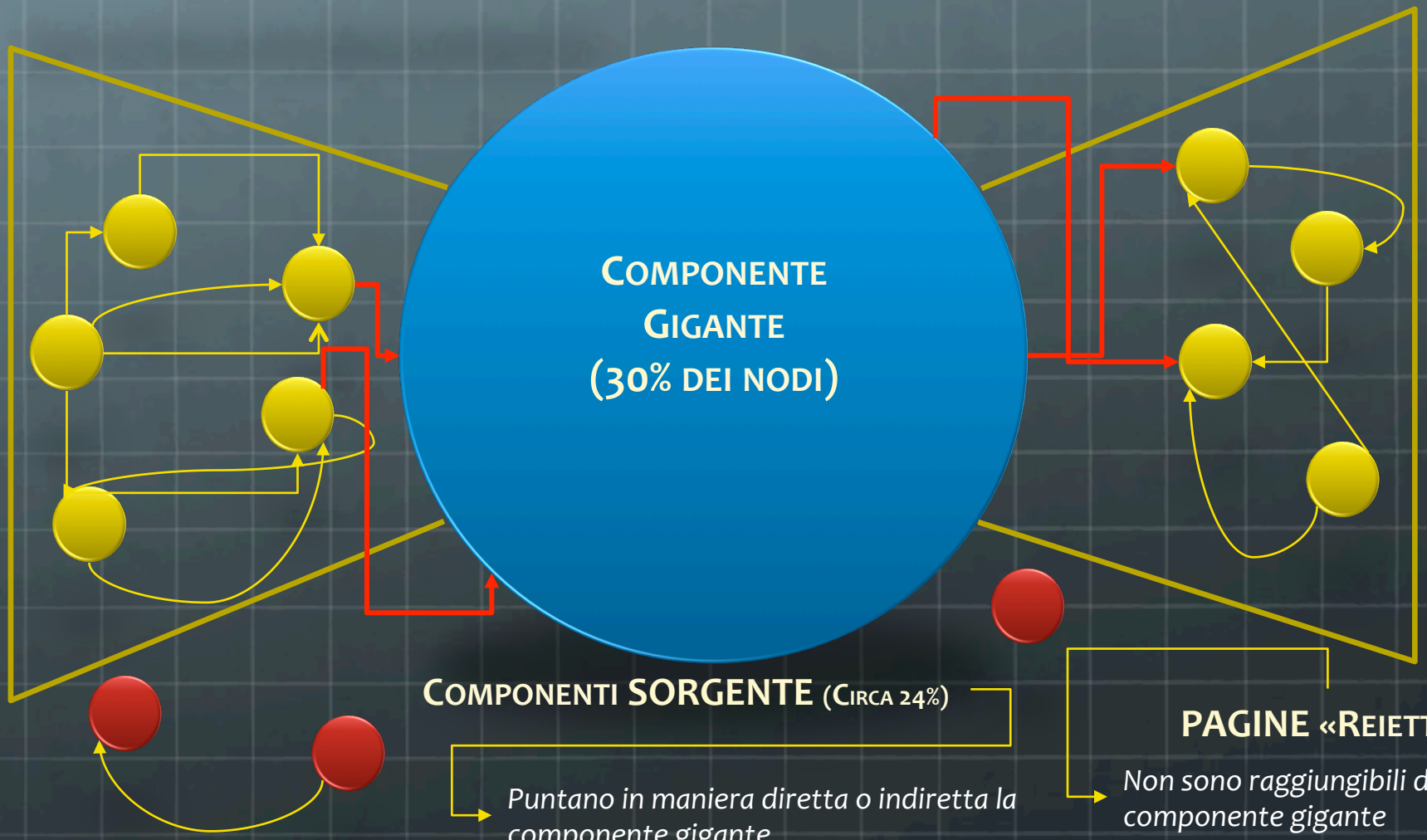
GRAFO RIDOTTO G*

esempio



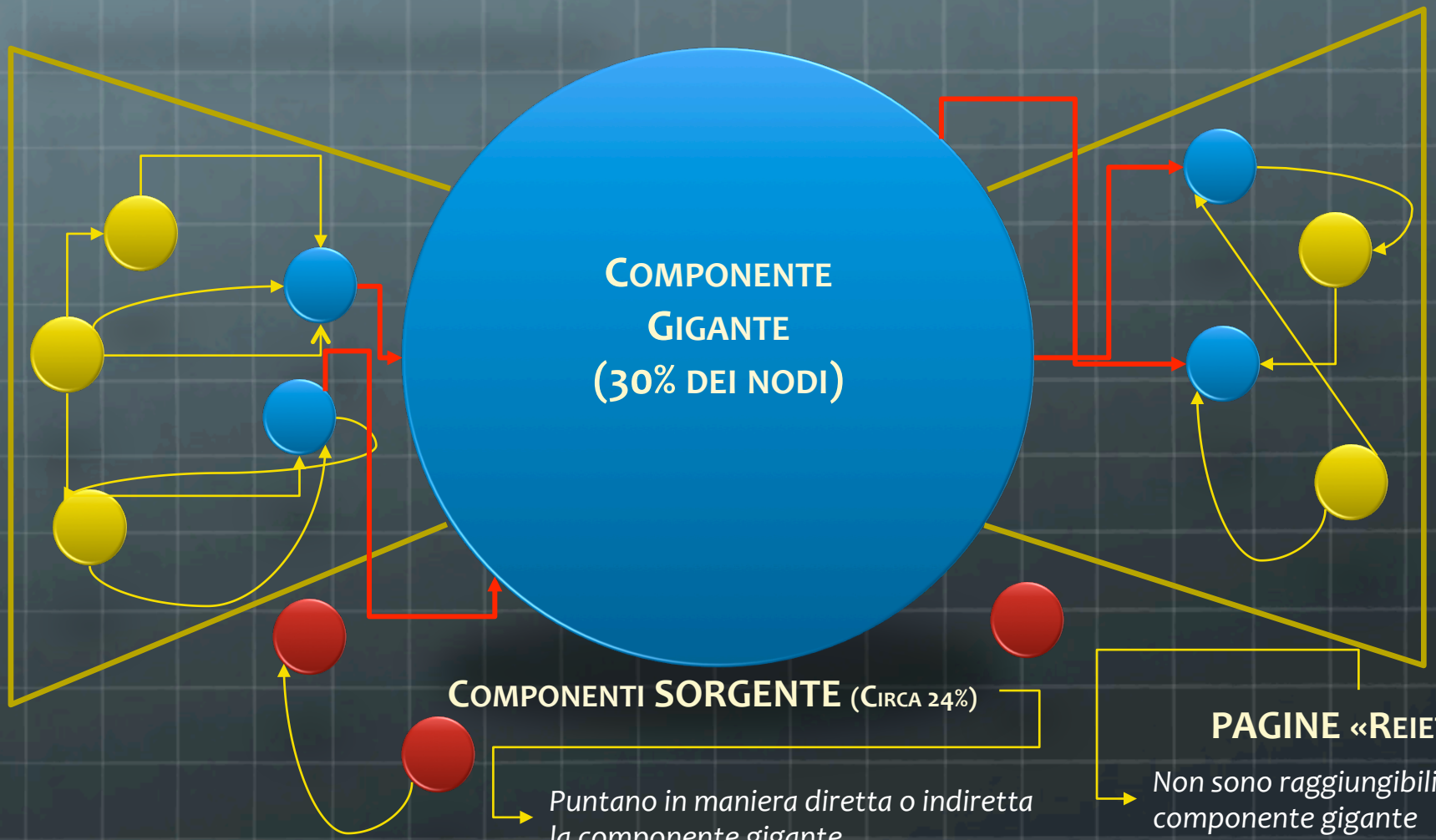
Algoritmo PageRank

Struttura a Caramella del Web



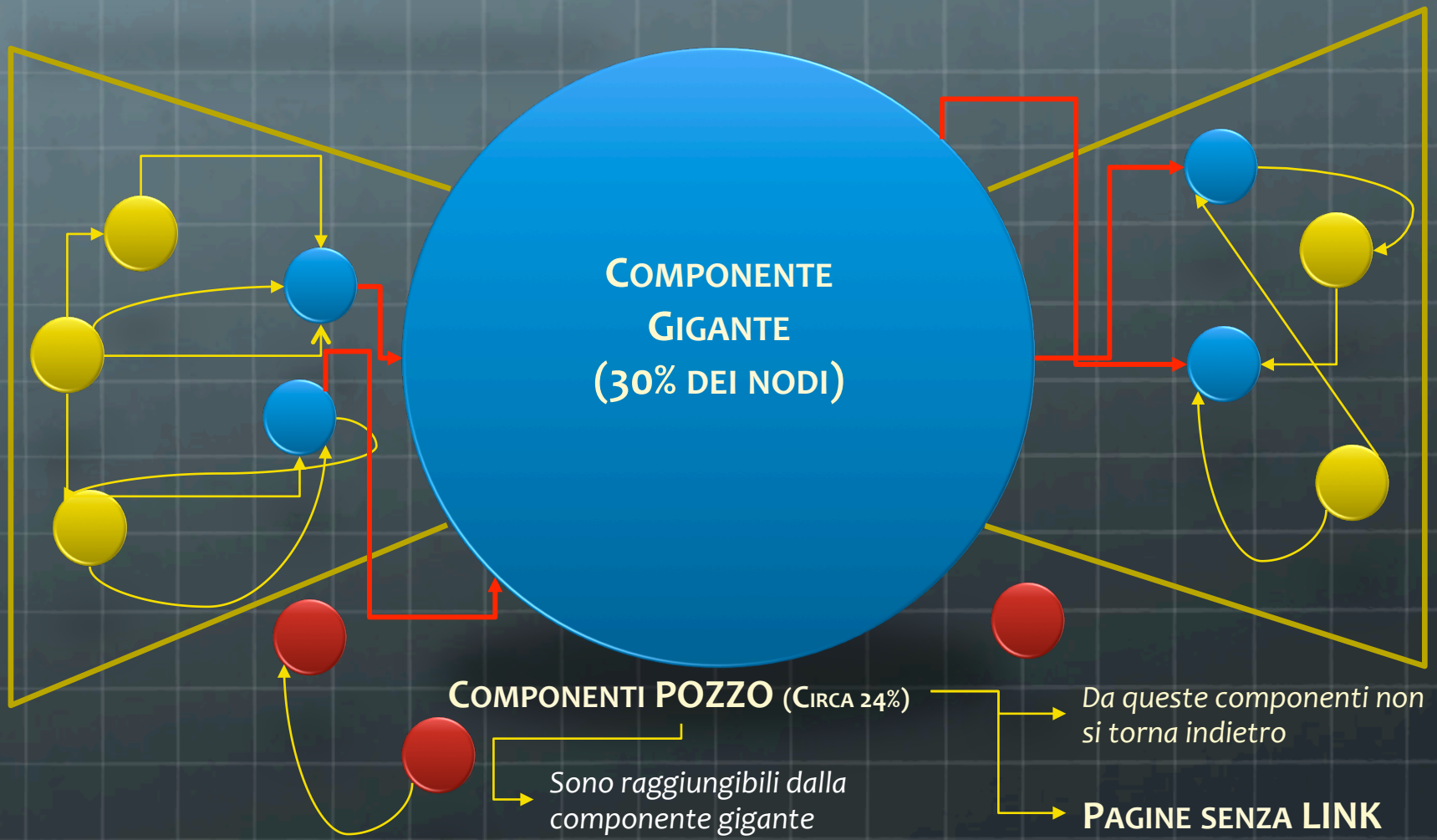
Algoritmo PageRank

Struttura a Caramella del Web



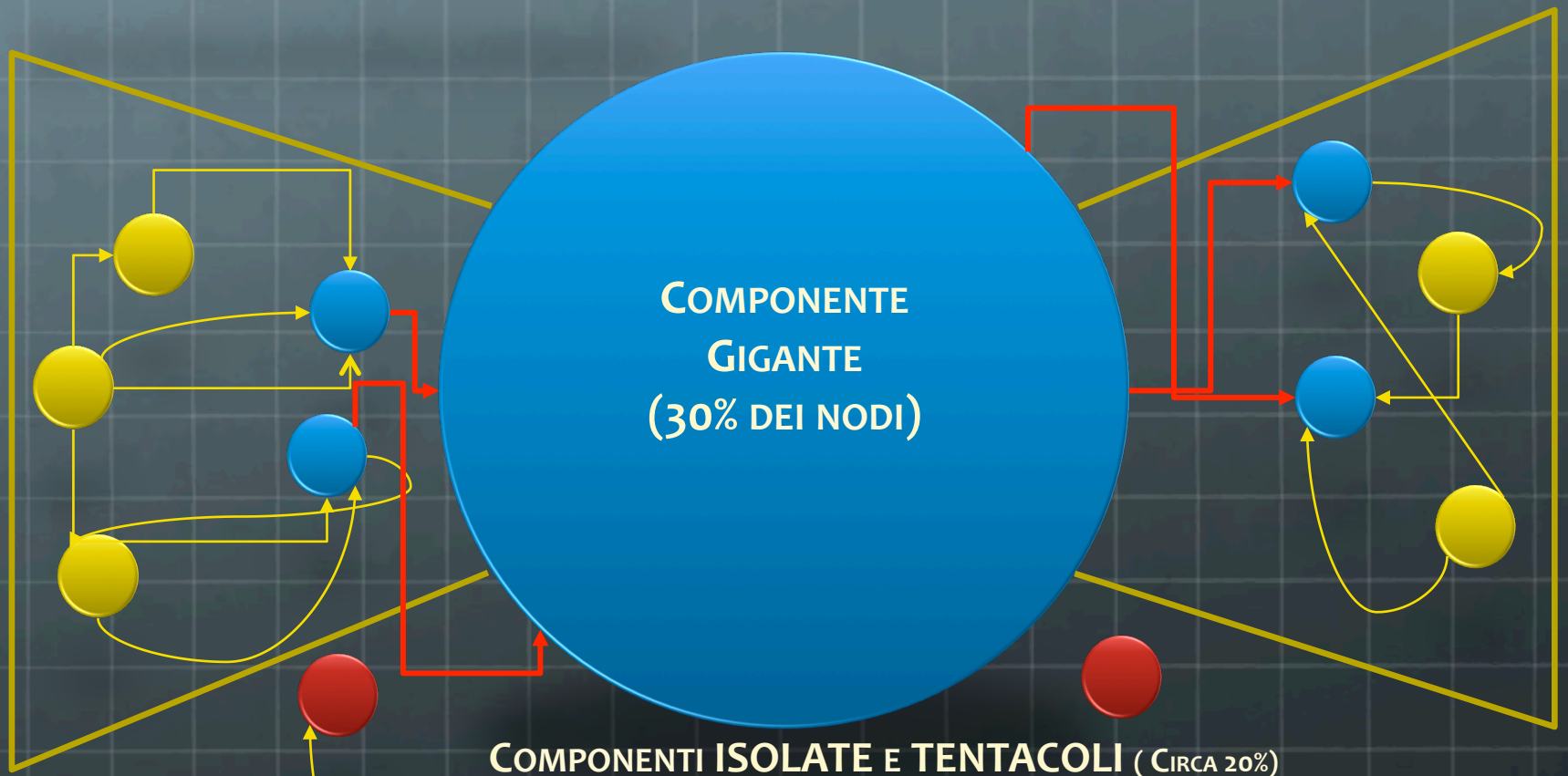
Algoritmo PageRank

Struttura a Caramella del Web



Algoritmo PageRank

Struttura a Caramella del Web



Componenti ISOLATE e TENTACOLI (Circa 20%)

Tentacoli *link per sorgenti e pozzi che non passano per la componente gigante*

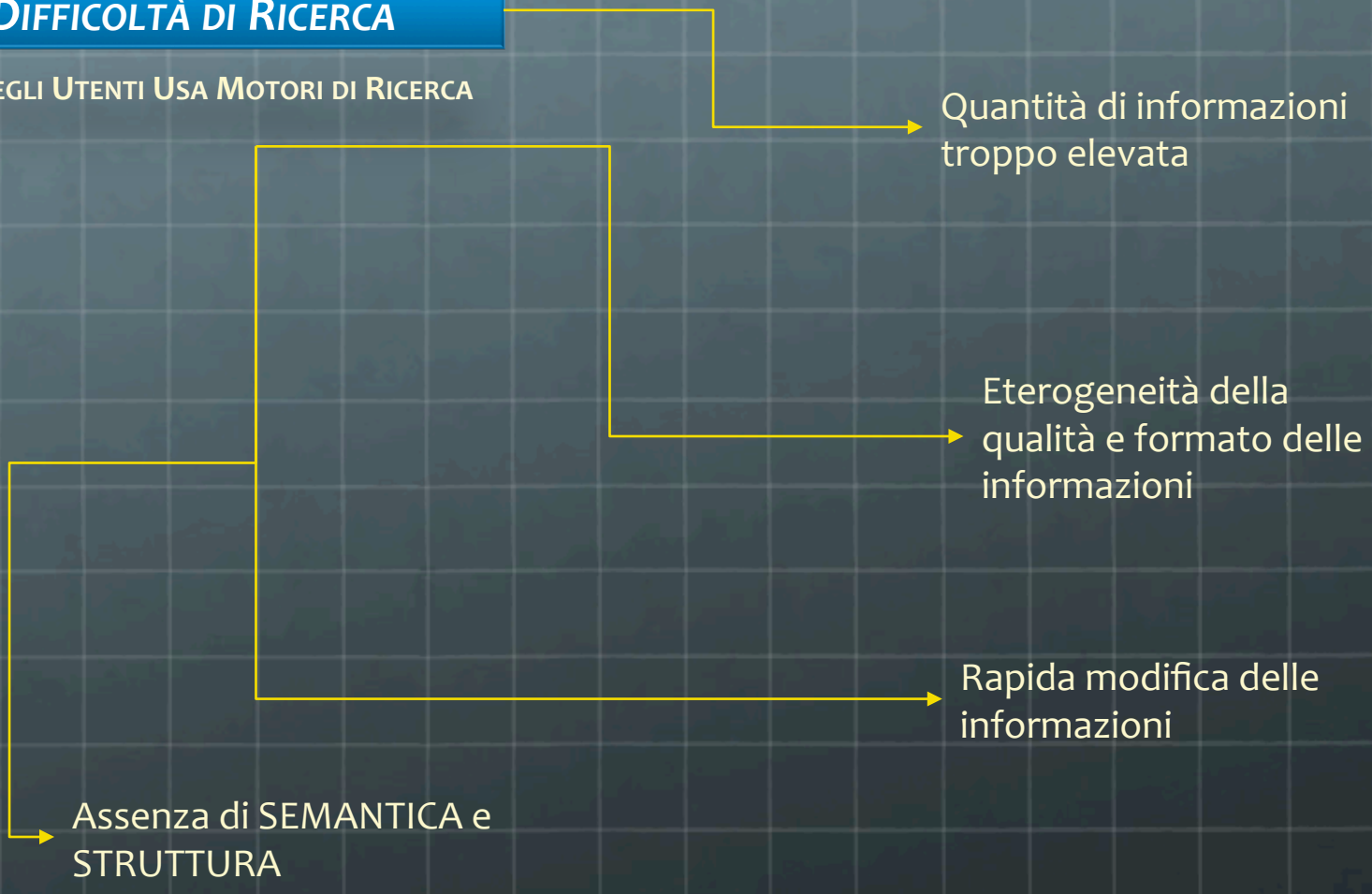
NON sono raggiungibili dalla componente gigante

Da queste componenti non si raggiunge la componente gigante

Algoritmo PageRank

DIFFICOLTÀ DI RICERCA

80% DEGLI UTENTI USA MOTORI DI RICERCA



Algoritmo PageRank

Motori di Ricerca

Azioni dei Motori di Ricerca

- Raccolta dati
- Elaborazione e catalogazione dei dati raccolti
- Elaborazione e risposta alle interrogazioni (Query) degli utenti

Algoritmo PageRank

MOTORI DI RICERCA

RACCOLTA DATI

Raccolta del contenuto delle pagine Web (informazioni di tipo testuale ma anche immagini (google immagini e documentazione)

robots.txt

Si usa uno **SPIDER O CRAWLER O ROBOT**

SPIDER	MOTORE DI RICERCA
googlebot	Google
fast	Fast - Alltheweb
scooter	Altavista
mercator	Altavista
Ask Jeeves	Ask Jeeves
teoma_agent	Teoma
ia_archiver Yahoo!	Alexa - Internet Archive
Slurp	Yahoo
Romilda	Facebook

Simulatore di Spider

PROBLEMI

Spider traps

Aggiornamento frequente delle pagine

Quantità di dati e larghezza di banda

Pagine nascoste (*pagine isolate della struttura a caramella*)

Mancanza di standard condivisi e rispettati

Algoritmo PageRank

Motori di Ricerca

Elaborazione e catalogazione dei dati raccolti

Parsing, (analisi: estrazione di informazioni)

Rilevazione delle ridondanze (presenza di MIRRORING)

Rilevazione di presenza di SPAMMING

Indicizzazione dei dati

In sketch comico del **Monty Python's Flying Circus** che ha come luogo un locale nel quale **ogni** pietanza proposta dalla cameriera era a base di Spam (*un tipo di carne in scatola*).

Reperimento e analisi delle informazioni per il calcolo del **RANKING**

Algoritmo PageRank

Motori di Ricerca

Elaborazione e risposta alle interrogazioni (Query) degli utenti

Ricerche testuali raffinate **AND** (*Pozzo* **NEAR** *Pizza*),il forse cercavi…

Suggerimenti ontologici

Ontologia fondamentale o primitiva per ha come obiettivo quello di descrivere "ciò che esiste" secondo un insieme di entità ritenuto non ulteriormente definibile (vocabolari)

Sistemi di catalogazione automatica

Analisi linguistiche (frequenze, relazioni…)

Analisi dei profili utente (ad esempio i bookmarks)

I contenuti indicizzati sono classificati rispetto ai concetti definiti all'interno dell'ontologia

ALGORITMO PAGERANK

MOTORI DI RICERCA

ELABORAZIONE DEI DATI

obiettivo

→ INDICIZZAZIONE DEI DOCUMENTI RACCOLTI

esigenze

- INDICIZZAZIONE → RENDE EFFICIENTE E VELOCE LA RISPOSTA ALLE QUERY
- INDICIZZAZIONE → RENDE POSSIBILE IL **RANKING** DEI DOCUMENTI

Algoritmo PageRank

RANKING

RANKING definizione

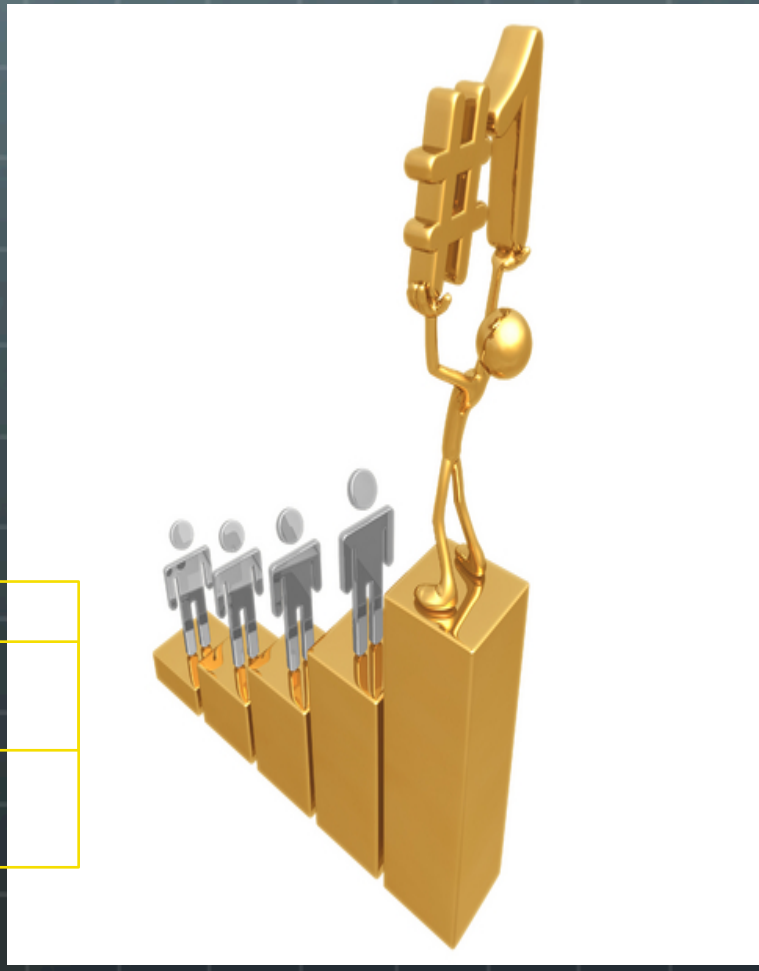
DATO UN INSIEME DI PAGINE **P** E UNA QUERY **Q** IL RANKING È DEFINITO DA UNA FUNZIONE:

$R_Q: P \rightarrow \mathbf{R}$ (*INSIEME DEI NUMERI REALI*)

CHE ASSOCIA AD OGNI PAGINA UN NUMERO REALE CHE INDICA LA ***«RILEVANZA»*** *DI QUELLA PAGINA NEL CONTESTO DI QUELLA QUERY.*

tecniche

- Analisi del contenuto testuale (ALTAVISTA)
- Analisi della struttura dei link (GOOGLE)

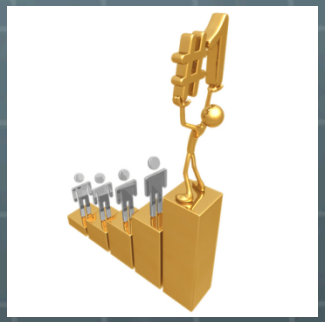


Algoritmo PageRank

RANKING

RANKING — Tecniche

LATENT SEMANTIC INDEX
Analisi del contenuto testuale
(ALTAVISTA)



procedimento

Roma	Pioggia

t=2

t numero di termini presi in considerazione, appartenenti ad un vocabolario o individuati durante la raccolta delle pagine

Ad ogni pagina **P** è associato un vettore

con

$(d_P)_j$ = ***numero di occorrenze del termine j in P***

Oggi la pioggia è stata abbondante in tutta Italia.
Roma con la pioggia è piacevole da visitare

Pagina P

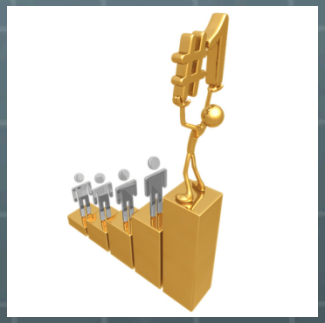
$(d_P)_1 = 1$ numero di occorrenze di **Roma** in **P**
$(d_P)_2 = 2$ numero di occorrenze di **Pioggia** in **P**

Algoritmo PageRank

RANKING

RANKING — Tecniche

LATENT SEMANTIC INDEX
Analisi del contenuto testuale
(ALTAVISTA)



procedimento

Ad ogni query **Q** è associato un vettore

con

$(d_P)_j$ = 1 se il termine j compare in P
$(d_P)_j$ = 0 se il termine j non compare in P

Roma	Pioggia

Oggi la pioggia è stata abbondante in tutta Italia.
Roma con la pioggia è piacevole da visitare

Pagina P

Pagine su Roma

Query Q

$(d_P)_1$ = 1 **Roma** è nella query **Q**
$(d_P)_2$ = 0 Pioggia non è nella query **Q**

Algoritmo PageRank

RANKING

RANKING — Tecniche

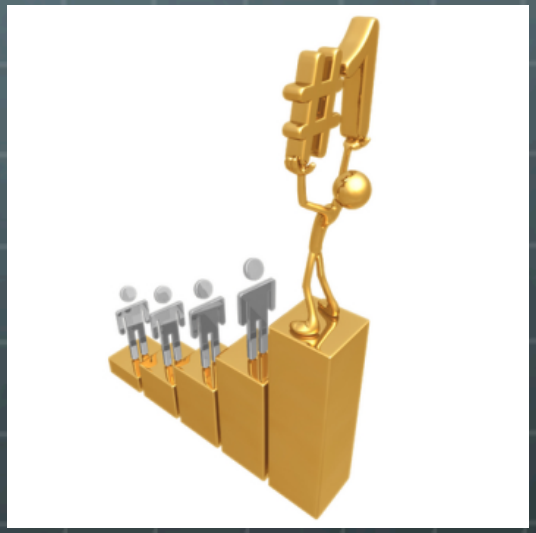
LATENT SEMANTIC INDEX (LSI)
Analisi del contenuto testuale
(ALTAVISTA)

Assunto INGENUO

Pagina autorevole se contiene frequenti termini concernenti l'argomento

Migliori prestazioni

Estensione del testo con i testi delle pagine collegate e i termini del contesto



$$\cos d_P d_Q = \frac{\hat{d}_P \times \hat{d}_Q}{\|d_P\| \cdot \|d_Q\|}$$

CONSIDERAZIONE: *LSI funziona bene su query multiple, le normali query sono semplici, 2 o 3 elementi al massimo.*

Algoritmo PageRank

RANKING

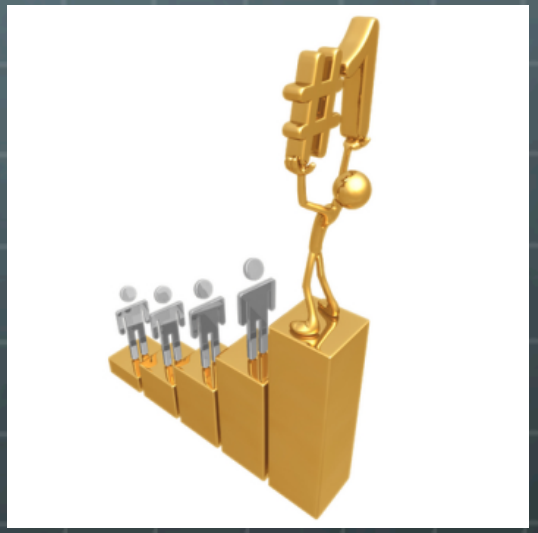
RANKING — Tecniche →

PAGERANK
Analisi dei link (GOOGLE)

procedimento →

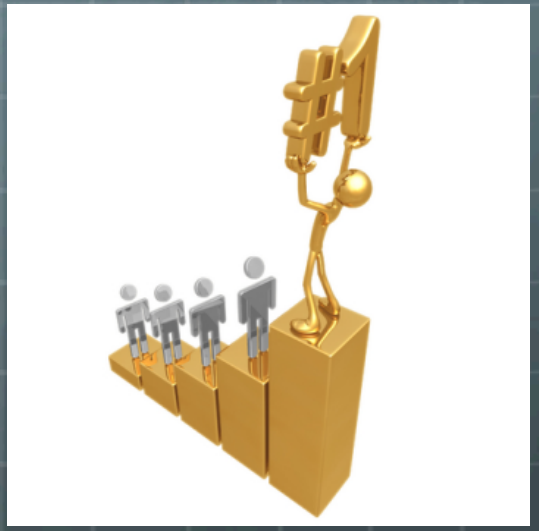
Ad ogni pagina j viene assegnato un valore reale, un **rank R_j** statico, indipendente cioè dalla query.

Data la query Q si ordinano i risultati in base al rank delle pagine individuate



L'importanza delle pagine è determinata **ESCLUSIVAMENTE** in relazione ai link che presenta o di cui è target.
*L'assunto è che il contenuto **NON è AUTODESCRITTIVO** e l'importanza di una pagina è il risultato di un processo **ESOGENO** (esterno all'ambito di riferimento)*

Algoritmo PageRank



RANKING

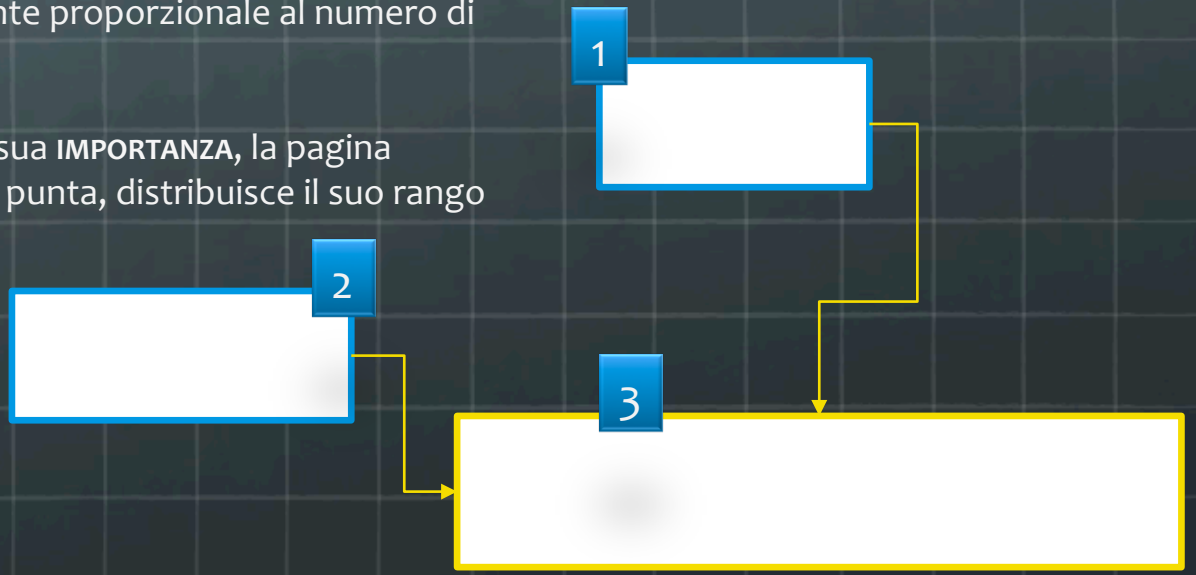
RANKING — Tecniche

PAGERANK
Analisi dei link (GOOGLE)

principio

L'importanza di una pagina è direttamente proporzionale al numero di pagine che la puntano

Se R_j è il **rank** (*RANGO*) di una pagina, la sua **IMPORTANZA**, la pagina trasmette l'importanza alle pagine che punta, distribuisce il suo rango in maniera uniforme



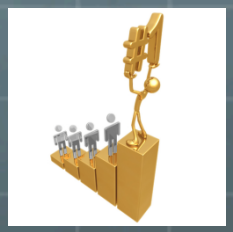
Algoritmo PageRank

RANKING

Principio ispiratore

STOCASTICO (greco stochastikós, "congetturale", derivato di stocházesthai, "fare congetture«)

RANKING Tecniche



PAGERANK
Analisi dei link (GOOGLE)

problema

La funzione ha un risultato unico solo se il grafo è connesso

soluzione

Si introduce un fattore che equivale a inserire link random (casuali) al grafo

d fattore di **dumping** (deciso da Google), fattore di spargimento, passa da una pagina all'altra ed é valore di PageRank minimo attribuito ad ogni pagina in archivio.

Nella documentazione originale d=0,85

PROCESSO STOCASTICO una famiglia di variabili aleatorie dipendenti dal tempo, definite su un unico spazio campione finito e che assumono valori in un insieme definito spazio degli stati del processo

Processo **STOCASTICO MARKOVIANO** o (processo di Markov): *processo stocastico nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende unicamente dallo stato del sistema immediatamente precedente (PROPRIETÀ DI MARKOV) e non dal come si è giunti a tale stato.*

DEFINIZIONE INTUITIVA: *un processo stocastico é un insieme ordinato di variabili casuali, indicizzate dal parametro t, spesso detto tempo. (Quantità di pioggia)*

Algoritmo PageRank

RANKING

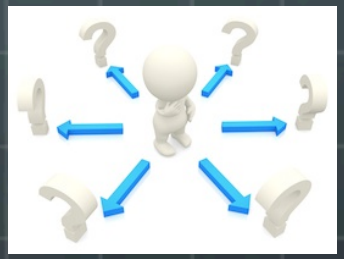
- Caffeine, algoritmo 2010 (10 giugno)
- Hummingbeard, algoritmo 2013 (27 settembre)

VANTAGGI

- *valore del* **RANK** *calcolato in maniera accurata*
- *processo iterativo converge molto rapidamente (tempi rapidi di esecuzione)*
- **RANK** *calcolato indipendentemente dalle query*

SVANTAGGI

Googlebombing



- *si possono pubblicare insiemi di pagine «TRAPPOLA» «ARTEFATTE» che influiscono sul ranking delle pagine*
- *è un limite l'indipendenza dalla query*

L'algoritmo di PageRank può essere corretto usando aggiustato usando un secondo ranking basato sul contenuto, come ad esempio **LSI (LATENT SEMANTIC INDEX).**