Sommario

[Pagerank: Concetti introduttivi 1](#_Toc480659722)

[Il problema di PageRank 2](#_Toc480659723)

# Pagerank: Concetti introduttivi

Nel motore di ricerca di google, **Pagerank** svolge un ruolo fondamentale nel valutare (rank) i risultati delle nostre ricerche. L’algoritmo quantifica l’importanza di ogni pagina web basandosi sulla struttura dei link presenti nella pagina. Il paper che stiamo studiando introduce l’argomento presentando il problema originale per poi spostarsi alla proposta di diversi algoritmi distribuiti randomizzati tutti volti a calcolare il PageRank.

Gli algoritmi di ricerca sono diventati strumenti indispensabili nel web in quanto è importante non solo fornire risultati attinenti alla ricerca, ma anche valutare le pagine in maniera appropriata in modo da facilitare l’accesso alle informazioni. L’idea di base di pagerank è che ogni pagina può essere immaginata come un elettore che esprime il proprio voto utilizzando i link in essa presenti. I voti però non hanno tutti lo stesso peso, infatti le pagine web più popolari esprimono con i propri link voti di peso maggiore. Questo problema si traduce matematicamente nel trovare l’autovettore corrispondente al più grande autovalore di una matrice stocastica associata alla struttura web. Attualmente, il web è vastissimo, infatti si parla di oltre 8 miliardi di pagine ed attualmente il calcolo del pagerank è eseguito centralmente da google che utilizza i propri crawler o spider per esplorare le pagine. Il problema principale di tale metodo è che impiega molto tempo, si stima infatti che è richiesta circa una settimana per finire. Il paper che stiamo studiando invece propone un metodo distribuito che ha 3 caratteristiche fondamentali

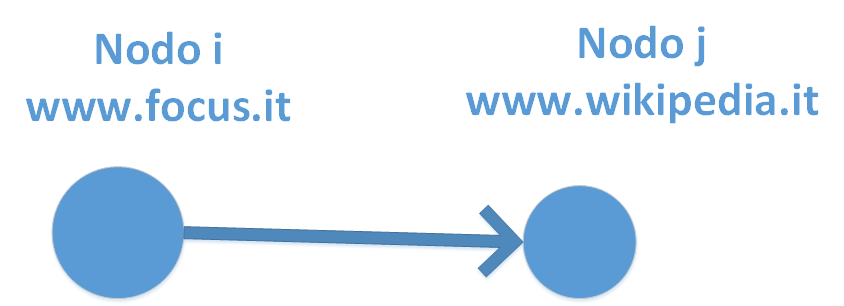
* Ogni pagina calcola automaticamente il proprio PageRank comunicando con le pagine che sono connesse ad essa tramite link diretti
* Le pagine iniziando la comunicazione con le altre in istanti di tempo random, senza dover usare un agente centralizzato
* Il carico computazionale richiesto ad ogni pagina è molto basso.

La soluzione proposta nel paper si basa sul raggiungimento del consenso tra gli agent (le pagine web), ovvero che gli agent raggiungano lo stesso valore in un tempo finito. Questa soluzione ovviamente si discosta dal metodo originale che si basa sul calcolo del determinato autovettore di una matrice stocastica utilizzando il **metodo delle potenze (Algoritmo iterativo che mi trova autovettore ed autovalore dominanti)**

Concludiamo il paragrafo introduttivo chiarendo alcune notazioni ampiamente usate nel seguito. Date due matrici X, Y ϵscriviamo che X ≤ Y se xij ≤ yij per i = 1…n e j = 1…m. Diciamo inoltre che un vettore di probabilità è un vettore non negativo tale che ij=1 per j=1….n. Indicheremo con S la matrice con tutti gli elementi 1.

# Il problema di PageRank

Consideriamo una network di n pagine ognuna rappresentata da un numero intero incrementale. Graficamente questa rete può essere vista come un **digrafo** *G= (ν,ε)* in cui *ν = {1,2…n}* è il set dei nodi, mentre è il set degli edges tra le pagine che rappresentano i link tra esse. Diciamo che il vertice i è collegato al vertice j tramite un edge se la pagina i ha un link verso la pagina j.



L’obiettivo di pagerank è dare un valore ad ogni pagina. Tale valore viene indicato con *xi\** e può assumere valori in [0,1]. Quindi diremo che una pagina i è più importante della pagina j se *xi\** > xj\*

Questo valore è calcolato per ogni pagina come somma dei contributi di tutte le pagine che hanno un link ad essa.

In cui Li è il set degli IN-Neighbor di i, mentre nj è il numero totale di link in uscita dalla pagina j. Se rappresentiamo il problema in forma vettoriale, cioè

*x\* = [x\*1,…,x\*n]T* Allora il problema di PageRank può essere scritto come segue:

*x\* = Ax\**in cui A = (aij) ϵ Rnxn è la matrice delle adiacenze chiamata **link matrix** e definita come segue:

Osserviamo che per come è stato definito, x\* è un autovettore di A con autovalore 1. Affinchè questo autovettore esista e sia unico è necessario che il web rappresentato come grafo sia **strongly connected**, ma noi sappiamo già che il web non potrà mai esserlo. Nel web esistono i nodi orfani che sono in pratica nodi che non hanno link verso l’esterno, come potrebbero ad esempio essere i tanti file caricati sul web. Per questi nodi introduciamo dei link artificiali, cioè semplicemente introduciamo in questi nodi dei link alle pagine che hanno un link verso quel nodo orfano. In questo modo la matrice A diventa stocastica, cioè per ogni j. Questo comporta che esisterà almeno un autovalore pari ad 1.Infine per garantire l’unicità dell’autovalore 1è stata introdotta una versione modificata della matrice dei collegamenti, come segue:

Il valore di m in genere scelto è pari a *m = 0.15*. Questa matrice M è positiva e stocastica, essendo anche primitiva, dal teorema di **Perron-Frobenius**[[1]](#footnote-1) segue che l’autovalore 1 di molteplicità 1 è l’unico autovalore di massimo modulo, inoltre il corrispondente autovettore è positivo, quindi da questo momento in poi ridefiniamo x\* utilizzando M anziché A, come segue

1. Provo la nota a piè di pagina [↑](#footnote-ref-1)