# 1.Introduzione

L’algoritmo *Pagerank* quantifica l’importanza di ogni pagina web basandosi sulla struttura dei link dell’intera rete.L’idea principale che sta alla base del funzionamento di Pagerank è che link puntano ad una pagina partendo da una pagina importante, rendono la pagina linkata importante. Ogni pagina, più concretamente ogni pagina “vota” le pagine ad essa collegata. Il problema viene matematicamente formulato dal calcolo di un autovettore corrispondente al più grande autovalore di una matrice stocastica. Attualmente il processo viene completato da un algoritmo centralizzato che data la dimensione della rete (oltre 8 miliardi di pagine) impiega circa una settimana. Allo scopo di risolvere il problema si sono propone un approccio basato su algoritmi distribuiti randomizzati.

L’algoritmo è caratterizzato da 3 aspetti:

1. Ogni pagina può calcolare il suo PageRank localmente comunicando con le pagine a cui è direttamente collegata
2. Ogni pagina decide di avviare la comunicazione con le altre in modo randomizzato.
3. La capacità computazionale richiesta per ogni pagina è molto bassa

L’approccio al calcolo è legato alla teoria del consenso in modo da tale che più nodi scambino i loro valori con i suoi vicini in modo tale da raggiungere il consenso, ovvero che tutti i nodi raggiungano lo stesso valore.

# 2. Pagerank

Consideriamo una rete di n pagine web indicizzate da 1 a n. La rete è rappresenta da un grafo *G*=(*V*, *E*) in cui V è il set dei nodi ed E è il set dei collegamenti (link) appartenente a VxV. (i,j) è un collegamento se esiste un link in uscita che mi conduce da i a j. Il Pagerank di una pagina i è un numero reale in [0,1] denotato come . In pratica una pagina i ha maggiore importanza di una pagina j, se

Il valore di ogni pagina (PageRank) è calcolato come Dove è l’insieme delle pagine che hanno un link entrante in i, mentre è il numero di link che escono dal nodo j. È buona norma normalizzare i valori in modo tale che . Riscriviamo i valori in forma vettoriale, cioè:

e secondo questa notazione possiamo riscrivere il problema di Pagerank nella seguente forma:

e dove la matrice A = ( chiamata matrice dei link in cui gli elementi sono definiti come:

Notiamo che il vettore è l’autovettore corrispondente all’autovalore 1 della matrice A. In generale questo autovettore esiste ed è unico se il grafo è fortemente connesso. Seppure la rete reale non sia fortemente connessa a causa di alcune pagine in seguito assumeremo che tutte le pagine sono connesse. Per questo motivo la matrice A diventa stocastica rispetto alle colonne e quindi esisterà sempre l’autovalore 1. Per garantire l’unicità dell’autovalore 1 introduciamo una versione modificata della matrice dei link A, definita come segue:

con m parametro tipico di 0.15. M è positiva stocastica e per il teorema di Perron la matrice è primitiva, in particolare la sua molteplicità è 1 ed è unico. Quindi ridefiniamo il vettore come segue:

Come detto, il calcolo dell’autovettore per matrici di grandi dimensioni è complesso. Per questo motivo viene usata la ricorsione:

Dove x(k) è un vettore di probabilità e notiamo che Sx(k) = **1.** Questa formula è importante perché ci permette di usare direttamente la matrice sparsa A (gli elementi sono quasi tutti uguali a 0).

Per il lemma 2.3 del paper il vettore x(k) converge al valore del consenso per

# 3. Un approccio distribuito randomizzato

Proponiamo un approccio distribuito al calcolo del vettore Lo schema base del protocollo è il seguente:

All’istante k, la pagina i inizia l’aggiornamento del valore del Pagerank mandando il proprio valore a tutte le pagine collegate e richiedendo a sua volta i loro valori. Per implementare lo schema distribuito assumiamo che le pagine che attivano l’aggiornamento sono determinate in modo casuale. Questo è determinato dal processo aleatorio . Ogni nodo può iniziare il processo con uguale probabilità, ovvero

In particolare consideriamo lo schema di aggiornamento distribuito nella seguente forma

Con matrice dei link distribuita. Definiamo la media temporale dei valori di x y(k) come

La media temporale y(k) converge al valore del consenso nel senso di media quadrata

Questo tipo di convergenza è detta ergodicità per processi stocastici.

## 3.1 Matrici di link distribuite

Definiamo le matrici dei link distribuite in modo tale che:

* la i-esima riga e colonna coincidono con quella di A
* Gli elementi sulla diagonale sono 1 - l=1…n e li
* Tutti gli altri elementi sono 0

Formalmente:

Queste matrici sono stocastiche in quanto la matrice originale è stocastica. Per chiarire le proprietà sulla matrice di link consideriamo il seguente semplice schema di update

Quindi noi siamo interessati ad osservare la dinamica mediata. Definiamo come matrice media dei link distribuita e la media di x(k) come e la sua ricorsione è

**Lemma 3.2** La matrice ha le seguenti proprietà:

* Esiste un autovettore corrispondente all’autovalore 1 per entrambe le matrici A e

## 3.2 Convergenza quadratica media

Come nel caso centralizzato con la matrice dei link A, risolviamo la possibile non unicità dell’autovalore 1 riscrivendo lo schema distribuito come:

Dove è M definito come:

A questo punto il problema diventa trovare un valore di tale che ed M condividano l’autovettore dell’autovalore 1 che necessariamente è il vettore del consenso. Sia e la sua dinamica è espressa da con . Ma al contrario di A non esiste una relazione in forma chiusa tra M ed , tuttavia se utilizziamo con m = 0.15 abbiamo il seguente lemma:

**Lemma 3.3**

* < m
* =
* Per l’autovalore è 1, è semplice, unico, massimo in modulo ed il suo autovettore è il vettore del consenso

Tuttavia data la natura aleatoria degli non possiamo assicurare la convergenza di x(k) al consenso, tuttavia il seguente teorema ci dimostra che possiamo raggiungere il valore del consenso tramite una media quadratica

**Teorema 3.4**Nello schema di aggiornamento distribuito, il valore di PageRank è ottenuto dalla media temporale y come ]🡪0 per k 🡪∞