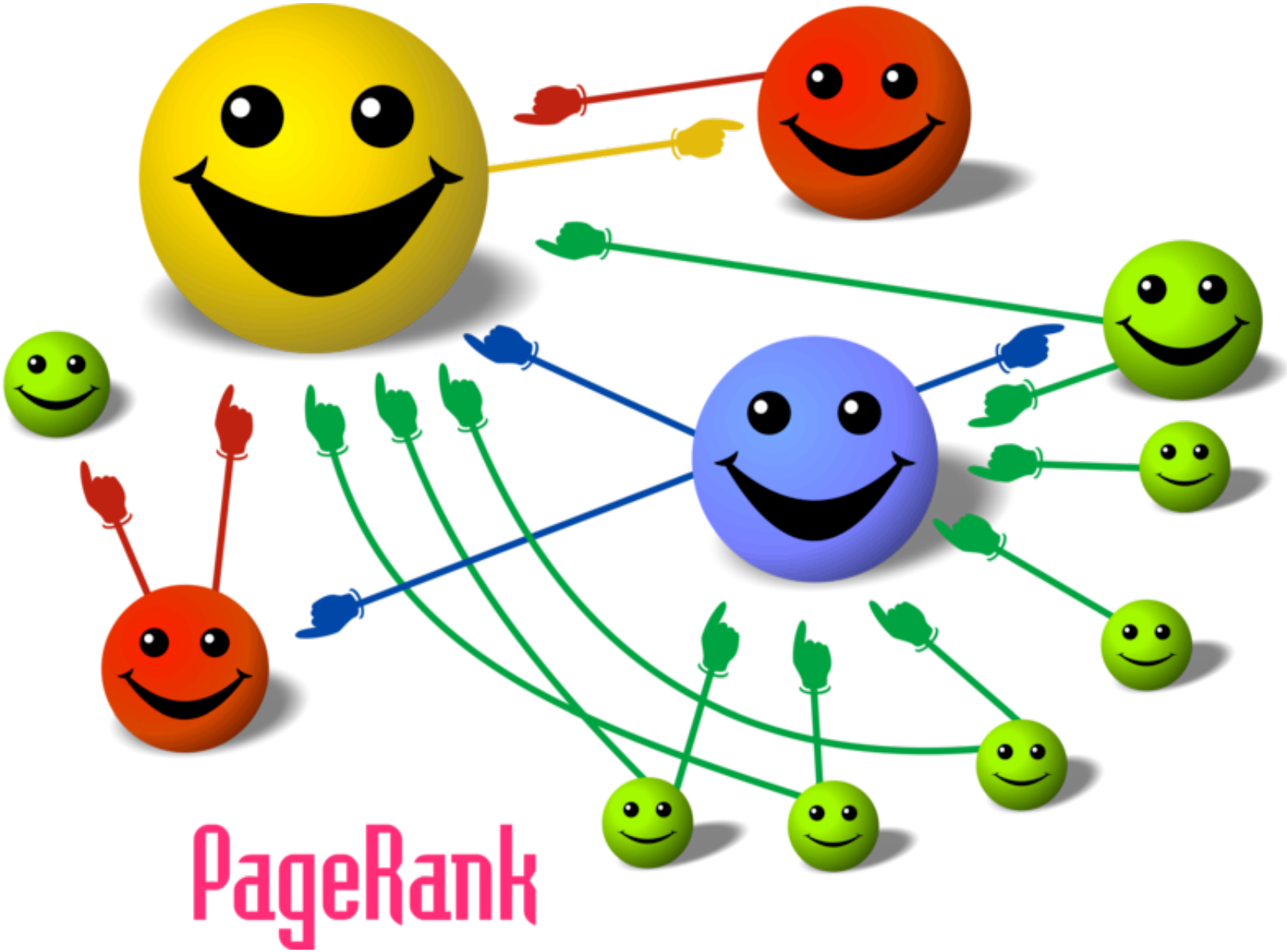
Il ranking di pagine web



Consideriamo l'insieme delle pagine web con i relativi link. Questo può essere considerato come un grafo orientato (con miliardi di nodi). Concentriamoci su l'esempio più piccolo della pagina precedente. Una rappresentazione alternativa è quella con la matrice di adiacenza:

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Il problema consiste nel dare un valore ai vari nodi (le pagine) utilizzando la struttura del grafo (o della matrice).

Il principio base è

Una pagina è importante se è citata da pagine importanti

Può sembrare una tautologia ma in realtà si presta ad una precisa definizione e trattazione matematica. Se x_i è l'importanza della pagina i, si scrive un'equazione del tipo

$$x_j = \sum_{i=1}^{N} x_i p_{ij}$$

ovvero

$$x^T = x^T P$$

Quindi l'importanza di una pagina è data dalla somma delle importanze delle pagine che la citano.

Questa formula si presta ad alcune obiezioni.

Innanzitutto bisogna normalizzare P per righe in modo che l'importanza di una pagina venga divisa tra tutti i suoi link, nel fare questo bisogna anche tenere conto delle righe nulle (ovvero delle pagine senza out link).

$$Q = \{q_{ij}\}, \quad q_{ij} = \frac{p_{ij}}{\max\left(1, \sum_{k=1}^{N} p_{ik}\right)}$$

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

L'interpretazione probabilistica

Se nella matrice Q non fossero presenti righe nulle, allora questa sarebbe una matrice stocastica e una soluzione dell'equazione

$$x^T = x^T Q$$

rappresenterebbe la probabilità di arrivare in una certa pagina scegliendo i link di uscita in modo casuale con probabilità uniforme.

C’è però il problema dei “dangling nodes” ovvero delle pagine senza link in uscita (la prima riga della matrice). Si suppone quindi che dai nodi senza out-link si esca in modo equiprobabile verso ogni altra pagina e la matrice Q viene quindi modificata con una correzione di rango 1:

$$R = Q + de^T$$

dove

$$e^T = (1,1,...,1), \qquad d_i = \begin{cases} 0, & \exists j : q_{ij} \neq 0 \\ 1 & \text{altrimenti} \end{cases}$$

$$\mathbf{R} = \begin{pmatrix} \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} & \frac{1}{11} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Un ultimo problema è rappresentato dalla possibilità (praticamente una certezza) che la matrice vi siano sottoinsiemi di nodi non connessi con il resto. La soluzione è supporre che con probabilità (1-α) si possa saltare ad una qualsiasi pagina, ovvero si considera la matrice

$$S = \alpha\left(Q + de^T\right) + (1-\alpha)\frac{ee^T}{N}$$

Vediamo come appare *S* (moltiplicata per *N* per occupare meno spazio sul foglio)

$$\mathbf{N\,S} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1-\alpha & 1-\alpha & 1+10\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+10\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1+\frac{9\alpha}{2} & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+\frac{8\alpha}{3} & 1-\alpha & 1+\frac{8\alpha}{3} & 1-\alpha & 1+\frac{8\alpha}{3} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1+\frac{9\alpha}{2} & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1+10\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha \\
1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1+10\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha & 1-\alpha
\end{pmatrix}$$

La matrice S è

- Stocastica (è positiva e la somma delle righe è sempre 1)
- Irriducibile
- Ha un solo autovalore uguale ad 1 e tutti gli altri sono in modulo minori di 1.

Si può quindi applicare il teorema di Perron-Frobenius che afferma, tra l'altro, che l'autovettore associato all'autovalore massimo è tutto positivo, ovvero la soluzione dell'equazione

$$x^T = x^T S$$

è unica e può essere normalizzata rendendola un vettore di probabilità.

Soluzione del problema

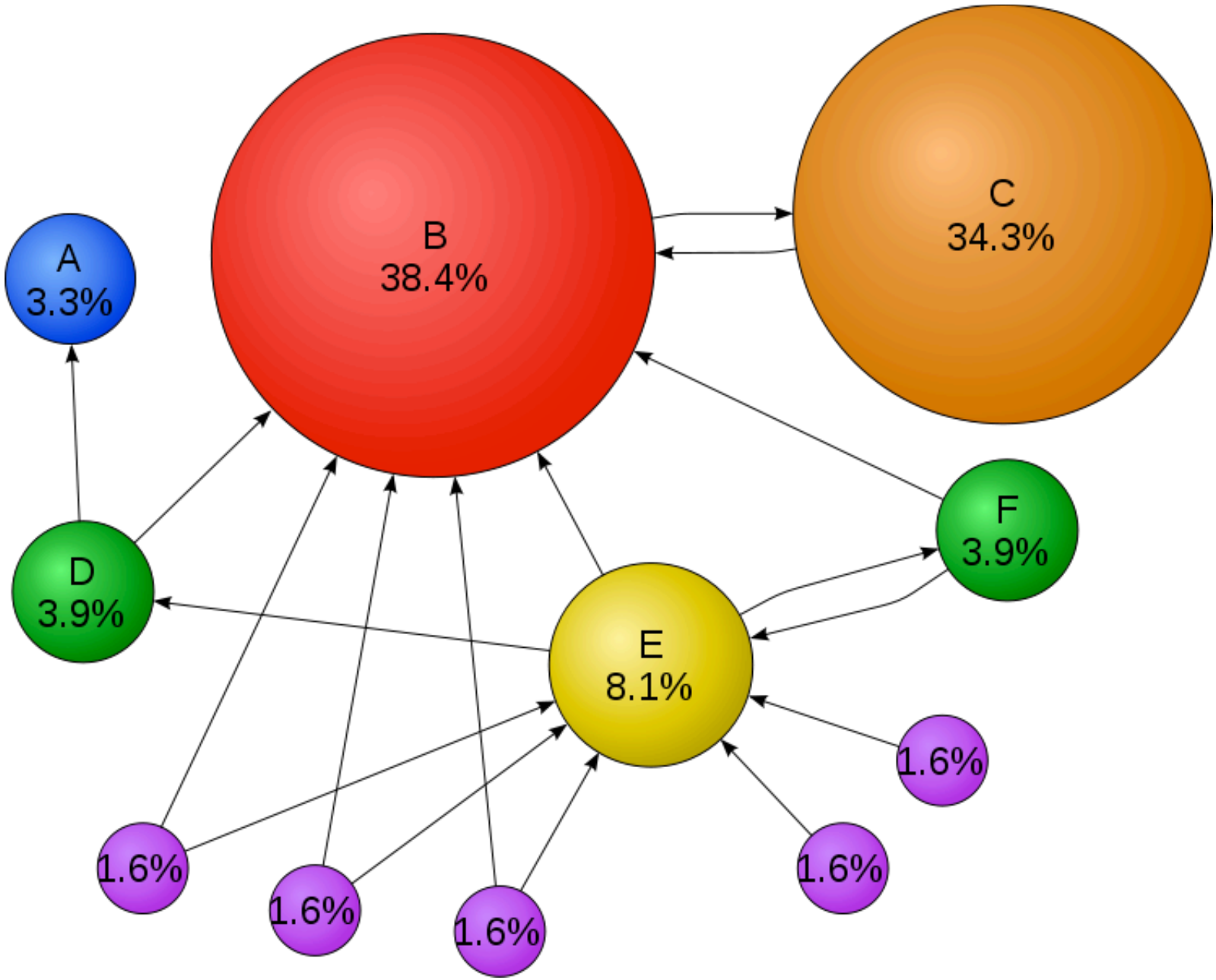
Per risolvere il problema in genere si sceglie un valore di α abbastanza lontano da 1 (tipicamente 0.85) ottenendo una matrice reale piena.

$$\begin{pmatrix}
0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 & 0.0909091 \\
0.0136364 & 0.0136364 & 0.863636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.863636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.438636 & 0.438636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.29697 & 0.0136364 & 0.29697 & 0.0136364 & 0.29697 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.438636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.863636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 \\
0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.863636 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364 & 0.0136364
\end{pmatrix}$$

La soluzione che si ottiene è il vettore

$$x = \{0.033, 0.384, 0.343, 0.039, 0.081, 0.039, 0.016, 0.0161695, 0.016, 0.016, 0.016\}$$

e l’importanza dei nodi può essere anche evidenziata graficamente


A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%

Aspetti matematici e dettagli implementativi

La matrice S gode di alcune importati proprietà, essendo stocastica

$$Se = e, \quad \|S\|_\infty = 1, \quad \rho(S) = 1$$

Ovvero l'autovalore di modulo massimo vale esattamente 1, gli altri autovalori sono tutti in modulo minori di 1, tanto più lontani da 1 quanto più piccolo è il valore di α.

Quindi la soluzione cercata si può ottenere con un metodo iterativo, specificamente con il metodo delle potenze

$$x^{(i+1)} = S^T x^{(i)}, \quad \left\|x^{(0)}\right\|_1 = 1$$

Partire con un vettore a norma-1 unitaria permette di ottenere il risultato già normalizzato come vettore di probabilità; comunque l'iterazione conserva in ogni caso la norma-1.

$$\left\|x^{(i+1)}\right\|_1 = \sum_{k=1}^{N} x_k^{(i+1)} = \sum_{k=1}^{N}\sum_{j=1}^{N} x_j^{(i)} S_{jk} = \sum_{j=1}^{N} x_j^{(i)} \sum_{k=1}^{N} S_{jk} = \sum_{j=1}^{N} x_j^{(i)} = \left\|x^{(i)}\right\|_1$$

Per quanto riguarda la struttura le matrici P, Q, ed R sono sparse mentre S è piena, questo potrebbe sembrare un grave problema sia dal punto del tempo di esecuzione che del numero di operazioni da effettuare, in realtà l'algoritmo di risoluzione può facilmente preservare la sparsità.

Si ha infatti

$$x^T S = \alpha x^T Q + \alpha x^T d e^T + \frac{(1-\alpha) x^T e}{N} e^T =$$

$$= \alpha x^T Q + \left[\alpha x^T d + \frac{(1-\alpha)}{N} \|x\|_1 \right] e^T$$

dove la quantità tra parentesi quadre è uno scalare.

Digressione: operazioni su matrici sparse

Una matrice $N \times N$ con M elementi diversi da zero con M tipicamente maggiore di N ma molto minore di N^2 si dice **sparsa.**

Le matrici sparse possono essere **strutturate** (per esempi le **matrici a banda**) o **non strutturate**. Le matrici sparse non strutturate hanno origine in molti campi della matematica applicata, della fisica. dell'ingegneria e dell'informatica.

Nel caso in esame la matrice del web è sparsa, non strutturata

I problemi tipici che sorgono trattando le matrici sparse sono la memorizzazione e le operazioni, tipicamente il calcolo del prodotto matrice-vettore.

Memorizzare in modo efficiente una matrice sparsa significa non memorizzare gli zeri conservando contemporaneamente tutte le informazioni sulla posizione degli elementi non nulli.

Memorizzazione con vettori di posizione.

Si usano due vettori interi I e J che indicano le posizioni di riga e colonna dell'elemento corrispondente nel vettore reale V, vediamo l'esempio per la matrice Q.

Spazio occupato: $2M$ interi e M reali.

I	J	V
2	3	1
3	2	1
4	1	0.5
4	2	0.5
5	2	0.33333
5	4	0.33333
5	6	0.33333
6	2	0.5
6	5	0.5
7	2	0.5
7	5	0.5
8	2	0.5
8	5	0.5
9	2	0.5
9	5	0.5
10	5	1
11	5	1

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Memorizzazione per righe

Si usano due vettori interi: I di N+1 posizioni che indica la posizione in V del primo elemento della corrispondente riga, J di M posizioni che indica le posizioni colonna dell'elemento corrispondente nel vettore reale V, vediamo l'esempio per la matrice Q.

Spazio occupato: (N+M+1) interi e M reali.

J	V
3	1
2	1
1	0.5
2	0.5
2	0.33333
4	0.33333
6	0.33333
2	0.5
5	0.5
2	0.5
5	0.5
2	0.5
5	0.5
2	0.5
5	0.5
5	1
5	1

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$I = \{1, 1, 2, 3, 5, 8, 10, 12, 14, 16, 17, 18, 19\}$$

Esempi in Java (NB in Java gli indici partono da 0)

Prodotti Ax e $A^{\mathrm{T}}x$, memorizzazione con vettori di posizione.

```
int[] I = new int[M], J = new int[M];
double[] V = new double[M];
   ...
public double[] times(double[] x, int N) {
   double[] r = new double [N];
   for(int k=0; k<M; k++)
      r[I[k]]+=V[k]*x[J[k]];
   return r;
}

public double[] timesT(double[] x, int N) {
   double[] r = new double [N];
   for(int k=0; k<M; k++)
      r[J[k]]+=V[k]*x[I[k]];
   return r;
}
```

Prodotti Ax e $A^{\mathrm{T}}x$, memorizzazione per righe.

```
int[] I = new int[N+1], J = new int[M];
double[] V = new double [M];
   ...
public double[] times(double[] x, int N) {
   double[] r = new double[N];
   for(int i=0; i<N; i++)
      for(int k=I[i]; k<I[i+1]; k++)
         r[i]+=V[k]*x[J[k]];
   return r;
}

public double[] timesT(double[] x, int N) {
   double[] r = new double[N];
   for(int i=0; i<N; i++)
      for(int k=I[i]; k<I[i+1];k++)
         r[J[k]]+=V[k]*x[i];
   return r;
}
```