The PageRank Algorithm

Maximilien Danisch

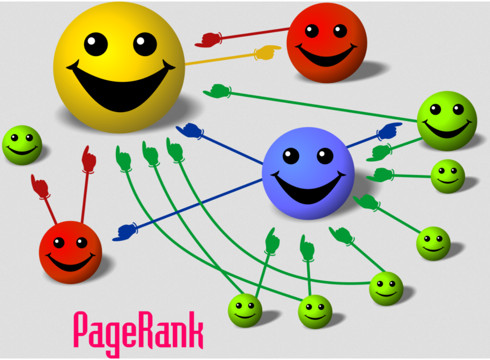
Sorbonne Université

first_name.last_name@lip6.fr

Outline

What is PageRank?

- PageRank is an algorithm used by Google Search to rank websites in their search engine results.
- It counts the number and quality of links to a page.
- The underlying assumption is that an important website is likely to receive links from other important websites.



Random walks

Pagerank is based on random walks:

- a walker starts at a node chosen uniformly at random
- it follows one of the out-links chosen uniformly at random
- go to 2

Score of a node v = probability to have the random walker on node v after an infinite number of steps.

Problem: what if the graph has a dead-end? has a spider trap? is cyclic?

Random walks

Pagerank is based on random walks:

1. a random walker starts at node u
2. it then teleports to a random node with probability α
3. if it does not teleport then:
   - it follows one of the outlinks chosen uniformly at random
   - if the node has no outlinks it teleports to a random node
4. go to 2

Analogy with a surfer on the web...

PageRank(v) $=$ probability to have the random walker on node v after an infinite number of steps.

Outline

The transition matrix of a graph

Definition: given a directed graph G with n nodes and m directed edges, its transition matrix T is define as follows.

- T is an n by n matrix with m non-zero values
- for each directed edge (u, v) in G, $T_{vu} = \frac{1}{d^{out}(u)}$

Definition: if G has no dead-end (nodes with $d^{out} = 0$), then the PageRank vector P is given by the following equation:

$$P_{t+1} = (1 - \alpha) \times T \times P_t + \alpha \times I$$

where I is the vector such that each entry equals to $\frac{1}{n}$.
Usually $0.1 \leq \alpha \leq 0.2$.

Question: P is the top eigenvector of which matrix?

The transition matrix of a graph

If G has dead-ends then the augmented transition matrix T' should be used instead of T:

- for each directed edge (u, v) in G, $T'_{vu} = \frac{1}{d^{out}(u)}$
- if $d^{out}(u) = 0$, then $\forall v$, $T'_{vu} = \frac{1}{n}$

Definition: the PageRank vector P is given by the following equation:

$$P_{t+1} = (1 - \alpha) \times T' \times P_t + \alpha \times I$$

where I is the vector such that each entry equals to $\frac{1}{n}$. Usually $0.1 \leq \alpha \leq 0.2$.

Question: What if the graph has many dead-ends?

Power iteration

$$P_{t+1} = (1 - \alpha) \times T' \times P_t + \alpha \times I$$

Algorithm 1 Power iteration to compute PageRank

function $\textsc{PowerIteration}(G, \alpha, t)$
    $T \leftarrow$ transition matrix of graph G
    $P \leftarrow \frac{1}{n} \times I$
    for i from 1 to t **do**
        $P \leftarrow \textsc{MatVectProd}(T, P)$
        $P \leftarrow (1 - \alpha) \times P + \alpha \times I$
        $P \leftarrow \textsc{Normalize2}(P)$ $\triangleright \forall i \in [\![1, n]\!], P[i] += \frac{1 - ||P||_1}{n}$
    return P

Outline

1. What is PageRank?
2. Computation using the power iteration method
3. Personalized PageRank

Personalized PageRank

$$P_{k+1} = (1 - \alpha) \times T'_{P_0} \times P_k + \alpha \times P_0$$

Algorithm 2 Power iteration to compute rooted PageRank

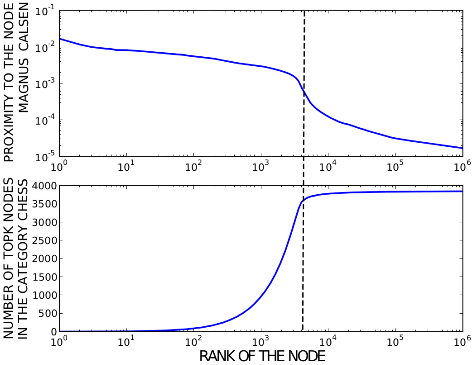
function $\textsc{PowerIteration}(G, P_0, \alpha, t)$
  $T \leftarrow$ transition matrix of graph G
  $P \leftarrow \frac{1}{n} \times I$
  for i from 1 to t **do**
    $P \leftarrow \textsc{MatVectProd}(T, P)$
    $P \leftarrow (1 - \alpha) \times P + \alpha \times P_0$
    $P \leftarrow \textsc{Normalize2}(P)$ $\triangleright$ $\forall i \in [\![1, n]\!], P[i] += P_0[i]\frac{1-||P||_1}{n}$
  return P

Definition: the *Rooted PageRank in u* is the personalized PageRank such that $P_0[u] = 1$.

Rooted PageRank as a "proximity metric"

- The distance may not be a good "proximity metric". Why?
- Rooted Pagerank might be better.

Experiments in the Wikipedia network:



https://tel.archives-ouvertes.fr/tel-01207046