PageRank: a billion dollar formula

Yuanyuan LIN

The Chinese University of Hong Kong

Nov 1, 2018

PageRank

- PageRank assigns a numerical weighting to each one of a set of hyperlinked webpages, with the purpose of measuring its relative importance.

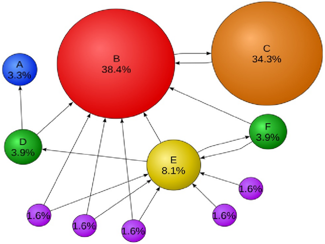
PageRank

- PageRank assigns a numerical weighting to each one of a set of hyperlinked webpages, with the purpose of measuring its relative importance.
- This numerical weight assigned to a webpage (E)is referred to as the PageRank of $E$and denoted by$PR(E)$.

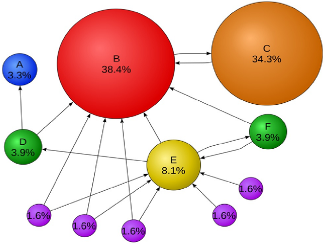
PageRank

- PageRank assigns a numerical weighting to each one of a set of hyperlinked webpages, with the purpose of measuring its relative importance.
- This numerical weight assigned to a webpage (E)is referred to as the PageRank of $E$and denoted by$PR(E)$.
- The name "PageRank" is a trademarkof Google, named after its Co-founder Larry Page.

An illustrative example



An illustrative example



PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.
- A websurfer moves from page to page forms a MC. Once a websurfer is on page i, regardless of how he gets there, he has chance d to continue to one of these L_i linked pages equally likely, and chance $1 - d$ to a random page out of the totally N pages.

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.
- A websurfer moves from page to page forms a MC. Once a websurfer is on page i, regardless of how he gets there, he has chance d to continue to one of these L_i linked pages equally likely, and chance $1 - d$ to a random page out of the totally N pages.
- Transition probabilities:

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.
- A websurfer moves from page to page forms a MC. Once a websurfer is on page i, regardless of how he gets there, he has chance d to continue to one of these L_i linked pages equally likely, and chance $1 - d$ to a random page out of the totally N pages.
- Transition probabilities:
  - $P_{ij} = (1 - d)/N$, if webpage i does not outward links to webpage j.

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.
- A websurfer moves from page to page forms a MC. Once a websurfer is on page i, regardless of how he gets there, he has chance d to continue to one of these L_i linked pages equally likely, and chance $1 - d$ to a random page out of the totally N pages.
- Transition probabilities:
  - $P_{ij} = (1 - d)/N$, if webpage i does not outward links to webpage j.
  - $P_{ij} = d/L_i$, if web i outward links to web j.

PageRank: as limit of Markov Chain

- Consider webpages $1, 2, \ldots, N$. Imagine them as the state space of a MC. PR_i is the PageRank of page i, which, in fact, is π_i, the limiting probability of a MC.
- Suppose page i has outbound link to L_i web pages.
- A websurfer moves from page to page forms a MC. Once a websurfer is on page i, regardless of how he gets there, he has chance d to continue to one of these L_i linked pages equally likely, and chance $1 - d$ to a random page out of the totally N pages.
- Transition probabilities:
    - $P_{ij} = (1 - d)/N$, if webpage i does not outward links to webpage j.
    - $P_{ij} = d/L_i$, if web i outward links to web j.
- Let $\mathbf{P} = (P_{ij})$ be the $N \times N$ transition probability matrix.

The Damp Factor: d

- d is called the damp factor, estimated to be about 85%.

The Damp Factor: d

- d is called the damp factor, estimated to be about 85%.
- The damp factor makes the MC regular.

The Damp Factor: d

- d is called the damp factor, estimated to be about 85%.
- The damp factor makes the MC regular.
- Without the regularity, the limit law of the MC will not hold. And the MC converges to black holes (pages without outbound links.)

The limit law of Markov Chain

It is known that

$$(\pi_1, \ldots, \pi_N) = (\pi_1, \ldots, \pi_N)\mathbf{P}, \qquad \pi_1 + \ldots + \pi_N = 1,$$

that is

$$\pi_i = \pi_1 P_{1i} + \pi_2 P_{2i} + \ldots + \pi_N P_{Ni},$$

or

$$\pi_i = \sum_k \pi_k (1-d)/N + \sum_{k \text{that outbound links } i} \pi_k d/L_k$$

- The above equation gives the PageRank PR_i, which is the solution of

$$PR_i = (1-d)/N + d \sum_{k \text{ that outbound links } i} PR_k/L_k$$

The limit law of Markov Chain

It is known that

$$(\pi_1, \ldots, \pi_N) = (\pi_1, \ldots, \pi_N)\mathbf{P}, \qquad \pi_1 + \ldots + \pi_N = 1,$$

that is

$$\pi_i = \pi_1 P_{1i} + \pi_2 P_{2i} + \ldots + \pi_N P_{Ni},$$

or

$$\pi_i = \sum_k \pi_k (1-d)/N + \sum_{k \text{that outbound links } i} \pi_k d / L_k$$

- The above equation gives the PageRank PR_i, which is the solution of

$$PR_i = (1-d)/N + d \sum_{k \text{ that outbound links } i} PR_k / L_k$$

Solving for the PageRank

- The total number of webpages N is very large (could be millions).

Solving for the PageRank

- The total number of webpages N is very large (could be millions).
- L_i, the number of outbound links, small for every i.

Solving for the PageRank

- The total number of webpages N is very large (could be millions).
- L_i, the number of outbound links, small for every i.
- The PageRank is the eigenvector of $I - \mathbf{P}^\top$.

Solving for the PageRank

- The total number of webpages N is very large (could be millions).
- L_i, the number of outbound links, small for every i.
- The PageRank is the eigenvector of $I - \mathbf{P}^\top$.
- Computationally difficult/slow to find the eigenvector for $I - \mathbf{P}^\top$.

Solving for the PageRank

- The total number of webpages N is very large (could be millions).
- L_i, the number of outbound links, small for every i.
- The PageRank is the eigenvector of $I - \mathbf{P}^\top$.
- Computationally difficult/slow to find the eigenvector for $I - \mathbf{P}^\top$.
- The limit law of MC:

$$P_{ij}^n \to \pi_j$$

, regardless of i.

Solving for the PageRank

- The total number of webpages N is very large (could be millions).
- L_i, the number of outbound links, small for every i.
- The PageRank is the eigenvector of $I - \mathbf{P}^\top$.
- Computationally difficult/slow to find the eigenvector for $I - \mathbf{P}^\top$.
- The limit law of MC:

  $$P_{ij}^n \to \pi_j$$

  , regardless of i.
- Use iteration, beginning with $PR_i = 1/N$ for all i, converge to the PageRank. The iteration converges fast, thanks to small L_i.