Page Rank Measure

[Carrie, Kazman 97]

indegree number of pages that reference a page

intuition : P_1 points to P_2 means author of P_1 recomends P_2.

[Brin,Page,..]

recursive variation of indegree.

Slide 1

The rank of a page depends on

- quality of pages that recomend it
- weight of such pages

Mathematical formula?

Random walk

$X = \{S_1, \ldots, S_n\}$ is a set of states.

For time t, there is a transition probability is $p_{ij}(t)$ representing th eprobability of going from S_i to S_j at time t.

A random walk on X corresponds to a (infinite) sequence of states, one for each step of the walk.

After walking long enough the system will reach some equilibrium state with fixed probabilities of being in a particular state?

Slide 2

The system is described by a linear equation $P(t) = (p_{ij})$, if the system at time t has state probabilities $\pi_t = (\pi_1, \ldots, \pi_n)$, in the next step will have state probabilities

$$\pi_{t+1} = \pi_t P(t)$$

In a Markovian random walk

the transition probability at each step is independent of time and history, thus it only depend on the previous state and the transitionprobability.

For each pair of states the transition probability is p_{ij}

Slide 3

The system is described by a linear equation $P = (p_{ij})$, if the system at time t has state probabilities $\pi = (\pi_1, \ldots, \pi_n)$, in the next step will have state probabilities

$$\pi_{t+1} = \pi_t P = \pi_0 P^t$$

That usually gets th ename of Markov chain

Slide 4

In the case of a Markov chain, it is well know when the system will reach some equilibrium.

If all the states are reachable and P is time reversible, the system converges to a unique state distribution π.

Furthermore

$$\pi = \pi P,$$

and for any initial state π_0 the random walk converges to the stationary π.

Methods to compute/approximate the stationary distribution of a given Markov chain

- Compute the eigenvectors.
- Simulate the behaviour of the chain for as many steps as posible.

Slide 5

The accuracy of the second method depends on the rate of convergence towards the stationary distribution.

In the context of a page

Web has T pages

choose a parameer d, $0 < d < 1$.

tipical values $0.1 \leq d \leq 0.15$

Let pages $p_1, \ldots, p_k$ link to page p and let $C(p)$ be the number of links out of p. Define

Slide 6

$$R(p) = d\frac{1}{T} + (1-d)\sum_{i=1}^{k} \frac{R(p_i)}{C(p_i)}$$

R can be scaled to get $\sum_p R(p) = 1$ so that R is a probability distribution, divide $R(p)$ by $\sum_p R(p)$.

This distribution is used by the Google search engine

Interpretation of R in terms of surfer's random walks

Web surfer visits page p

- either jumps to another unrelated page with probability d
- or follow a link randomly selected from those on the current page. (prob $1-d$)

In the equilibrium distribution of such a random walk, the probability of p is $R(p)$.

Slide 7

Sampling pages according to their PageRank

- crawl a significant portion of the web, and determine the PageRank for the crawled pages.

  the measure so obtained may not be the real one, but we hope to approximate it
- proposal: perform a random walk, with equilibrium distribution = PageRank

  with prob $1-d$ follow a link — easy

  with prob d select a new page (unrelated?)
- How many steps? as much as possible!

  Although we will not perform a pure random markovian walk, we could expect a limiting distribution and hope that it is achieved quickly.

Slide 8

- Jump randomly (uniformly) to a page seen in the walk.

  bad performance, non-markovian

- Select a host uniformly at random from the set of hosts encountered in the walk (do not record the host of a page with no links)

  Select a page in the host, that have appeared in the walk

  still some bias, lot of space to record a long walk

  reduce size by limiting the information stored.

Slide 9