https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018

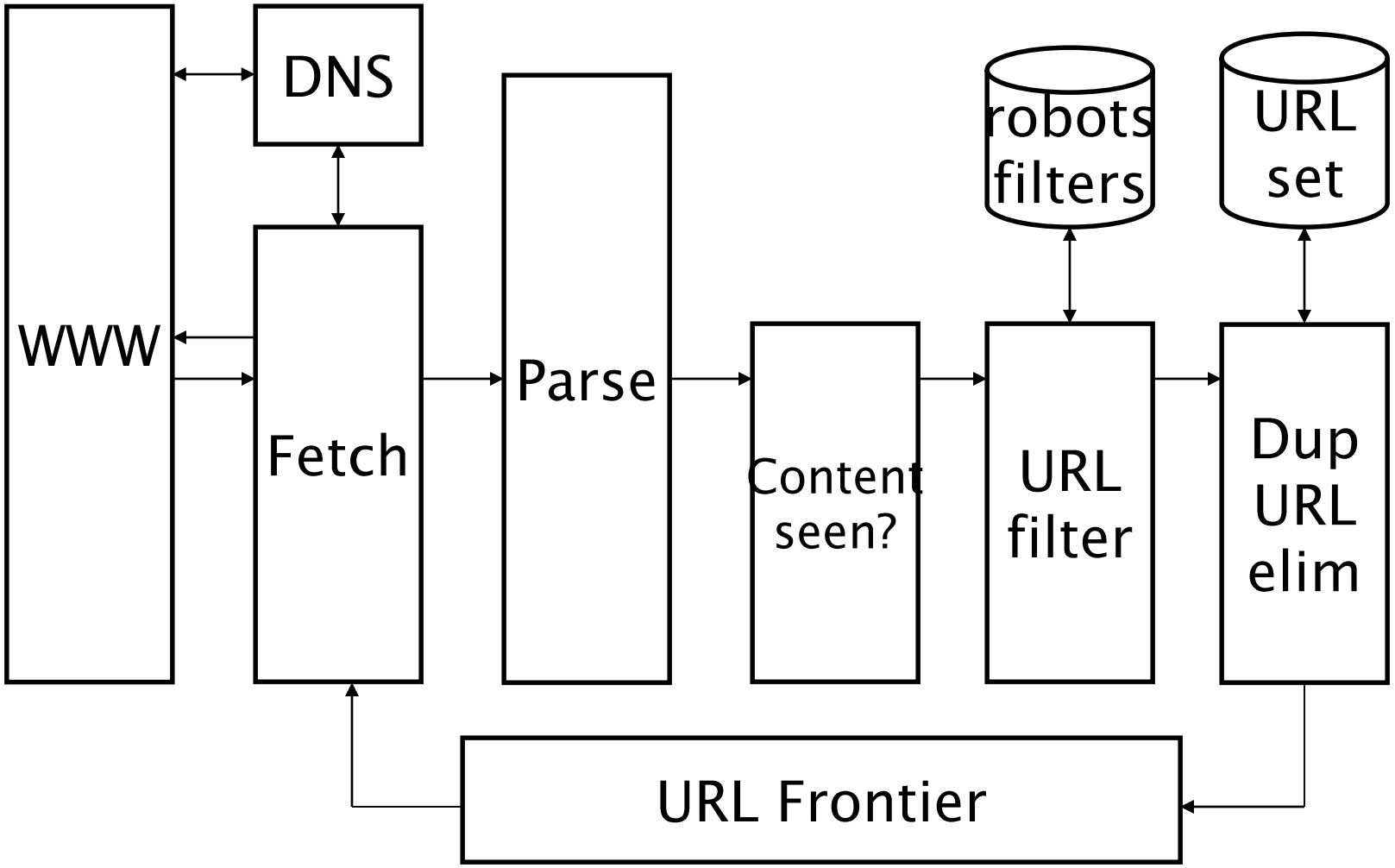
Indian Institute of Information Technology, Sri City

I still think Google uses **PageRank!**

– A Random User on the Web

Crawlers

Basic crawl architecture



Robots.txt

```
User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:
```

Challenges

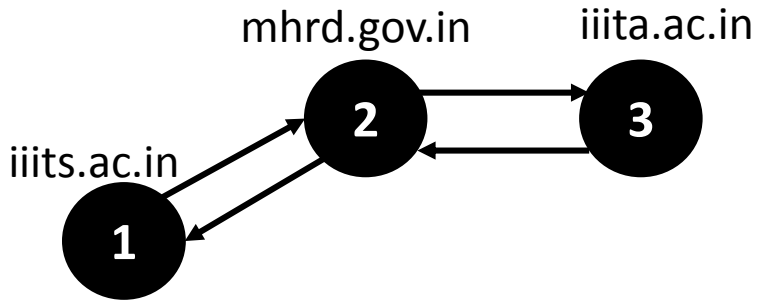
- We do not have a list of all URLs
- Link Extraction
- Avoiding Spider Traps
  - Dynamically create and respond to new URLs from every page within the same domain.
- Duplicate Sites
- Politeness
  - Access only once every n seconds.
- Storing page-related information (like popularity)

Link Analysis

Popularity as Search Parameter

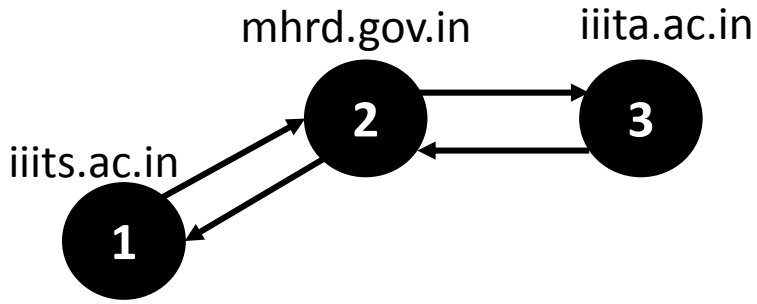
Which pages are popular?

The Web as a Graph



A Random Surfer

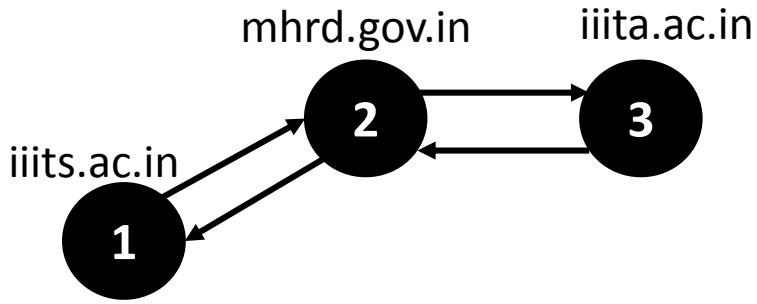
- May start from any node with 1/3 probability



- Can you represent this graph using Adjacency Matrix?

A Random Surfer

- May start from any node with 1/3 probability

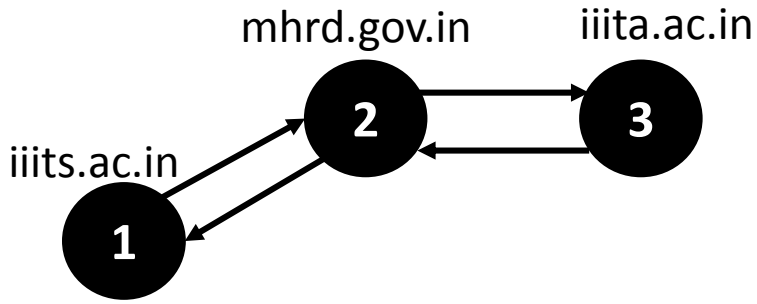


- Can you represent this graph using Adjacency Matrix?

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

A Random Surfer

- May start from any node with 1/3 probability



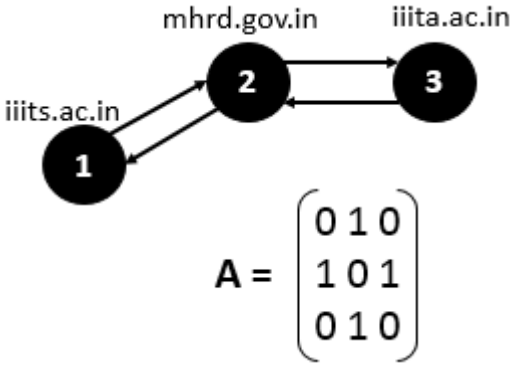
- May also teleport to any node with α probability

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

How can you compute the transition probabilities?

Transition Probabilities

- We can convert the Adjacency Matrix (A) to Transition Probability Matrix (P)

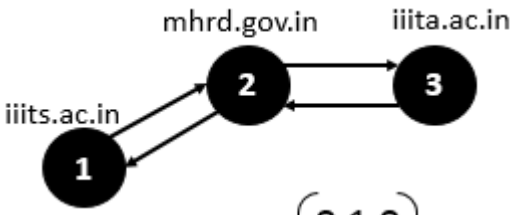


- If the random surfer is at 1,
  - and he did not teleport
    - Probability = 1- α
  - and he teleports
    - he may reach state3 with probability α/3
    - and may reach state 2 with probability α/3

- Transition Probability from 1 is (α/3, (1- α) + α/3, α/3)

Quiz

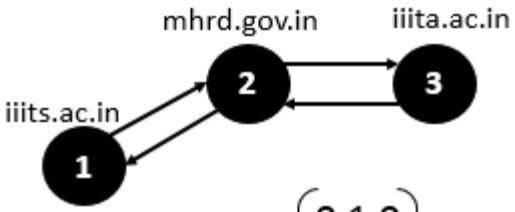
- If the teleportation probability, α = 0.5, Calculate the transition probability matrix for this network.



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

Quiz

- If the teleportation probability, α = 0.5, Calculate the transition probability matrix (P) for this network.



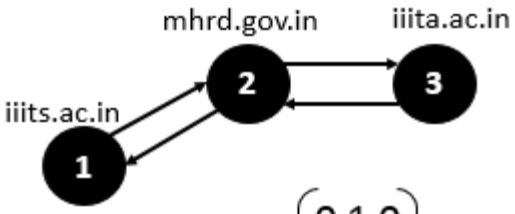
$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ ?? & ?? & ?? \\ ?? & ?? & ?? \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

Which page is more popular?

- If a random surfer at 1 can reach (1,2,3) with probabilities (1/6, 2/3, 1/6), where will he end up in the next time slot if choses to continue his walk?

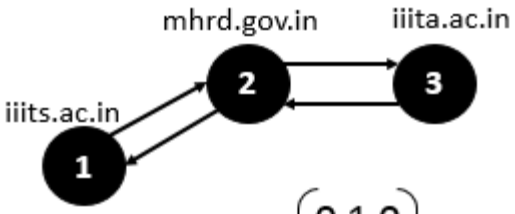


$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$P = (1/6, 2/3, 1/6) \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = (??, ??, ??)$$

Which page is more popular?

- If a random surfer at 1 can reach (1,2,3) with probabilities (1/6, 2/3, 1/6), where will he end up in the next time slot if choses to continue his walk?



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{P} = (1/6,\ 2/3,\ 1/6) \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = (1/3, 1/3, 1/3)$$

Steady State Probability

- If the random surfer keeps walking, the probabilities tend to converge!
  - Since we have an Ergodic Markov Chain
- In our case, we should get (5/18, 8/18, 5/18)
- So, the page 2 gets the highest rank.

Thank You