Chapter 4: Directions and Weights

4.1 and 4.2: The Web as an Information Network

These slides:

tinyurl.com/analytics20210315

License:

4.1 Directed Networks

- Edge goes *from* source node *to* target node
- Each node has a degree
    - in-degree: number of incoming links
    - out-degree: number of outgoing links
- Observed in information networks, such as:
    - email
    - wikipedia
    - journal publications

4.2 The web

The world wide web as an information network:

- **Node**: anything with a URL
- **Edge**: link to a URL (directed)

Browser and server communicate through HTTP protocol:

- *Client* sends URL (e.g. host.npr.org)
- *DNS* returns IP (e.g. 216.35.221.76)
- *Client* sends request (e.g.)

```
GET / HTTP/1.1
Host: npr.org
```

- *Server* returns content (e.g.)

```
HTTP 1.1 200 OK
Content-Type: text/html

<HTML>
<HEAD>
<Title>National Public Radio</Title>
 ...
</HTML>
```

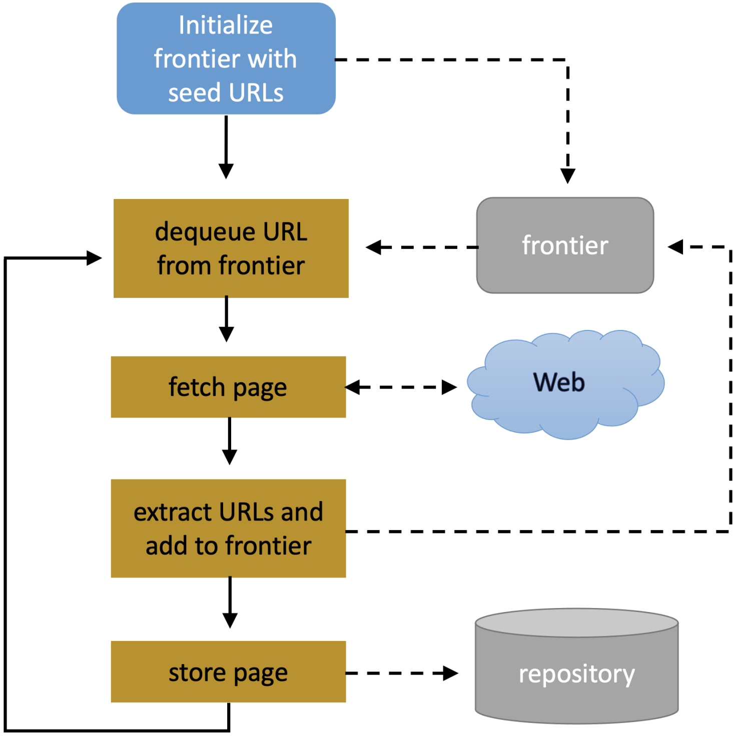
Web crawler

Breadth-first search algorithm running on the Web link graph

- Starts from (high-quality/relevant) seed pages
- Recursively extracts links and adds them to 'frontier'
- Technically challenging, but conceptually simple

The frontier:

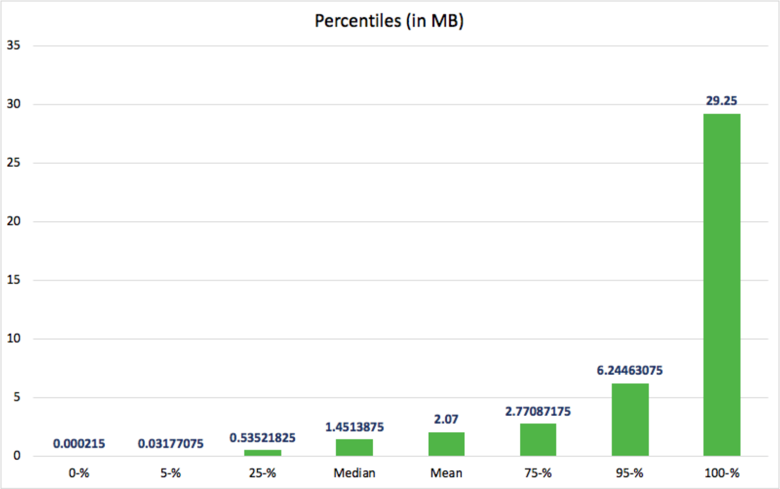
- First in, first out
- Pages at distance *n-1* before those at distance *n*
- Thus: high initial relevance to seed
- Often optimized, revisit known pages

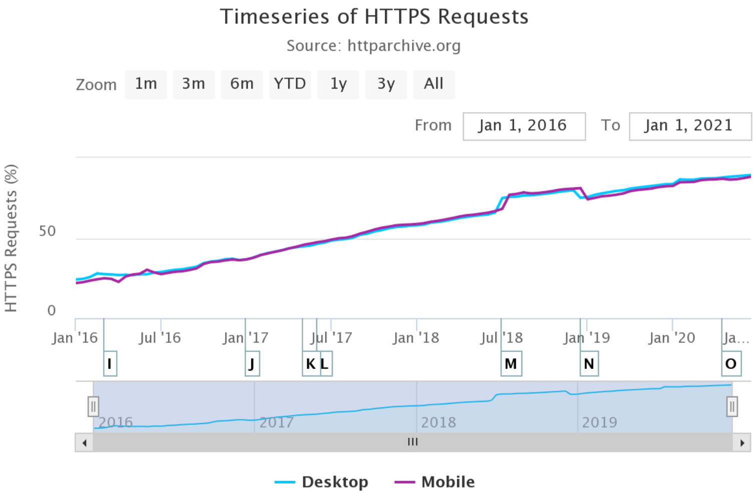

Initialize frontier with seed URLs
dequeue URL from frontier
frontier
fetch page
Web
extract URLs and add to frontier
store page
repository

Web structure and size

Web graph: the network of Web pages and hyperlinks

- Many (weakly) connected components
- Skewed size distribution



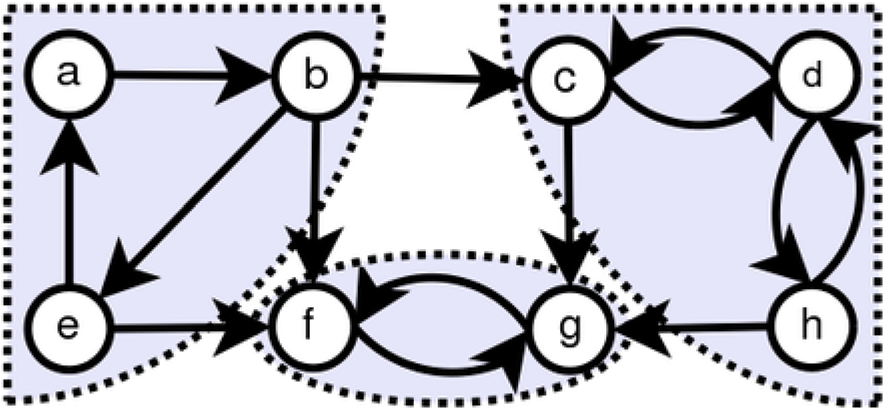


Copyright: pingdom.com

Copyright: HTTP archive

Bow-tie structure

- **Giant strongly connected component**: every node is reachable from every other node



Source: wikipedia

- **in-component**: those pages through which the GSCC can be reached
- **out-component**: those pages the GSCC can reach

Topical locality

Homophily in information networks

- the tendency of similar nodes to be connected
- "the capability to guess what a page is about by looking at the content of its neighbor pages"

Topical Locality:

Likeliness that a target page within a given distance from the source is about the same topic as the source, related to this happening by chance (i.e. how general the topic is)

- Text similarity is a proxy for topical relatedness
  - calculated with co-occurrence of keywords (cosine similarity)

Topical locality is the relationship between the structure of the information networks and the content of the nodes

Cosine similarity

- document *d* as high-dimensionality vector: each term in the vocabulary has a weight *w* in the vector
- weight typically corresponds to frequency of a term in the document
- compute the similarity between two documents by measuring the cosine between their vectors
- cosine close to 1: high similarity
- cosine close to 0: unrelated documents

Exercises: in-degree and out-degree

1. Go to google scholar and search for publications on **network science**. Pick two papers from the search results:
2. What is the in-degree of the paper in the citation network?
3. What is the out-degree (i.e. number of papers citing them)?