CS347

Lecture 12

May 21, 2001

Topics

- Web characterization
- Research Problems

The Web: A directed graph

- Nodes = static web pages (1+ billion)
- Edges = static hyperlinks (~10 billion)
- Web graph = Snapshot of web pages and hyperlinks
- Sparse graph: ~7 links/page on average
- Focus on graph structure, ignore content

Questions about the web graph

- How big is the graph? How many links on a page (outdegree)? How many links to a page (indegree)?
- Can one browse from any web page to any other? How many clicks?
- Can we pick a random page on the web?
  - Search engine measurement.

Questions about the web graph

- Can we exploit the structure of the web graph for searching and mining?
- What does the web graph reveal about social processes which result in its creation and dynamics?
- How different is browsing from a "random walk"?

Why?

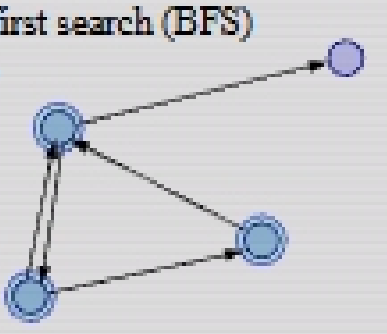
- Exploit structure for Web algorithms
  - Crawl strategies
  - Search
  - Mining communities
- Classification/organization
- Web anthropology
  - Prediction, discovery of structures
  - Sociological understanding

Web snapshots

- Altavista crawls (May 99/Oct 99/Feb 00)
- 220/317/500M pages
- 1.5/2.1B/5B hyperlinks
- Compaq CS2 connectivity server
  - back-link information
  - 10bytes/url, 3.4bytes/link, 0.15μs/access
  - given pages, return their in/out neighborhood

Algorithms

- Weakly connected components (WCC)
- Strongly connected components (SCC)
- Breadth-first search (BFS)
- Diameter



Challenges from scale

- Typical diameter algorithm:
  - number of steps ~ pages × links.
  - For 500 million pages, 5 billion links, even at a *very* optimistic 0.15μs/step, we need ~4 billion seconds.

  Hopeless.
- Will estimate diameter/distance metrics.

Scale

- On the other hand, can handle tasks linear in the links (5 billion) at a μs/step.
  - E.g., breadth-first search
- First eliminate duplicate pages/mirrors.
- Linear-time implementations for WCC and SCC.

May 1999 crawl

- 220 million pages after duplicate elimination.
- Giant WCC has ~186 million pages.
- Giant SCC has ~56 million pages.
  - Cannot browse your way from any page to any other
  - Next biggest SCC ~150K pages
- Fractions roughly the same in other crawls.

Tentative picture