CIS 833 – Information Retrieval and Text Mining

Lecture 16

Web Search & Crawling

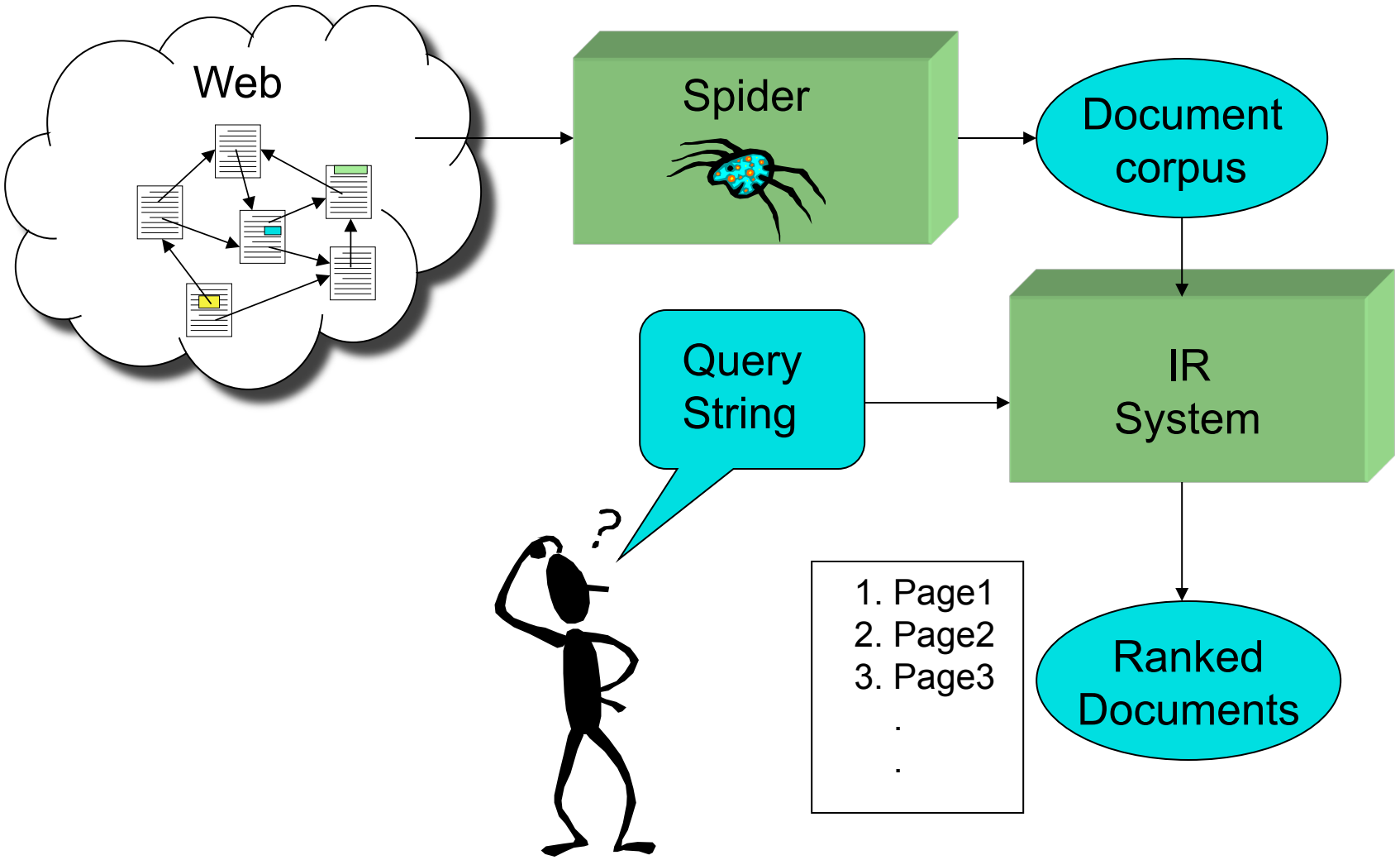
October 27, 2015

Credits for slides: Allan, Arms, Manning, Lund, Noble, Page.

Next

- Web Search
  - Textbook Chapter 19 – Web search basics
  - Textbook Chapter 20 – Web crawling
  - Textbook Chapter 21 – Web analysis
  - Monika R. Henzinger, Hyperlink Analysis for the Web. IEEE Internet Computing, vol. 5, no. 1, pp. 45-50, Jan/ Feb., 2001.

Web Search Using IR

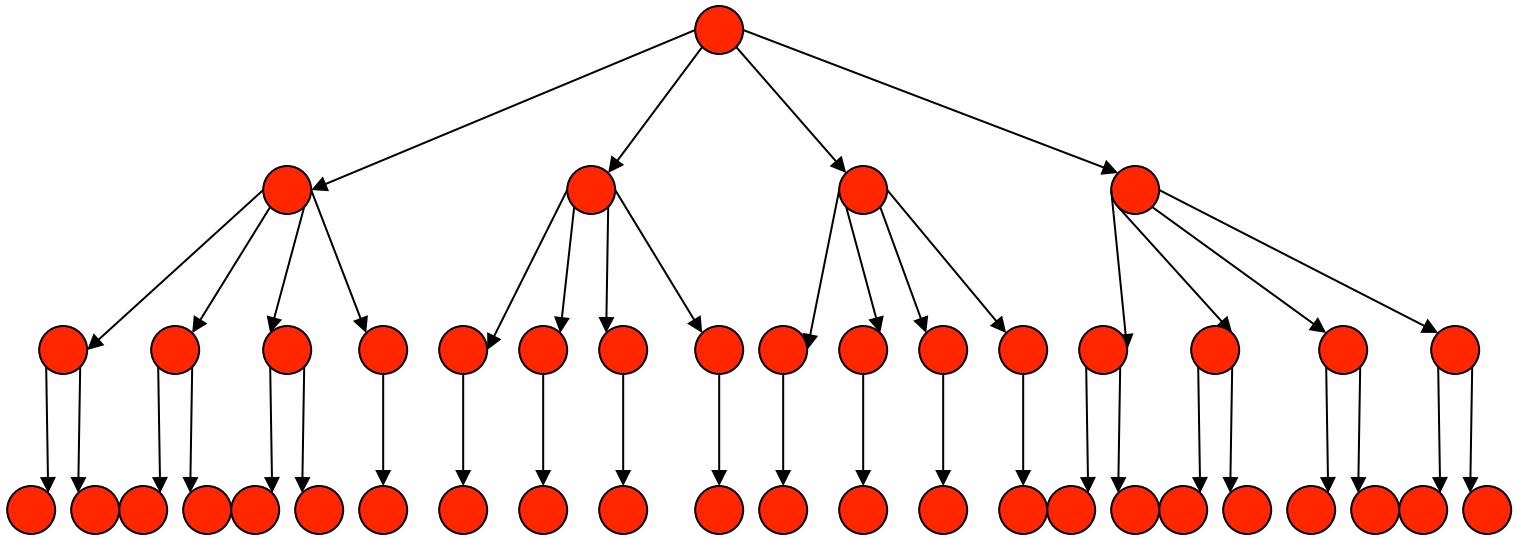


Spiders (Robots/Bots/Crawlers)

- Start with a comprehensive set of root URLs (seeds) from which to start the search.
- Follow all links on these pages recursively to find additional pages.
- Index/Process all **novel** found pages in an inverted index as they are encountered.
- May allow users to directly submit pages to be indexed (and crawled from).

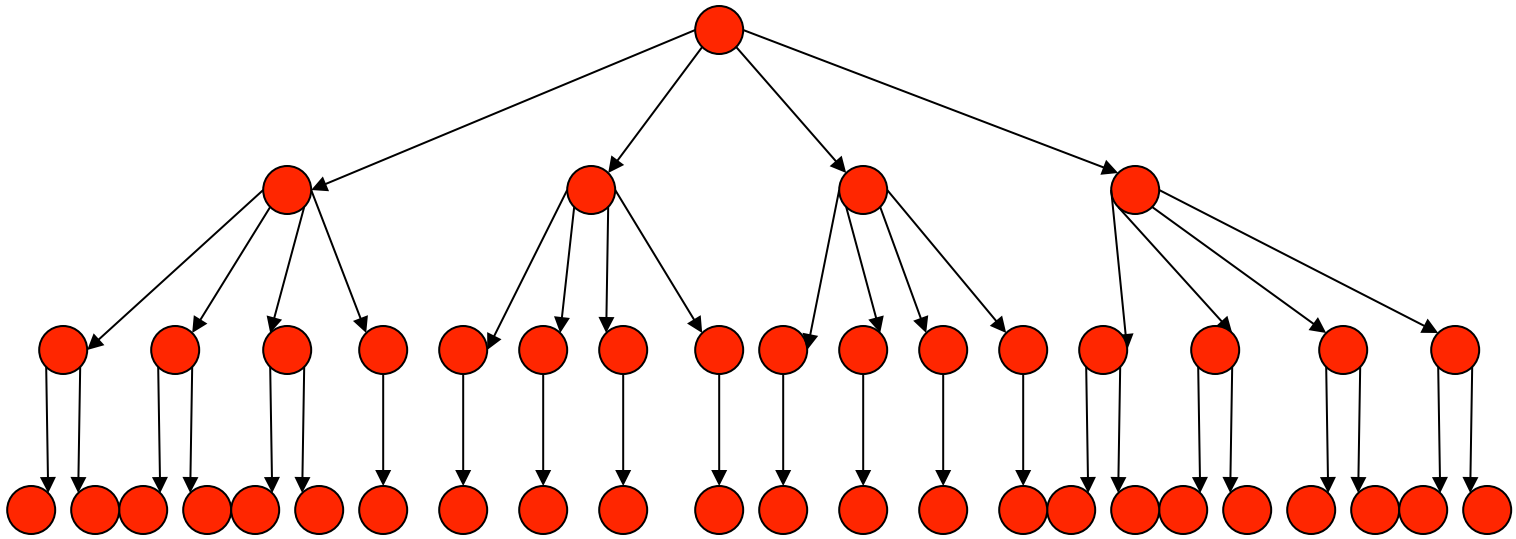
Search Strategies - BFS

Breadth-first Search (BFS)



Search Strategies - DFS

Depth-first Search (DFS)



Search Strategy Trade-Offs

- Breadth-first explores uniformly outward from the root page but requires memory of all nodes on the previous level (exponential in depth). Standard spidering method.
- Depth-first requires memory of only depth times branching-factor (linear in depth) but gets "lost" pursuing a single thread.
- Both strategies can be easily implemented using a queue of links (URLs).

Queueing Strategy

- How new links are added to the queue determines search strategy.
- FIFO (append to end of Q) gives breadth-first search.
- LIFO (add to front of Q) gives depth-first search.
- Heuristically ordering the Q gives a “focused crawler” that directs its search towards “interesting” pages.

Spidering Algorithm - BFS

Initialize queue (Q) with initial set of known URLs (seeds).
Until Q empty or page or time limit exhausted:
    Pop URL, L, from front of Q.
    If L is not to an HTML page (.gif, .jpeg, .ps, .pdf, .ppt…)
        continue loop.
    If already visited L, continue loop.
    Try to download page, P, for L.
    If cannot download P (e.g. 404 error, robot excluded)
        continue loop.
    Index P (e.g. add to inverted index or store cached copy).
    Parse P to obtain list of new links N.
    Append N to the end of Q.

Avoiding Page Duplication

- Must detect when revisiting a page that has already been spidered (web is a graph not a tree).
- Must efficiently index visited pages to allow rapid recognition test.
  - Tree indexing (e.g., trie)
  - Hashtable
- Index page using URL as a key.
  - Must canonicalize URLs (e.g., delete ending "/")
  - Not detect duplicated or mirrored pages
- Index page using textual content as a key.
  - Requires first downloading page

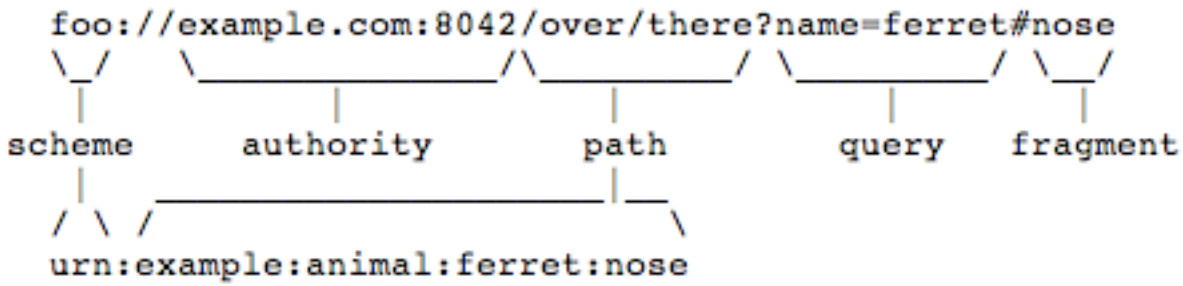
Link Extraction

- Must find all links in a page and extract URLs:
  -
  -

- Must complete relative URLs using current page URL:
  - to
    http://people.cis.ksu.edu/~dcaragea/teaching/CIS833
  - to
    http://people.cis.ksu.edu/~dcaragea/teaching/CIS833/syllabus.html

URL Syntax

- A URL has the following syntax:
  - <scheme>://<authority><path>?<query>#<fragment>
- A *query* passes variable values from an HTML form and has the syntax:
  - <variable>=<value>&<variable>=<value>…
- A *fragment* is also called a *reference* or a *ref* and is a pointer within the document to a point specified by an anchor tag of the form:
  - <A NAME="<fragment>">



http://www.rfc-editor.org/rfc/rfc3986.txt

Link Canonicalization

- Equivalent variations of ending directory normalized by removing ending slash.
  - http://people.cis.ksu.edu/~dcaragea/teaching/
  - http://people.cis.ksu.edu/~dcaragea/teaching
- Internal page fragments (references) removed:
  - http://people.cis.ksu.edu/~dcaragea/publications.html#Books
  - http://people.cis.ksu.edu/~dcaragea/publications.html

Anchor Text Indexing

- Extract anchor text (between <a> and) of each link followed.
- Anchor text is usually descriptive of the document to which it points.
  - Evil Empire
  - IBM

- Add anchor text to the content of the destination page to provide additional relevant keyword indices.
- Used by Google.

Anchor Text Indexing (*cont.*)

- Helps when descriptive text in destination page is embedded in image logos rather than in accessible text.
- Many times anchor text is not useful:
  - “click here”
- Increases content more for popular pages with many in-coming links, increasing recall of these pages.
- May even give higher weights to tokens from anchor text.

Restricting Spidering

- You can restrict spider to a particular site.
  - Remove links to other sites from Q.
- You can restrict spider to a particular directory.
  - Remove links not in the specified directory.
- Obey page-owner restrictions (robot exclusion).

Explicit and Implicit Politeness

- Explicit politeness: specifications from webmasters on what portions of site can be crawled
  - robot exclusion
- Implicit politeness: even with no specification, avoid hitting any site too often
  - common heuristic: insert time gap between successive requests to a host that is >> time for most recent fetch from that host

Robot Exclusion

- Two components:
  - Robots Exclusion Protocol: Site wide specification of excluded directories – robots.txt.
  - Robots META Tag: Individual document tag to exclude indexing or following links.

Robots.txt

- Protocol for giving spiders (“robots”) limited access to a website (originally from 1994)
  - www.robotstxt.org/
- Website announces its request on what can(not) be crawled

Robots Exclusion Protocol

- Site administrator puts a "robots.txt" file at the root of the host's web directory to specify access restrictions.
  - http://www.ebay.com/robots.txt
  - http://www.cnn.com/robots.txt
  - http://www.k-state.edu/robots.txt
- File is a list of excluded directories for a given robot (user-agent).
  - Exclude all robots from the entire site:

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

Robot Exclusion Protocol Examples

- Exclude specific directories:

```
User-agent: *
Disallow: /tmp/
Disallow: /cgi-bin/
Disallow: /users/paranoid/
```

- Exclude a specific robot:

```
User-agent: GoogleBot
Disallow: /
```

- Allow a specific robot:

```
User-agent: GoogleBot
Disallow:
```

Robot Exclusion Protocol Details

- Only use blank lines to separate different User-agent disallowed directories.
- One directory per “Disallow” line.
- No regular expression can be used as patterns for directories.

Robots META Tag

- Include META tag in HEAD section of a specific HTML document.
  - <meta name="robots" content="none">
- Content value is a pair of values for two aspects:
  - index | noindex: Allow/disallow indexing of this page.
  - follow | nofollow: Allow/disallow following links on this page.
- Special values:
  - all = index,follow and none = noindex,nofollow
- Examples
  - <meta name="robots" content="noindex,follow">
  - <meta name="robots" content="index,nofollow">

Robot Exclusion Issues

- META tag is newer and less well-adopted than “robots.txt”.
- Standards are conventions to be followed by “good robots.”
- Companies have been prosecuted for “disobeying” these conventions and “trespassing” on private cyberspace.

Multi-Threaded Spidering

- Bottleneck is network delay in downloading individual pages.
- Best to have multiple threads running in parallel each requesting a page from a different host.
- Distribute URLs to threads to guarantee equitable distribution of requests across different hosts to maximize throughput and avoid overloading any single server.
- Early Google spider had multiple co-ordinated crawlers with about 300 threads each, together able to download over 100 pages per second.

Directed/Focused Spidering

- Sort queue to explore more “interesting” pages first.
- Two styles of focus:
  - Topic-Directed
  - Link-Directed

Topic-Directed Spidering

- Assume desired topic description or sample pages of interest are given.
- Sort queue of links by the similarity (e.g., cosine metric) of their source pages and/or anchor text to this topic description.
  - Related to Topic Tracking and Detection

Link-Directed Spidering

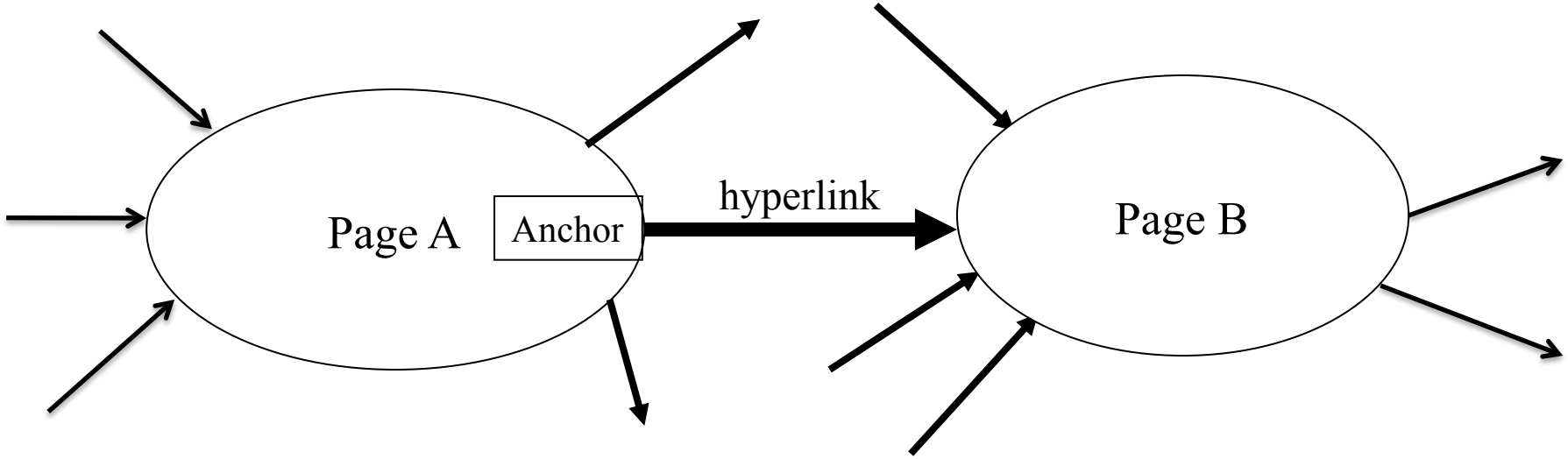
- Monitor links and keep track of in-degree and out-degree of each page encountered.
- Sort queue to prefer popular pages with many in-coming links (*authorities*).
- Sort queue to prefer summary pages with many out-going links (*hubs*).

Keeping Spidered Pages Up to Date

- Web is very dynamic: many new pages, updated pages, deleted pages, etc.
- Periodically check spidered pages for updates and deletions:
  - Just look at header info (e.g., META tags on last update) to determine if page has changed, only reload entire page if needed.
- Track how often each page is updated and preferentially return to pages which are historically more dynamic.
- Preferentially update pages that are accessed more often to optimize freshness of more popular pages.

Link Analysis

The Web as a Directed Graph



Assumptions:

A hyperlink between pages denotes

- author perceived relevance (quality signal) and/or
- similar topic

The anchor of the hyperlink

- describes the target page (textual context)

Bibliometrics: Citation Analysis

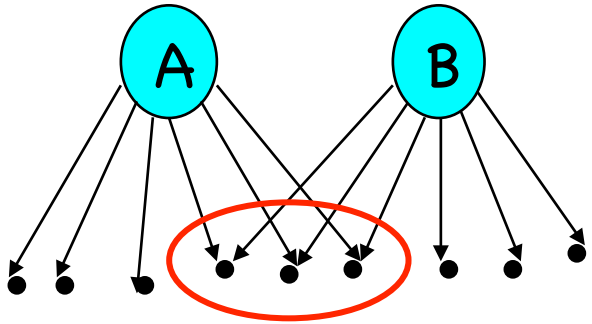
- Many standard documents include *bibliographies* (or *references*), explicit *citations* to other previously published documents.
- Using citations as links, standard corpora can be viewed as a graph.
- The structure of this graph, independent of content, can provide interesting information about the similarity of documents and the structure of information.

Impact Factor

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals.
- Measure of how often papers in the journal are cited by other scientists.
- Computed and published annually by the Institute for Scientific Information (ISI).
- The *impact factor* of a journal *J* in year *Y* is the average number of citations (from indexed documents published in year *Y*) to a paper published in *J* in year *Y*-1 or *Y*-2.
- Does not account for the quality of the citing article.

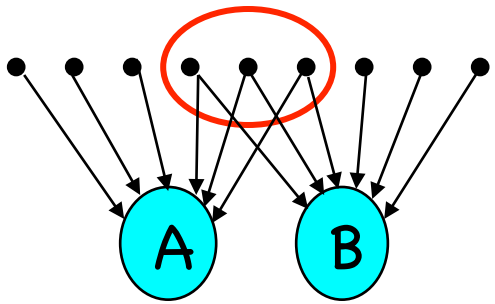
Bibliographic Coupling

- Measure of similarity of documents introduced by Kessler in 1963.
- The bibliographic coupling of two documents *A* and *B* is the number of documents cited by *both A* and *B*.
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



Co-Citation

- An alternate citation-based measure of similarity introduced by Small in 1973.
- Number of documents that cite both *A* and *B*.
- Maybe want to normalize by total number of documents citing either *A* or *B?*



Citations vs. Links

- Web links are a bit different than citations:
  - Many links are navigational.
  - Many pages with high in-degree are portals not content providers.
  - Not all links are endorsements.
  - Company websites don't point to their competitors.
  - Citations to relevant literature is enforced by peer-review.