Link Analysis

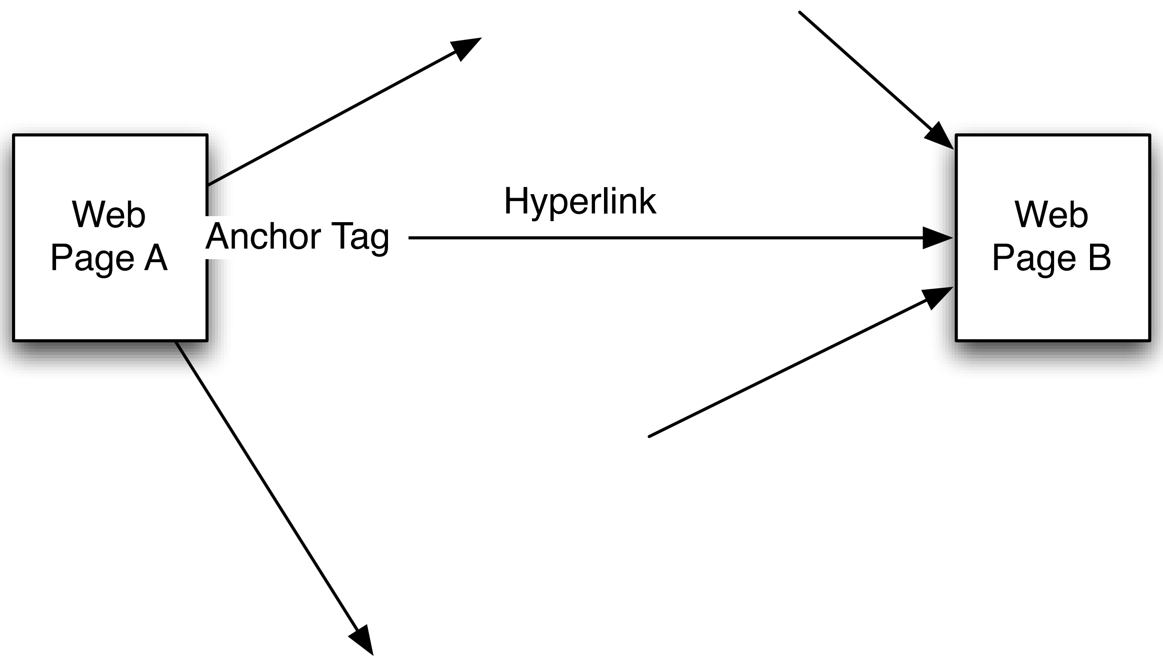
Introduction to Information Retrieval
Informatics 141 / CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Outline

- The web as a directed graph

The web as a directed graph



- Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)
- Assumption 2: The anchor of the hyperlink describes the target page (textural context)

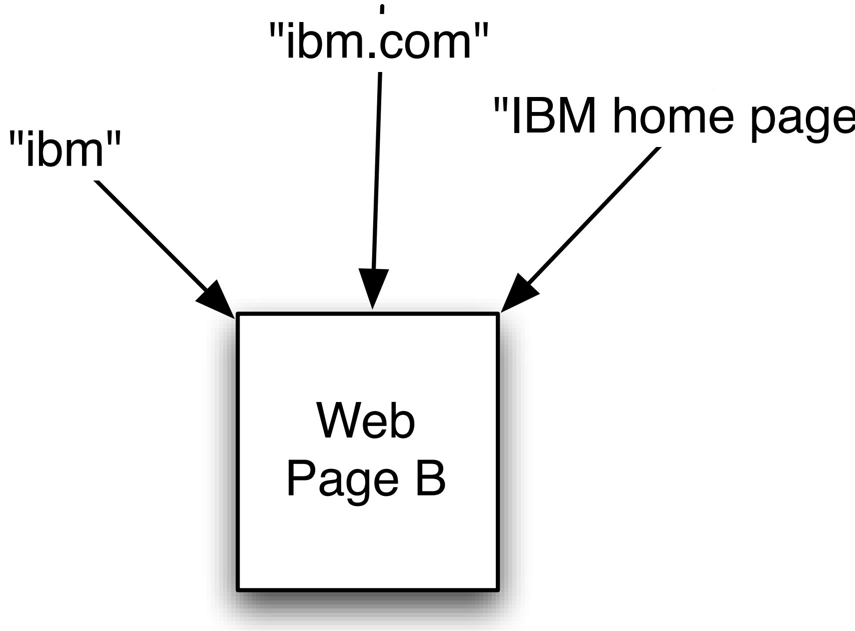
The web as a directed graph

- Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)
- Assumption 2: The anchor of the hyperlink describes the target page (textural context)

- Where might these assumptions not hold?

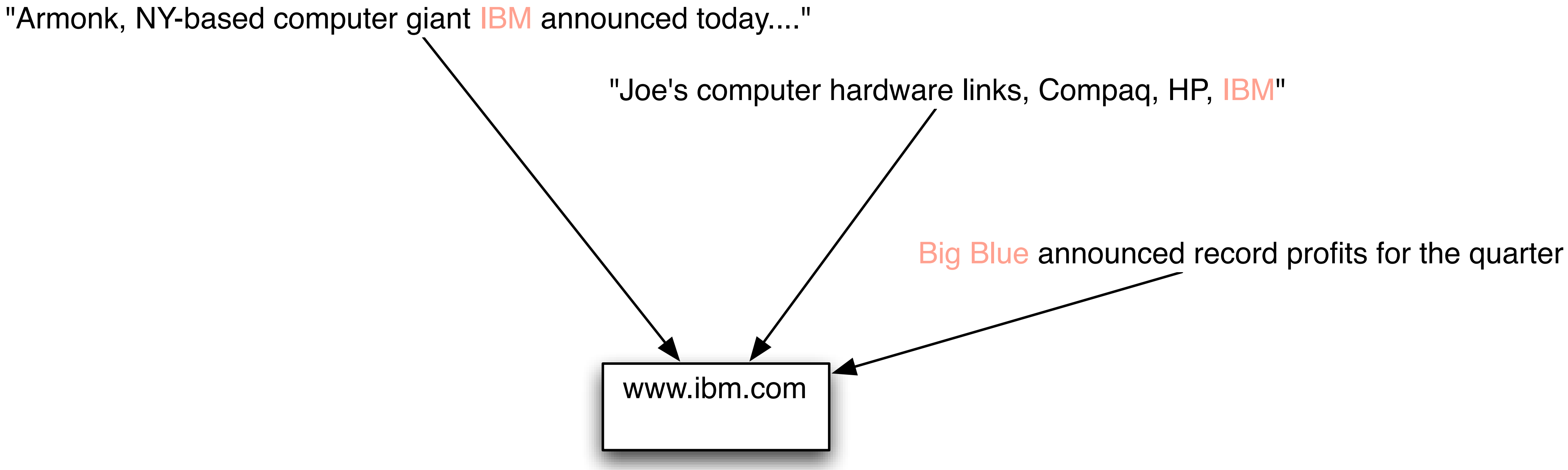
The web as a directed graph

- Anchor Text
  - WWW Worm -McBryan94
- For IBM how do you distinguish between
  - IBM's home page (mostly graphics)
  - IBM's copyright page (high TF for "ibm)
  - Rival spam page (high TF for "ibm")
  - ?
- A million pieces of anchor text with "ibm" send a strong signal



Indexing anchor text also

- When indexing a document D
  - include anchor text from links pointing to D



"Armonk, NY-based computer giant IBM announced today...."

"Joe's computer hardware links, Compaq, HP, IBM"

Big Blue announced record profits for the quarter

www.ibm.com

Indexing anchor text

- Anchor text is often a better description of a page’s content than the page itself.
- Can be weighted more highly than the text
  - If enough anchor text is available
  - Same technique as zone weighting
    - create a “zone” for anchor text
- Indexing anchor text can have unexpected side effects
  - Google bombs, miserable failure
  - nigritude ultramarine follow-on

Indexing anchor text

- Anchor text is often a better description of a page’s content than the page itself.
- Can be weighted more highly than the text
  - If enough anchor text is available
  - Same technique as zone weighting
    - create a “zone” for anchor text
- Indexing anchor text can have unexpected side effects
  - Google bombs, miserable failure
  - nigritude ultramarine follow-on

Anchor text

- Other applications
  - Weighting links in the graph
  - Generating page descriptions from anchor text

PageRank

- Citation analysis:
  - Analysis of citations in the scientific literature
  - Example citation:
    - “Miller (2001) has shown that physical activity alters the metabolism of estrogens”

PageRank

- Two ways of measuring similarity of two scientific articles:
  - Cocitation similarity: The two articles are cited by the same articles
  - Bibliographic coupling similarity: The two articles site the same articles

next...

LUCI