CS347

Lecture 7

April 30, 2001

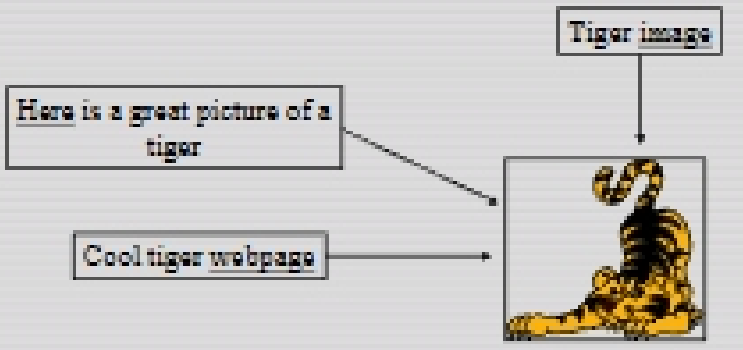
Topics du jour

- Finish up web ranking
- Peer-to-peer search
- Search deployment models
  - Service vs. software
  - External vs. internal-facing search software
- Review of search topics

Tag/position heuristics

- Increase weights of terms in titles
- Increase weights of terms in <h > tags
- Increase weights of terms near the beginning of the doc, its chapters and sections - key phrases

Anchor text



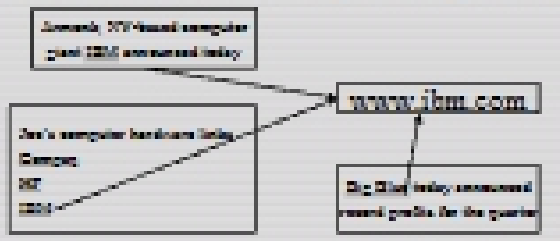
The text in the vicinity of a hyperlink is descriptive of the page it points to.

Two uses of anchor text

- When indexing a page, also index the anchor text of links pointing to it.
- To weight links in the hubs/authorities algorithm from the last lecture.
- Anchor text usually taken to be a window of 6-8 words around a link anchor.

Indexing anchor text

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



Indexing anchor text

- Can sometimes have unexpected side effects - *e.g., evil empire*.
- Can index anchor text with less weight.

Weighting links

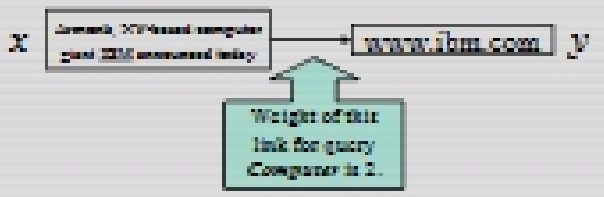
- In hub/authority link analysis, can match anchor text to query, then weight link.

$$h(x) \leftarrow \sum_{x \quad y} a(y) \qquad a(x) \leftarrow \sum_{y \quad x} h(y)$$

$$\Rightarrow \quad h(x) = \sum_{x \quad y} w(x,y) \cdot a(y) \qquad a(x) = \sum_{y \quad x} w(x,y) \cdot h(y)$$

Weighting links

- What is $w(x,y)$?
- Should increase with the number of query terms in anchor text.
  - Say 1+ number of query terms.



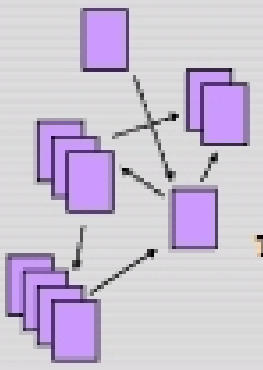
Weighted hub/authority computation

- Recall basic algorithm:
  - Iteratively update all $h(x)$, $a(x)$;
  - After iteration, output pages with highest $h()$ scores as top hubs; highest $a()$ scores as top authorities.
- Now use weights in iteration.
- Raises scores of pages with "heavy" links.

Do we still have convergence of scores? To what?

Web sites, not pages

- Lots of pages in a site give varying aspects of information on the same topic.



Treat portions of web-sites as a single entity for score computations.

Link neighborhoods

- Links on a page tend to point to the same topics as neighboring links.
  - Break pages down into *pagelets* (say separate by tags) and compute a hub/authority score for each pagelet.