PageRank

CISC489/689-010, Lecture #20

Wednesday, April 29th

Ben Carterette

Web Search

- Problem:
  - Web search engines easily hacked
  - I want to sell something; I'll just add a few popular keywords to my page over and over and over again
  - All the retrieval models we've discussed will score that page higher for those keywords
- Other problems:
  - No hackers, but top-ranked pages are coming from deep within a site, or from pages that change often, or pages about very obscure topics
  - Not really useful

Possible Solution

- Leverage link structure
- Maybe if many pages are linking to a page, that page is more "important"
- Idea:
  - Count the number of inlinks to the page
  - Assign it "importance" based on that number
- Any problem with this?

Hacking Link Counts

- I can just make a bunch of pages that link to my spam page
- Inlink count will be high even though my page is not important
- Better idea:
  - Recursively use the importance of the linking pages when calculating the importance of the page

PageRank

- Google's PageRank is probably the best known algorithm
- Intuitive idea: "random surfer" model
  - If you start on a random page on the internet and just start clicking links randomly,
  - What is the probability you will land on page u?
  - If one page has a higher landing probability, the pages it links to have higher landing probabilities as well
  - Higher probability = more authority = better PageRank

Illustration