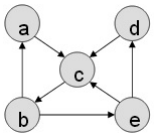
CSCI 104L Lecture 26: Extra Topics

PageRank

The key insight that resulted in the dominance of the Google Search Engine was to think of the world wide web as a graph.

- If page A links to page B, this is an endorsement from A of B.
- Not all links are created equal. If A links to B, to determine how good B is requires us to first know how good A is. Sounds circular, but there is a nice solution to the problem.
- Number of links is important. If A and B are equally good pages, but A has one outgoing link to C, and B has 10,000 outgoing links, one of them to D, we would say that the link from A to C confers more support (since it was chosen more carefully).



rA = 0.5*rB
rB = rC
rC = rA+rD+0.5*rE
rD = 0.5*rE
rE = 0.5*rB
rA+rB+rC+rD+rE=1

To solve this system of equations requires lots of linear algebra, and the actual PageRank has a few extra optimizations. There is a simpler way to explain the solution without linear algebra:

- You have a web-surfer who starts at a page chosen uniformly at random, and chooses an outgoing link uniformly at random.
- If the surfer gets stuck (no outgoing links) they move to a page chosen uniformly at random. The surfer can get stuck in a less obvious trap (two pages linking only to each other), so there is usually a small probability (around 15%) that at each step the surfer moves to a page chosen uniformly at random.
- We can iteratively calculate the probability the surfer is at any given page at any given step. At the first step, each page is equally likely. We use those values to calculate the probability the surfer is at any given page on the second step, and repeat.
- Typically this process will stop at some small number of iterations (such as 20), but in theory you would want to find the limit as the number of iterations approaches infinity.

This system is called PageRank, and started as a research project by two grad students at Stanford. It was so much better than the competition, that it quickly became a serious product that made them filthy stinkin' rich.