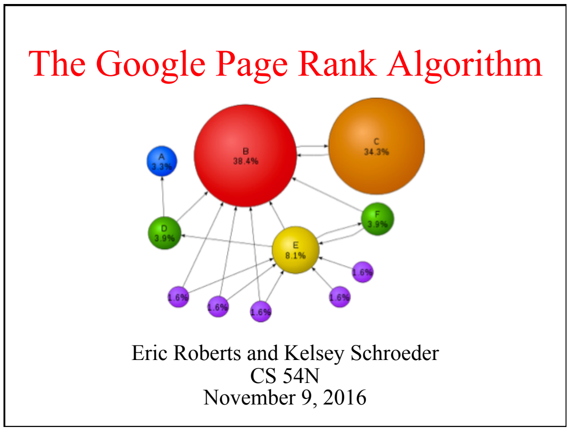
Eric Roberts
CS 54N

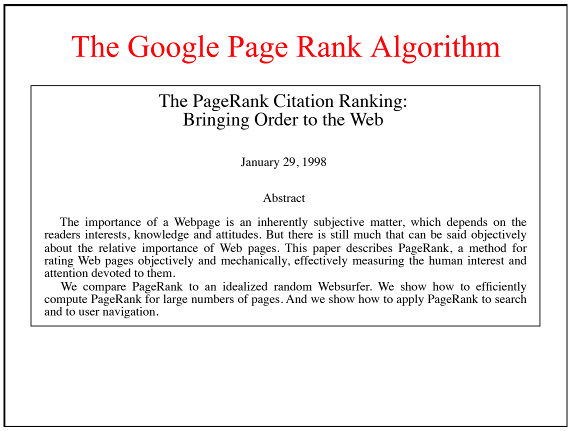
Handout #24
November 9, 2016

The Google PageRank Algorithm



The Google Page Rank Algorithm

Eric Roberts and Kelsey Schroeder
CS 54N
November 9, 2016



The Google Page Rank Algorithm

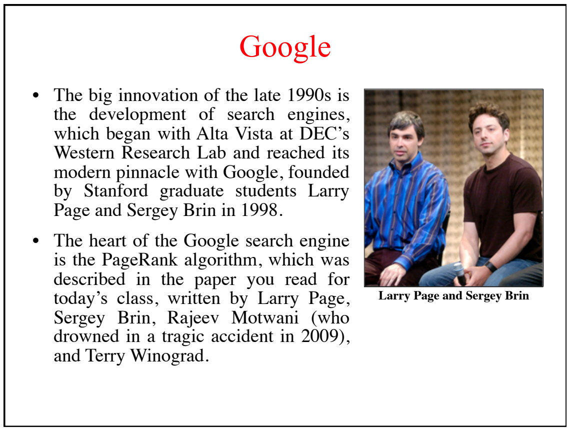
The PageRank Citation Ranking:
Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Webpage is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

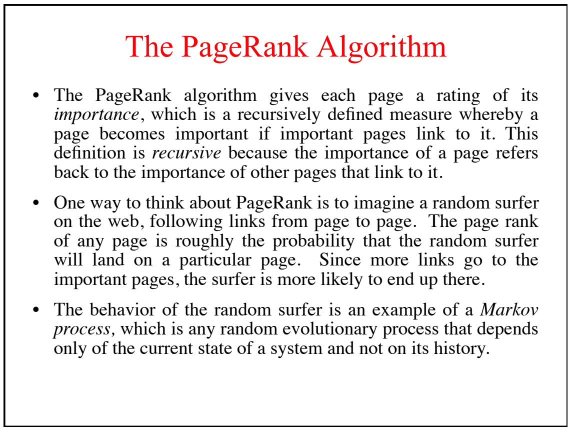
We compare PageRank to an idealized random Websurfer. We show how to efficiently compute PageRank for large numbers of pages. And we show how to apply PageRank to search and to user navigation.



Google

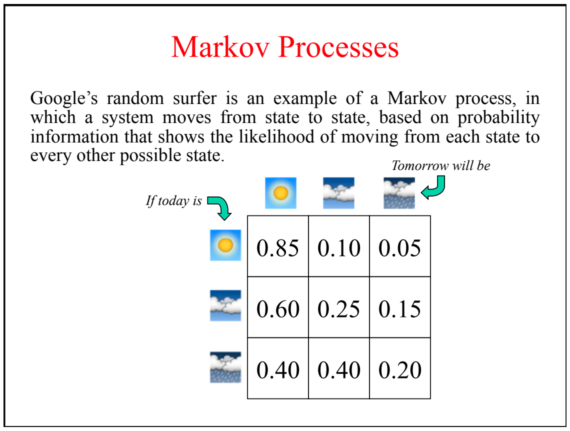
- The big innovation of the late 1990s is the development of search engines, which began with Alta Vista at DEC's Western Research Lab and reached its modern pinnacle with Google, founded by Stanford graduate students Larry Page and Sergey Brin in 1998.
- The heart of the Google search engine is the PageRank algorithm, which was described in the paper you read for today's class, written by Larry Page, Sergey Brin, Rajeev Motwani (who drowned in a tragic accident in 2009), and Terry Winograd.

Larry Page and Sergey Brin



The PageRank Algorithm

- The PageRank algorithm gives each page a rating of its *importance*, which is a recursively defined measure whereby a page becomes important if important pages link to it. This definition is *recursive* because the importance of a page refers back to the importance of other pages that link to it.
- One way to think about PageRank is to imagine a random surfer on the web, following links from page to page. The page rank of any page is roughly the probability that the random surfer will land on a particular page. Since more links go to the important pages, the surfer is more likely to end up there.
- The behavior of the random surfer is an example of a *Markov process*, which is any random evolutionary process that depends only of the current state of a system and not on its history.

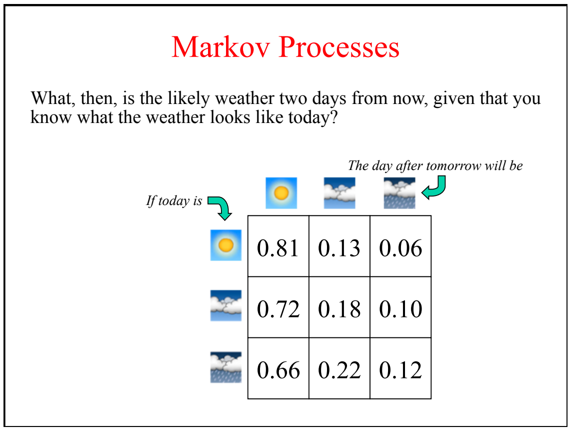


Markov Processes

Google's random surfer is an example of a Markov process, in which a system moves from state to state, based on probability information that shows the likelihood of moving from each state to every other possible state.

If today is / *Tomorrow will be*

	☀	⛅	🌧
☀	0.85	0.10	0.05
⛅	0.60	0.25	0.15
🌧	0.40	0.40	0.20



Markov Processes

What, then, is the likely weather two days from now, given that you know what the weather looks like today?

If today is / *The day after tomorrow will be*

	☀	⛅	🌧
☀	0.81	0.13	0.06
⛅	0.72	0.18	0.10
🌧	0.66	0.22	0.12

Markov Processes

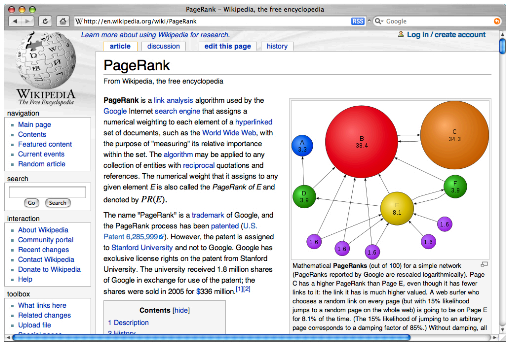
What if you then repeat the process for ten days?

Ten days from now will be

If today is

	☀	⛅	🌧
☀	0.77	0.14	0.07
⛅	0.77	0.14	0.07
🌧	0.77	0.14	0.07

Google’s PageRank Algorithm



PageRank – Wikipedia, the free encyclopedia

http://en.wikipedia.org/wiki/PageRank

Learn more about using Wikipedia for research.

PageRank

From Wikipedia, the free encyclopedia

PageRank is a link analysis algorithm used by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is also called the *PageRank of E* and denoted by $PR(E)$.

The name "PageRank" is a trademark of Google, and the PageRank process has been patented (U.S. Patent 6,285,999). However, the patent is assigned to Stanford University and not to Google. Google has exclusive license rights on the patent from Stanford University. The university received 1.8 million shares of Google in exchange for use of the patent; the shares were sold in 2005 for $336 million.[1][2]

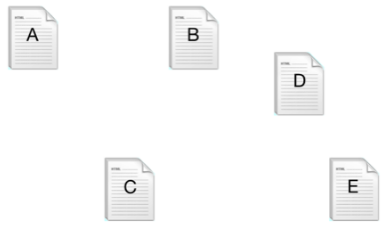
Contents [hide]

1 Description

Mathematical **PageRanks** (out of 100) for a simple network (PageRanks reported by Google are rescaled logarithmically). Page C has a higher PageRank than Page E, even though it has fewer links to it: the link it has is much higher valued. A web surfer who chooses a random link on every page (but with 15% likelihood jumps to a random page on the whole web) is going to be on Page E for 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all

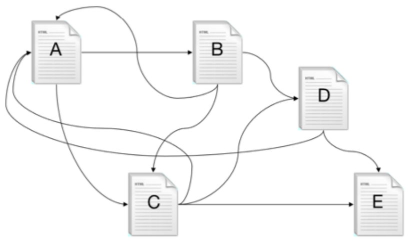
The Page Rank Algorithm

1. Start with a set of pages.



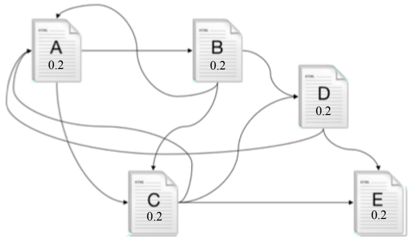
The Page Rank Algorithm

2. Crawl the web to determine the link structure.



The Page Rank Algorithm

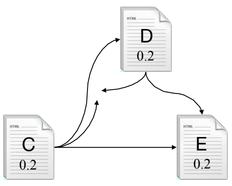
3. Assign each page an initial rank of 1 / *N*.



The Page Rank Algorithm

4. Successively update the rank of each page by adding up the weight of every page that links to it divided by the number of links emanating from the referring page.

- In the current example, page E has two incoming links, one from page C and one from page D.
- Page C contributes 1/3 of its current page rank to page E because E is one of three links from page C. Similarly, page D offers 1/2 of its rank to E.

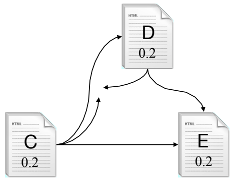


- The new page rank for E is

$$\mathrm{PR(E)} = \frac{\mathrm{PR(C)}}{3} + \frac{\mathrm{PR(D)}}{2} = \frac{0.2}{3} + \frac{0.2}{2} \approx 0.17$$

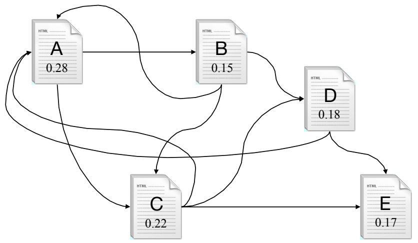
The Page Rank Algorithm

5. If a page (such as E in the current example) has no outward links, redistribute its rank equally among the other pages in the graph.
   - In this graph, 1/4 of E's page rank is distributed to pages A, B, C, and D.
   - The idea behind this model is that users will keep searching if they reach a dead end.



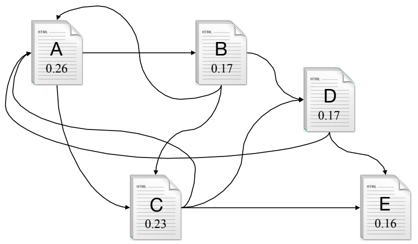
The Page Rank Algorithm

7. Apply this redistribution to every page in the graph.



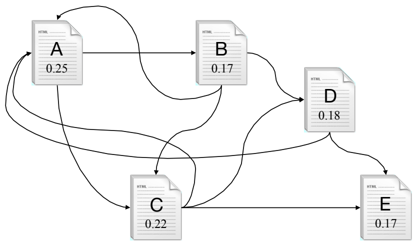
The Page Rank Algorithm

8. Repeat this process until the page ranks stabilize.



The Page Rank Algorithm

9. In practice, the Page Rank algorithm adds a *damping factor* at each stage to model the fact that users stop searching.



PageRank as a Two-Player Game

- One of the challenges for the designers of any search engine is ensuring that a commercial interest can't artificially increase its ranking by creating many others pages whose only purpose is to link to that company's home page.
- Adopting the PageRank algorithm makes it harder for authors to manipulate the system because the ranking of a page depends on the prestige of important pages that are typically outside the control of those who are seeking to game the system.
- Preventing users from manipulating their own web rankings is an ongoing problem for all search engine companies. To help ensure that the rankings remain fair, Google must keep the details of the ranking algorithms secret and change them often enough to outwit the would-be saboteurs.

Exercise: Quoted Word Sequences

- In the movie *Enigma*, Claire Romilly first meets Tom Jericho on a train while she is solving a cryptic crossword. She muses aloud about the clue—*Roast mules go topsy-turvy*—and Tom provides the answer.
- When you enter a set of search terms, Google allows you to search for a sequence of consecutive words by enclosing those words in quotation marks. In this example, searching for roast or mules is useless; searching for the quoted string "roast mules" brings the answer up immediately.
- Given that indexing all pairs of words would be prohibitively expensive in terms of storage, how can Google make this feature work?
- Hint: In addition to the URLs of the pages on which a search term appears, Google records the position on that page.