The PageRank Citation Ranking: Bringing Order to the Web

wbg231

January 2023

1 abstract

- the importance of webpages is subjective,
- this paper is about page rank, a method for rating web pages, with math and measuring human interest in them
- compare page rank to a random web surger

introduction

- the web is big in diverse, there are also many types of users so hard to recommend search pages efficiently
- bit the web has meta data including link information that can be leveraged into structure of the web

diversity of pages

- webpages have really diverse uses, quality of information
- also there is profit incentives are people will try to play the system must be aware of that when building a search engine

page rank

- page rank is a system to compute the relative importance of every web page based on the graph of the web

ranking every page on the web

link structure of the web

- the web is a graph of forward links and incoming ie backwards link
- we can never know all backwards links to a page but we can know all forward links form it
- just counting the number of back nodes to a page is not a good measure of importance
- a page has a high rank if the sum of the ranks of it's backlinks are high.
- this covers both caes when a page has many backlinks and when a page has a few highly ranked back links

definition of page rank

- let u be a webpage $F_u and B_u$ be the set of forward and backward links pointing to that page
- let rank be defined as $R(u) = \sum_{v \in B_u} \frac{R_v}{N_v}$ so it an average of th rank of bank links
- this is a recursive formula
- if we put this in a matrix (that we first make a probaility matrix) from and iterate over a state vector, we end up with a steady state that coresponds to eigenvector coresponding to eigenvalue 1
- this can be throught of as a random surfer randomly going from one page to another
- additionally we add a parameter that jumps the surfer randomly to another point in the graph to avoid loops
- this steady state vector (is a probaility vector that can then directly rank the importance of pages)

dangeling links

- these are pages with no outgoing links
- we remove dangeling links from the system until we have a steady state, then we know the rank of pages that lead to them at least, then we add them back as links from there outgoing page (which must be renormalized)

searching with page rank

- page rank does well with underspecfied queries
- title matching and page rank work well togther

personalized page rank

- this can be achived by changing the teleportation parameter to send users to pages we know they like to visit or have interacted with a lot in the past
- there is a textbook reading but it more or less covers the same material as this so i am going to skip it for now.