Page Rank

Dreycey Albin

September 2019

The key need for the page rank algorithm occurred with the birth of the internet, where scalable sorting algorithms for websites were needed. Current search engines algorithms at the time were focused on concepts learned from research on citations for publications, which neglected important information inherent to website popularity. At the time, most sorting algorithms based popularity on the number of links to a particular page. This was a problem because there could be websites spawned purely for the purpose of increasing another websites rank, or likewise, the links could be stemming from websites without visitors [2].

The key solution given in the page rank algorithm is a method that still counts incoming links, but it accounts for the popularity of the linker website (or parent node). It does this by using a recursive approach to count all of the links that website has (again, accounting for the weight of those websites too). This traces back through a network of linked websites until converging, where the number of weighted incoming edges is known for all websites in the network. This effectively gets around the pitfalls of the current approaches at the time [2].

There were a couple drawbacks mentioned, and ideas for getting around these drawbacks was also proposed. One such flaw is processing the dangling links. The dangling links have to be re-added once the algorithm has converged, which requires them to iterate as many times as it took to get rid of them in the first place. This was estimated to take. around 5 hours in the implementation published in the paper. One potential work around given in the paper was the ability to add more strict convergence criteria to speed the calculation [2].

It's interesting to read such a historic piece of work. Not because the idea was ground breaking at the time (honestly, it's a great idea, but seemed like it came right on time- not "before" it's time), but because it was the beginning of Google. I didn't realize this is where Google started, nor did I realize the paper worked as a way to popularize the Google search engine. I had originally assumed the paper was strictly focused on the algorithm, and that Google came some time after.

The paper by Ivan and Grolmusz (2011) explores the possibility of using the page rank algorithm on biological data. They test the algorithm on to different objectives: (1) metabolic networks in *Mycobacterium tuberculosis*; (2) Proteins involved in Melanoma. Both of these applications returned expected results, which validated the use of the algorithm in this biomedical domain, but there were additional unexpected findings that could give insight into novel findings for both applications. My favorite part of the paper would be their justification for the algorithm using a stability estimation for the page rank algorithm, as most biological data is filled with both false positives and false negatives [1].

References

[1] Gábor Iván and Vince Grolmusz. When the web meets the cell: using personalized pagerank for analyzing protein interaction networks. *Bioinformatics*, 27(3):405–407, 2010.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.