CS 549 - Paper Review on "The PageRank Citation Ranking: Bringing Order to the Web" (Page et al., 1999)

Ga Young Lee

April 1, 2021

1 What is the problem discussed in the paper?

With a myriad of available websites online, it is crucial to show the more relevant search results to users since their attention is limited. In this paper, Page and at al. (1993) suggest that the number of links connected to a website can be used as a proxy of its importance denoting its relative ranking as "link popularity."

2 Why is it important?

PageRank effectively utilizes a graph algorithm to show a systematic method of organizing a large amount of information in conjunction with related data. Therefore, PageRank's impact expands beyond search, informational retrieval, and data mining. For instance, Pagerank-index (Pi) is widely used as a measure of researchers' impact on their community as it reflects the underlying citations and collaboration within their networks.

3 What are the main ideas of the proposed solution for the problem?

Basically, the algorithm returns a probability distribution between 0 to 1 to show the likelihood of a user randomly searching for information online and arriving at a specific link considering the structure of a given network. To give a high-level overview of the network, each link is a node that is connected to other links with a weighted edge. The weight is determined by the number of links connected to a node and its significance in the network, which is denoted as "PageRank." The benefit of this approach is that the connectivity of the network can be represented in a matrix which is useful for calculating multiple iterations recursively.

4 What are the weaknesses of PageRank?

Despite its impact on numerous applications on the internet, PageRank has a couple of weaknesses; first, the algorithm fails to consider the recency of the information presented in the network. Specifically, the content of a webpage might change dynamically as users edit the content. However, this update is not reflected in the static probabilistic network that is built on the initial relations of the web pages. Therefore, PageRank requires multiple "reruns" of the algorithm to truly map out the internet, which can be computationally expensive. Additionally, the search space of PageRank is limited, so when a user introduces a word that is out of the scope, PageRank might not be able to query the proper information and build a graph. In short, PageRank can perform very well on a given dataset that its interconnectedness doesn't change over time, but its pitfall lies in its inability to consider dynamic changes that are inevitably occurring in the network.