Link Analysis

Lecture – 75

The Web Graph

So, in the 1990s, when the internet was booming with web pages that were linked to each other through hyperlinks, it became challenging to identify the best page with authoritative information. To tackle this problem, the speaker suggested hiring thousands of people to rate each web page and add relevant keywords to a database. Whenever someone searched for a keyword, the highest-rated page with that keyword would be displayed. However, this idea was not scalable. The much better way to tackle this problem paved the way for the birth of Google. What was this much better way?

Link Analysis

Lecture - 76

Collecting the Web Graph

The technique proposed for collecting the web graph is to take a random walk on the web by imagining a city that has never been visited before. Assume that the person is given a car, and there is no traffic in the city. They randomly take turns at every intersection, choosing a direction uniformly at random, and keep going for a period of time, say one month. At the end of one month, they will have covered almost all the roads in the city.

Similarly, the person can explore the web graph by taking a random walk on the web. The graph can be generated randomly with some probability p and n vertices. The person starts at a randomly chosen vertex and starts walking on the graph randomly. The question is when they will visit all the vertices in the graph. The answer depends on how well connected the graph is, and how big it is.

To estimate the time taken to visit all the vertices, we can write a Python code that performs multiple random walks and checks how many of them reach all the vertices. By increasing the number of random walks, we can get a better estimate of the expected time taken to visit all the vertices.

Link Analysis

Lecture - 77

Equal Coin Distribution

The strategy described in the lecture is known as the "PageRank" algorithm, which was initially developed by Google founders Larry Page and Sergey Brin to rank web pages. The algorithm assigns a score to each web page based on the importance of the pages that link to it. The more important a page that links to a particular page is, the more "PageRank" the page being linked to will have.

In the lecture, this algorithm was demonstrated using a social network of 30 people. Each person is given 100 gold coins and they are asked to distribute them to their friends. Each person distributes their coins equally among their friends, and this process is repeated multiple times until a stable state is reached, where the amount of gold coins each person has does not change anymore. The person with the most gold coins is considered to be the most important person in the network.

The intuition behind this algorithm is that important people are likely to have more connections to other important people, so they will receive more gold coins from their friends who are also important. In the context of the web graph, web pages that are linked to by many other important pages are likely to be more important themselves. By applying this algorithm to the web graph, Google is able to provide more relevant search results to its users.

Lecture - 78

Random Coin Dropping

The abbreviation GNU stands for "GNU's Not Unix". It is a recursive acronym which means that the abbreviation itself refers back to itself in a self-referential manner. It was created by Richard Stallman and the GNU Project to develop a free and open-source operating system similar to Unix. The name GNU represents the idea of creating a completely free software system that could be modified and shared by anyone.

Lecture – 79

Google Page Ranking Using Web Graph

The web graph is a network of web pages that are related by hyperlinks.

Google uses the web graph to provide relevant search results.

Google crawls the web by taking random walks and dropping coins on each node/page it visits.

The coins represent points that are accumulated by each node/page.

When a user searches for a keyword, Google retrieves all relevant pages and sorts them in descending order of the points accumulated by each page.

Larry Page and Sergey Brin created Google by implementing this algorithm of crawling the web and giving points to each page/node.

The web graph seemed like seemingly useless data but it made the search industry a billion-dollar industry.

The web graph data is easy to collect by crawling the web and taking random walks.

Google's search algorithm is simple yet effective and it avoids the need to hire thousands of people to manually sort and organize search results.

Lecture – 80

Implementing Page Rank Using Points Distribution Method – 1

Introduction:

The method involves assigning fixed equal number of points to every node in the graph

Nodes then distribute their points to their neighbors connected by an out link

After each iteration, the points that every node contains keeps changing

Convergence is attained when the points do not change even after distribution

At this point, every node will have a certain number of points which can be used to rank the nodes (page rank)

Steps for Implementation:

Create a directed graph (using networkx or creating one manually)

Assign 100 points to each node as initial points

Every node distributes its points to its neighbors until convergence is reached

Rank the nodes based on the points they have accumulated (page ranking)

Compare the page ranking obtained from the implemented method with the inbuilt page rank method from networkx.

Lecture – 84

Implementing PageRank Using Random Walk Method – 1

Random walk method involves choosing a node uniformly at random and selecting one of its neighbors randomly to move forward in the network

Increment random walk points of a node by 1 each time it is reached during the random walk

Teleportation is used when a node with no out links is reached; choose a node uniformly at random and resume the random walk from there

The basic idea behind the random walk method is that a node with a high number of in links will be reached more often, which is recorded by the number of random walk points

The ranking of nodes based on random walk points is the same as the ranking obtained from the page rank technique

Steps to implement:

Take a directed graph

Perform random walk and keep count of random walk points

Rank nodes based on random walk points

Compare results with inbuilt page rank method from network