

PAGE RANK: HOW DOES GOOGLE WORK? 04

Now we had a bunch of people and we ask them to take part in an activity and then we drew a network from them and did what is called Random walking on the network i asked you to do it right will show you how to do it right now but one should do this random walk and then see the points accumulated per node and then look at that node with the highest point and call that fellow a leader. Now you know what? This is precisely how the Google search engine works and this algorithm is called the page rank algorithm. So what is the connection? Look at this, look at this network and take a random walk on it. Which what are those nodes which will have highest number of visits, its going to be those nodes which are easily reachable when you are taking random walks. How do you reach a node? It should be easy to reach their neighbours. How can you reach their neighbours? It should be easy to reach their neighbours neighbours the most popular person's neighbours neighbours so if there is a easy in coming way to a node then that node have a very high weight so intuitively speaking a person is important if important people point to him. A node will get a lot of weight if it is adjacent to a few nodes and these few nodes are all important, by important we mean they are adjacent to a lot more nodes, a node is adjacent to many nodes and these nodes are adjacent to many nodes and so on. Correct? So we are just giving you a very toned down version of page rank because its a first course in programming we don't want to bother you with the lot of details but if you are interested you can take a look at our course on NPTEL called the SOCIAL NETWORKS course there we describe in detail one entire week we discuss on how page rank works. So getting back we have now told you how to take a random walk on a network increment the points individually on the nodes and look at the top nodes thus obtained they are the most important nodes now how does Google work based on this idea. It's very simple what Google does is as you can see in this animation it takes a website and on that website as you know given a website it will have a lot of hyper links to another website by hyper links we mean a place where you go and you click and it takes you to another page you see that's called the hyper link ok so i will take all possible web pages in the world that one can think of and i will put an arrow from a web page to another web page if there is a link from this web page to this web page and i will create this huge network by the way this is how the network of internet looks like where a line between two node where nodes represent that sides and the line between two represent that there is a link from this side to that side this was the snapshot of the internet long long back off course it has grown a bigger right now it looks even more complicated right now but just for completeness sake we are flashing the network of the internet this is called the world wide web network so now here what Google does is it takes all these pages of the world and runs a random walk on it. It's a huge network you see it takes a long time to run random walk off course with sophisticated computers you can do this. A huge network on billions of node you run random walk and you make an note of how many times you visit a node as you are taking random walk, as we are talking, as we are listening to this video Google actually is doing this. It is taking all the websites of the world and then taking random walks on it. Ok you would have heard of crawlers this is precisely what i mean by crawlers. There is a program which crawls the web like this and makes a note of how many times i visited the what node and ranks all the nodes based on the visits to the node ok when you type the word Mysore city right it will look at all

those websites which has the word Mysore and then it will realise that there are ten thousand websites that has Mysore in it but it will show you the top few sites that have accumulated top points due to this random walk. You may want to get that straight let me repeat so when type Mysore there are ten thousand hits ten thousand sites which has Mysore in it but then Google will fetch the top few sites to you how does it know what are the top few sites, top few sites based on the points accumulated because of random walk that's it this single idea has lead to a five hundred billion dollar industry today. Google has it stand tall mainly started of this algorithm which till date is the best possible searching technology off course there are several other parameters in search engine users but the basic heart the nucleus of the idea is whatever we are whatever we taught you so far. Now let us go head and look at the integrities of how to take a random walk and then how to crawl on a network.