Social Networks
Prof. S. R. S. Iyengar
Department of Computer Science
Indian Institute of Technology, Ropar

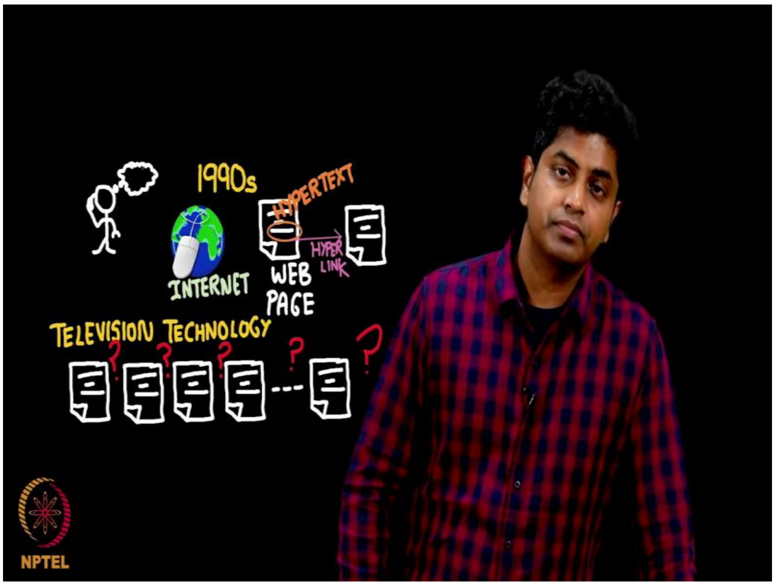
Link Analysis
Lecture – 75
The Web Graph

Let us down switch directions and talk about something that I spoke in at length in the introductory video.

(Refer Slide Time: 00:19)

It is about the origin of Google as an industry and the algorithm that they used to search the web. So, I will now call upon you all to imagine the following story.

(Refer Slide Time: 00:29)



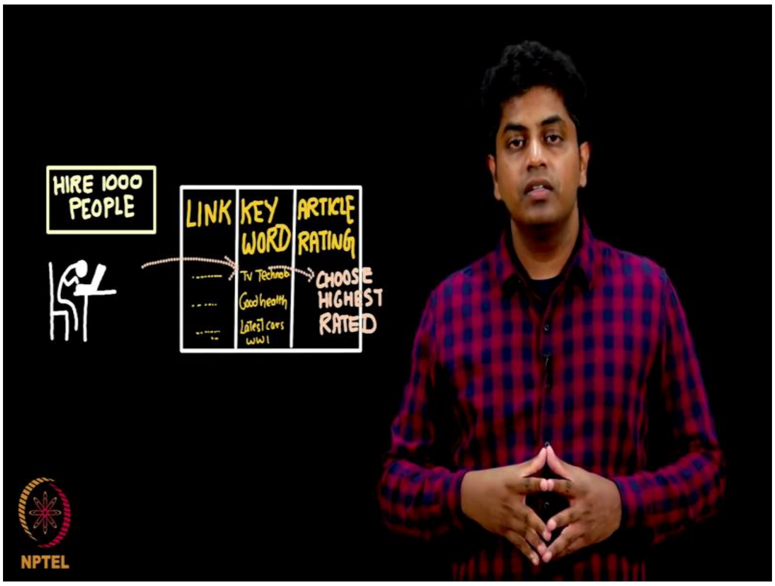
Please note you really have to imagine ~~otherwise,~~otherwise; you cannot get the gist what I am going to say. Please imagine that me and you are in the 1990s and I have this word with you that, see there is something called the internet which is fast developing. There is something called a web page where people put some information and there is something called hyper link, a hypertext where in the page you can link to another page. You click here, it just goes there right that is what you mean by hyperlink. By a link you mean something that links to somewhere else, but hyperlink means something that not just links hyper means what more than a link hyperlink is more than a link. Hyperlink means it just does not link you to a new place; it even takes you to that place.

So, a ~~hyperlinks~~hyperlink was the invention of hypertext was very big leap in computer science. So, there are pages with hyperlinks to other pages and now me and you are thinking of a start up idea. In the 1990 where we see that internet is booming with a lot of web pages which are linked ~~one page~~one-page links to the other page through a hyper link and we all start thinking. Now there are so many pages here thousands and thousands of pages. Today actually its billions of ~~billions~~billions of pages, billions of pages, but I am talking about 1990; me and you are having coffee table conversation in nineteen ninety.

Look there are so many web pages. How do you know what to search for? How do you know which page has the right information? Assume there are so many pages, web pages talking about television technology, the television technology. Which page is a authoritative source for me to refer? If there are 100 pages which talk about 100 different web pages written by 100 different technicians, technology experts. Which is the best possible webpage? Now I see I repeat we are in 1990s, I see that this probably becomes a problem is going to become a problem very soon. Now that there are few pages people can sort of go through all that page and decide which one is good and then read it, but very soon the web pages will be populated with hundreds of thousands of pages. How on earth will people know what to look for where right? What is the immediate idea, you think we both will get over this coffee table discussion in 1990?

I for a reason would give this dumb idea that fine we will hire some thousand people. Assume we had we had so much money to pay thousand people; I will say we will hire thousand people and ask them to go through each and every page and maintain a sheet which has the following entries following columns.

(Refer Slide Time: 03:39)



LINK	KEY WORD	ARTICLE RATING
-------	TV Techno	CHOOSE HIGHEST RATED
-- ---	Good health	
-- ----	Latest cars WW I	

First column will be the link second column will be the keywords which tells me what ~~is this link~~this link is all about. This is the link I will say it talks about television technology. There is another link, this talks about health tips for good health; another link that talks about latest cars, another link that talks about World War-I so on right. First column will have the link, second column will have ~~this keywords~~these keywords what I what is the link all about. The third column will be, I will read by me I mean ~~this~~

~~thousand employees~~these thousand employees I am going to hire, each one of them will read the link and give the article rating. Or this article on World War-I is 8 on 10 rating, 8 on 10 like how the give rating for movies. Another person sees another link on World War-I and says this is not so good; it is 3 on 10 so on and so on and so on. These thousand employees in my start up industry will go through every single possible page ever created, ~~somehow~~somehow, we get access to all the pages let us say and we rate all the pages and we also put keywords.

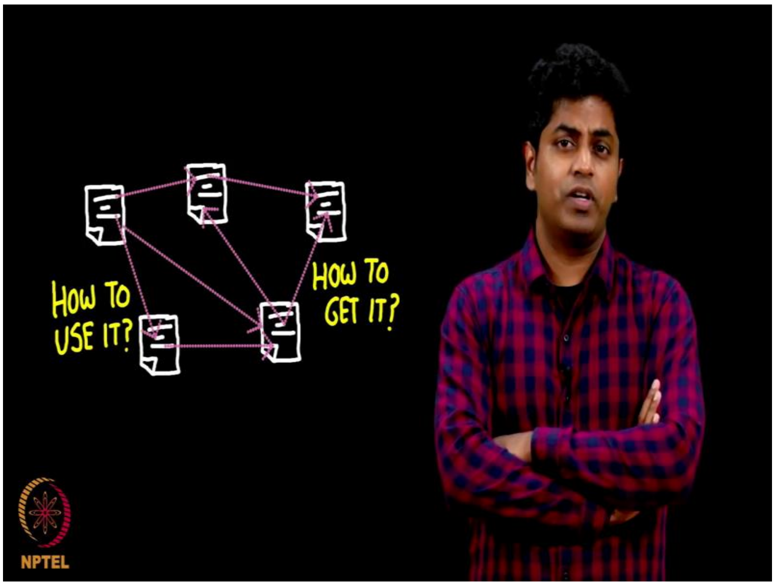
So, that whenever someone searches for a keyword television technology, then it will come to this database of all the entries that my thousand people whom I have hired have populated. It will look at the highest rated page which has the keyword television technology and display that to him. Maybe it will sort it in the order of rating, best rated link first next best second, third, fourth descending order. It will show it you wonderful idea, everybody will come to my page to search for information on what is the ~~so called~~so-called internet with so many pages. Very soon I will become a millionaire because everybody will be coming to my page, my page will become popular. And once you have a popular portal, you can make a lot of money through ads or whatever right.

So, here is my idea in 1990 let us say, but then nobody ever dreamt that there could be a much ~~much~~ better way of doing this. I am going to explain next: what is the much ~~much~~ better way which paved its way for the birth of as I said a ~~450 billion dollar~~450-billion-dollar industry called Google; obviously, my start up idea will not scale up. Thousand people can do how much of work? The number of pages that gets populated on the internet is in terms of millions. So, very soon I will run out of man hours I am I will be need of more man hours and less employees; obviously, thousand is not enough. You should think of some other way of doing it right. What is this way of doing? It is ~~very~~ simple I will tell you the simple idea which is behind searching and again I will try explaining it bit by bit with analysis examples and questions; very simple.

(Refer Slide Time: 07:23)

Larry Page and ~~Surgey~~Sergey Brin, they came out with his idea of maintaining what is called a web graph.

(Refer Slide Time: 07:35)



By a web graph you mean take a page a webpage, call it a node. Take all the web pages, each web page is represented by a node. You wire a link between this page and this page namely this node and that corresponding node by a link. If this page has a hyperlink to that page; for example, the same example I have been giving you people. This is my homepage and I link the homepage of my friend on my homepage. Then my homepage is a node, my friends homepage is a node; there is an edge from my home page to my friends homepage.

Please note this is a directed graph, why? Me linking to Prime Minister’s homepage does not imply that the Prime Minister’s homepage also links to mine. It is completely one sided correct. I like her, she does not mean she likes me. It is a very asymmetric relation. So, this is what is called a web graph where pages are denoted by nodes and a link represents a page pointing to another page through a hyperlink. The point is if you collect this web graph that is enough it does the trick of having thousand people or more. First question, how do I collect the web graph, what is the use of this web graph, how is this web graph going to solve the problem that I was trying to resolve with some thousand employees by hiring thousand employees right.

I have thousand employees in my organisation and now I am saying you can replace these thousand employees or even more by simply making node of the web graph. How do we get a web graph? What is the web graph? Think about it. It is a graph containing so many nodes and so many links. How do you even now the in some nook corner of the world, there will be a web page point into some other web page I am saying that you should have the all this information. How do you get this information? Do you have any idea how one can get this information? Any rough idea? Think about it.