**Cloud Computing for Data Analysis VIDEO CASE 06 : PageRank**

**Name: Ashish Bisht**

**Student Id: 801168390**

Watch following videos:

**Video 1:** https://youtu.be/A4-yw07Ku1s

**Video Case Questions:**

1. What is PageRank?

Ans1. The **PageRank algorithm** is a classic example of a **distributed iterative algorithm** for the purpose of ranking the relative importance of each of a set of linked pages, such as the World Wide Web. The goal of the algorithm is to compute a score for each page, representing its relative importance. All pages start out equal, so the initial score is set to be 1 for each page, and the previous score is 0. Creating the list of outbound links is achieved using the Crunch operations of grouping the input by the first field (page), then aggregating the values (outbound links) into a list.10

The iteration is carried out using a regular Java while loop. The scores are updated in each iteration of the loop by calling the pageRank() method, which encapsulates the PageRank algorithm as a series of Crunch operations. If the delta between the last set of scores and the new set of scores is below a small enough value (0.01), then the scores have converged and the algorithm terminates. The delta is computed by the computeDelta() method, a Crunch aggregation that finds the largest absolute difference in page score for all the pages in the collection.

2. What are all factors we need to consider for calculating a web page’s PageRank score?

Ans 2. The factors that we need to consider for calculating a web page’s PageRank score are the following:

1. **Link Back to High-Quality Sites:** Google PageRank is calculated primarily upon the quality of the sites you link back to. Remember, the larger the amount of high-ranking sites you link back to, the higher overall Google PageRank your website is going to achieve.

2. **Total Inbound Links:** Whenever the number of incoming links increase in number, the PageRank automatically goes high for that particular web page.

3. **Keyword in Title Tags:** A title tag is the single most important on-page optimization factor besides content itself.

4. **Page Age:** Page age refers to the date the article was posted, not the date the websites domain was created. The newer the date on the page, the more likely people are going to click on it.

5. **Quality of Internal Links Pointing to Page**: Internal links from authoritative pages on domain have a stronger effect than pages with no or low PageRank

3. When does a PageRank of a web page go high?

Ans 3. The PageRank goes high when the **incoming nodes/links** increase in number. When the PageRank of current pages increases, the PageRank of pages linked to the current pages also increases. And if a new page is linked further, eventually the PageRank also goes high. Thus, the webpage with the highest PR score comes under “**Very Important”** category, the second highest score comes under “**Important**” and the other web pages come under “**Average**” depending upon their PR score.

However, now should be careful of the fact that linking thousands of low quality websites can actually decrease the PageRank of your particular website and overall standing with Google. Google’s ranking mechanism will be able to detect that your site is listed on so many of these undesirable pages. In many cases, this is seen as an attempt to inflate your PageRank unfairly, and sites may be penalized as a result. Instead, spend your time submitting to credible directories and exchanging links with other relevant sites. The process can be time-consuming, but it is the only way to generate safe, site-specific traffic.