**Big Data Project (Option A)**

Submitted by:

Harshvardhan Pandey

Sujay Kulkarni

Under the guidance of:

Dr. Elham Khorasani

**Introduction:**

Presently a days the accessibility of effective and exact Web hunt like Google, which was the first web index ready to thrashing the spammers who had made futile inquiry. It gave the creative system known as PageRank.

PageRank is a capacity that allots a genuine number to every page in the web. The goal is that the higher the PageRank of a page, the more vital it is. It is for the most part figured by the equation which clarifies it by utilizing the approaching & active connections and additionally the consistent quality and the aggregate number of pages.

Project goal is to print the Page Ranks of the Wikipedia page by using the algorithms like Map Reduce, Pig and Hive.

**Methodology:**

In our project we try to solve the issues in two Steps, Step1 deals with extracting the link data sets from the provided input xml files by using the Mapreduce algorithm. Whereas the Step 2 deals with the removal of red links from the link data set and then assigning the page ranks for the data set by using Pig or Hive Script.

The Mapreduce algorithm we used in this project has a Mapper class, Driver class and XMLinputFormat class. We can generally extract link data set from xml files in many ways and one of them is to use Mapreduce along with Apache Mahout Library to parse xml files. We can use Mahout XmlinputFormat to extract everything between <page></page>tags and send it to our map function. Then in our map function we can use the built in java library to extract the value of the <revision>, <title> and the <text>. Once we extract the text we use java regular expressions to look for wikilinks. To get the link text and link title we used pattern and matcher functions with the following as the pattern input.

\\[\\[([^\\[\\]|]\*)[^\\[\\]]\*\\]\\]

We use Pig Script to remove the red links from these link data sets. To remove them we need to get the output of our map reduce which is the relation of two columns let’s call it R. we need to make a copy of this relation which is S, then we need to join R with S on the 2nd column of R and 1st column of S and then we only take the first and second columns out of join result

Once we remove the red links we will do the page rank computation on the link data set. In page rank computation we divide the computation part in to pieces, we define the constants, input and output links for the data and develop the table which consist of columns made out of these pieces. We take these column values and calculate the Page ranks by the first iteration, but the sum gradually decreases with the iterations, we further calculate the second iteration by taking the output of the first iteration. Like this we do ten iterations.

Once we are done with these iterations, we use the rank calculation of the last (ten) iteration and use the Order function to get it in descending order, once we get it then we filter it by rank(which is one of the column in the table with divided computation parts ) which is smaller than 5/N .

**Results & Conclusion:**

Result: According the task the normal yield after the 10 iterations is the dropping request of the page ranks of the link data set file.

References:

1. For Xml file http://en.wikipedia.org/w/index.php?title=Special:Export/
2. For references in coding : <http://xmlandhadoop.blogspot.com/>
3. Quick start guide to java regular expression:http://ocpsoft.org/opensource/guide-to-regular-expressions-in-java-part1/
4. Forum:http://stackoverflow.com/questions/8138153/matching-wikilinks-with-a-regex