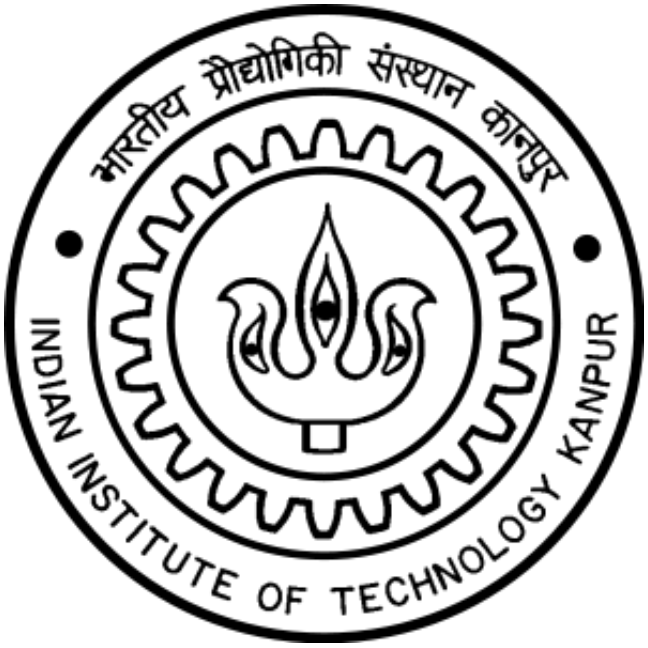
Indian Institute of Technology Kanpur

Assignment-2 report

Course Instructor: Prof. Arnab Bhattacharya

November 19, 2020



Abstract :

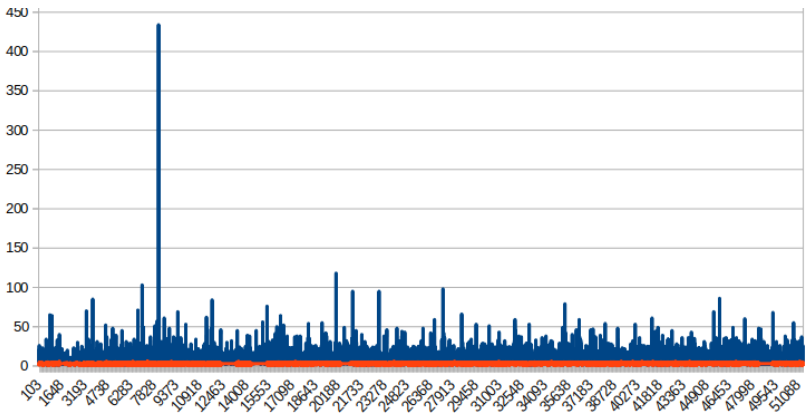
The Data-set we have here contains human navigation paths on Wikipedia, Where users are asked to navigate from source to target by clicking on Wikipedia links.
With our analysis we will try to analysis how humans navigate and find their target form source, and how humans approach is different from any available algorithms to find shortest path. How efficient humans approach is when they know their target. To compare with practical approaches we will use shortest path algorithm(*like bfs and floyd warshall*).

Analysis :

Here, we are using 4604 articles and 146 categories. we gave ids to articles from $A0001$ to $A4604$ and gave ids to categories from $C0001$ to $C0146$ based on alphabetical sorting.

Paths Traversed by Humans and Shortest path :

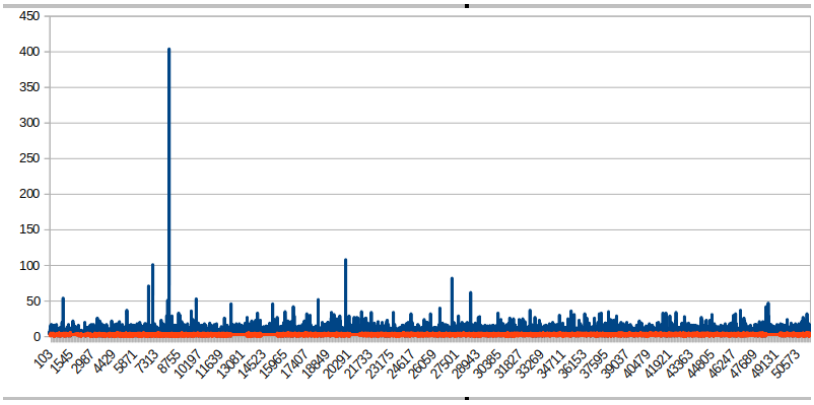
We have 51306 paths and we are comparing the length of path traversed by humans and the actual shortest path(*from shortest-path-distance-matrix file*).



Human Path Vs Shortest Path

In above Fig. Blue lines shows length of human path and Red lines shows length of shortest path.

On visualization of above graph we can see that the path traversed by humans in web browsing is higher in almost all the paths.



Human Path without back-links Vs Shortest Path

In above Fig. Blue lines shows length of human path(*without back-links*) and Red lines shows length of shortest path.

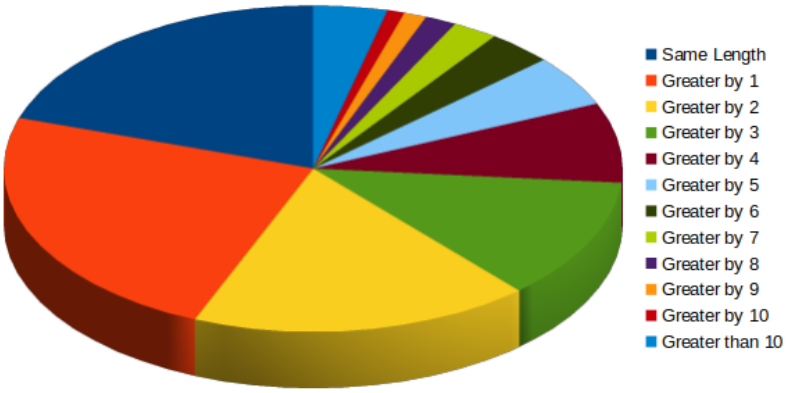
Averages	Human Path	Shortest Path
Finished Path	5.76	3
Finished Path(*without back-links*)	4.96	3

From these visualizations and the average matrix we can conclude:

1. The path traversed by humans is greater than shortest path in most of the times.
2. Also, we if we see the averages, we can conclude that humans traversed path are not so much longer all the time but yes it almost doubles of the shortest path.
3. The average of human paths(*without back-links)* is .80 less than the average of human paths(*with back-links*), from this we can conclude human takes almost 1 back links on an average in each path.

Difference in Length between Human Path and Shortest path :

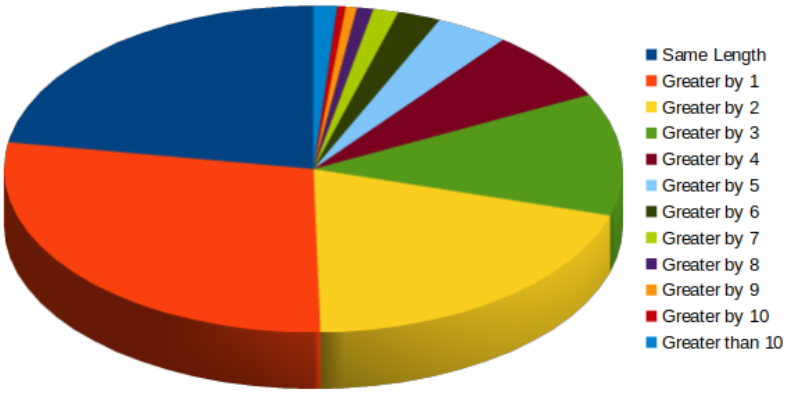
As we can see from above conclusions that not all path are very much large from shortest path. Here we are comparing percentages of difference between human path length and shortest path length.



Differences between Human and Shortest Path

Differences	Percentage
Exact, Greater by 1 or 2	61.57%
Greater than 2	38.43%

From these visualization we can see that more than 60 *percent* of paths are max greater by 2. There may be several reasons that some paths are very much greater than shortest path, like some path are much more difficult to interpret than others also, it may vary from human to human. Some human can interpret a particular task very easily in comparison to others.

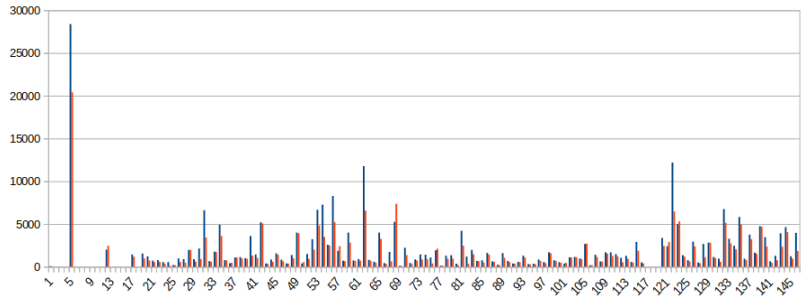


Differences between Human(*without back-links*) and Shortest Path

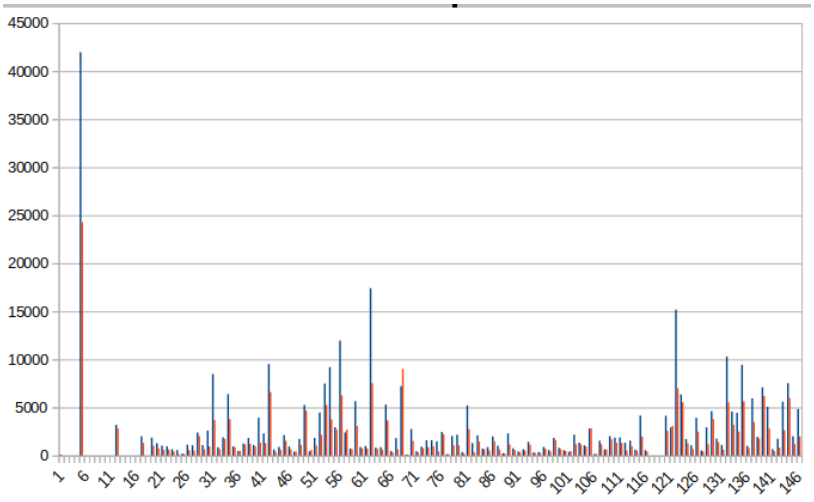
Differences	Percentage
Exact, Greater by 1 or 2	70.38%
Greater than 2	29.62%

We can see here the percentage go more than 70 *percent* if we didn't consider the back-links .

Number of Clicks on Each Category :



Comparison of First Click by Human And Shortest path

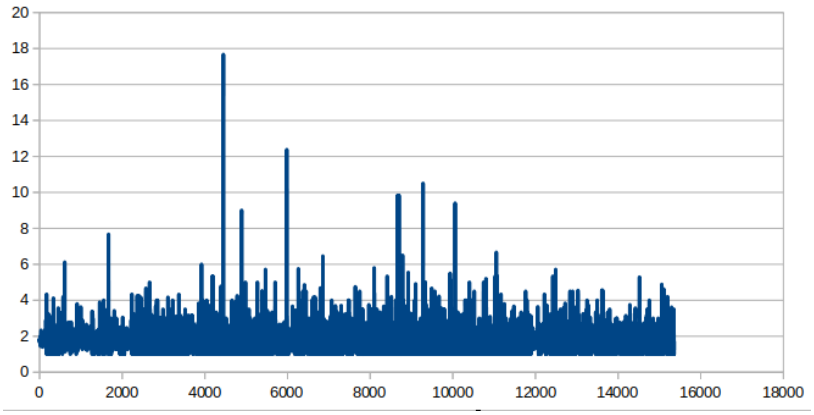


Comparison of Total Click by Human And Shortest path

From these visualization we can see that First click made my humans are mostly accurate except in some cases. While on the other hand if we see total number of clicks than there is significant amount of increase in almost all the categories.Also, from here we can conclude first click made by humans are very efficient while finding the target.

Ratio of Human path to Shortest path :

As we already seen that most of the Human path are greater than Shortest path by 1 or 2 length while other are comparatively much larger than shortest path.
Here, we plot ratio of human path to shortest path, but we here we take all the available human path for each category pair of source and destination and find the ratio based on all the paths for each pair.



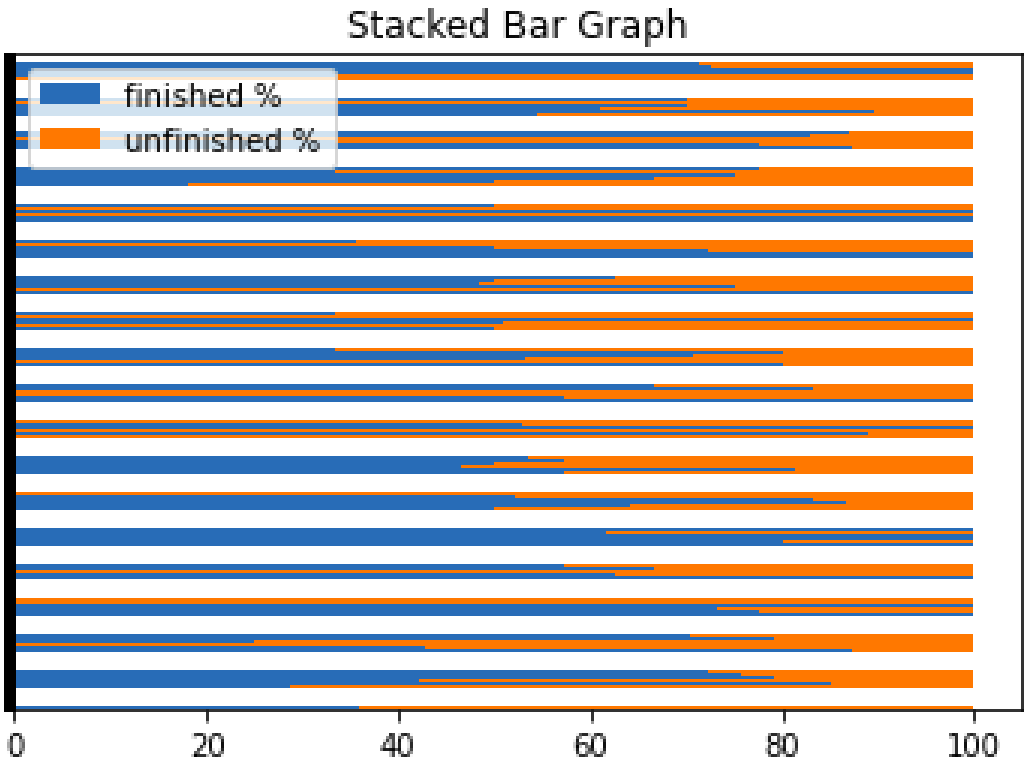
Average Ratio for human and shortest path for category pair

Max Value of average Ratio	17.66
Mean Value of average Ratio	1.72
Min Value of average Ratio	1

This visualization also conclude that only some paths are much greater than the shortest path , which may have different reasons like some targets are difficult to find and some are difficult to interpret by a particular human(*depends on individual*)

Comparison of Finished and unFinished path :

Here, we take all the category pairs for source and destination and analyse in how many paths target is achieved successfully and how many paths remain unfinished.



Percentage of Finished and unFinished paths

Here, these blue lines represent the *percentage of finished paths* and rest of the orange part *i.e.* $(100 - \%of finishedpath)$ is *percentage of unfinished path.*

Here, there may be several reasons for path to be remain unfinished like target may be very difficult to reach ,having same kind subcategories to click on, in that case user may end up in traversing loops again and again.

Conclusion :

The main aim of this report is to analysis the results of *WIKIPEDIA paths* for *human traversed* and *shortest path*. We have analysed our result for 4604 articles and 1460 categories.
Moreover, we mainly focused on Finished path for many aims but as a future expand we can extend this analysis to know the reason why human give up on some paths(*unfinished paths*).
Also, some paths are very much greater than others, we can take different clusters as a data set and see which factors affects the human path to be much larger than shortest path in *Web Browsing*.