CS685A Data Mining : Assignment-2

Harshvardhan Pratap Singh (20111410), hrshsengar20@iitk.ac.in

November 20, 2020

Abstract

In this assignment we studied dataset containing human navigation paths on Wikipedia, collected through the human-computation game Wikispeedia. This is a brief report for the analysis of results and conclusions.

1 Introduction

Purpose of this report is analysis and mining of Wikipedia data. Finding pattern in it, how articles are connected with each other like graph nodes based on various parameters like categories. What are some common pattern between the visit of users path and what is the shortest path to get some article. Some articles are showing connectivity with other articles as we can see in the connected components of the graph. How big those components are and what is the reachability in those components. Wikispeedia, users are asked to navigate from a given source to a given target article, by only clicking Wikipedia links. Analysis of all those navigation paths are included in this report.

2 Analysis

article-ids.csv Looking at the article-ids.csv, Articles are assigned unique ids for better use further.

category-ids.csv Looking at the category-ids.csv, Categories are assigned ids into a hierarchy. Starting from subject to all sub-categories in breath-first ordering

article-categories.csv For every article id form article-ids.csv we are assigning all the categories and sub-categories id which this article belongs to.

edges.csv edges.csv tells us about edges in the graph of articles. edges are of directed graph. this edges are mined from the available shortest path distances from the shortest-path-distance-matrix.txt which is containing all pair shotest path.

graph-components.csv graph-components.csv are giving information about the connected component present in the graph like how many nodes and edges are there and also what is the reachability by seeing at the diameter, If diameter is low then every article is well connected and easily reachable within the component.

finished-paths-no-back.csv and finished-paths-back.csv finished-paths-no-back.csv and finished-paths-back.csv are giving information about the length of paths finished by human to reach target from source and also gave shortest path between them and also the ratio between human path and shortest path. so that we can get to know how much inefficient human path is with respect to shortest path.

percentage-paths-no-back.csv and percentage-paths-back.csv percentage-paths-no-back.csv and percentage-paths-back.csv gives information with respect human path and shortest path and human path. analysis with respect to how inefficient human path is with respect to shortest path with every possible difference.

category-paths.csv Given all the finished human path category-paths.csv is giving us information about for each particular catogory, in how many path and how many times particular category is appearing. In both human path as well as shortest path.

category-subtree-paths.csv category-subtree-paths.csv is giving us information about category appearing as a super category of its sub category. Basically on high level how many times certain category or its subcategory is visited.

category-pairs.csv category-pairs.csv is giving us information each source and target appearing in every finished and unfinished path. Also we are calculating percentages for finished and unfinished path for each source and target category pair.

category-ratios.csv category-ratios.csv is giving us information ratio of human (without back clicks) to shortest paths. For finished paths for each source and target category pair.

3 Observations and Conclusions

When we see connectivity between articles we can clearly see that in undirected graph there is one big connected component containing 4589 nodes out of 4604 article nodes and components diameter is 5. So we can conclude that articles are well connected and can be easily accessible and reachable. also by seeing the ratio of human to shortest path we can say that there is not much difference in general. And also by looking at category-paths.csv we can say that what categories are most promonant and are appearing in most of the paths. By looking at the ratios of human and shortest path we can say that shortest path and human path are not varies much.