Indian Institute of Technology, Kanpur

CS685A: Data Mining

Assignment 2

Supervised by: Dr. Arnab Bhattacharya

November 16,2020

Submitted By: Aditya Jain
20111004
adityaj20@iitk.ac.in

Table of Contents

1 Introduction

I have done analysis on the dataset that contains the human navigation paths on Wikipedia, collected through the human-computation game "Wikispeedia". Wikispeedia, users are asked to navigate from a given source to a given target article, by only clicking Wikipedia links. A condensed version of Wikipedia (4,604 articles) is used.

1.1 Data collection

Data for this study is collected from `http://snap.stanford.edu/data/wikispeedia.html`. Following tsv files are incorporated for data processing to generate the results. These are available in wikispeedia paths-and-graph.tar.gz.

1. articles.tsv
2. categories.tsv
3. paths_finished.tsv
4. paths_unfinished.tsv
5. shortest-path-distance-matrix.txt

1.2 Data Correction and Pre-processing

Following changes are made inside the programs to correct the errors in the file.

- There exists 6 articles for which no category is defined in category.tsv, so I have assigned them to category "subject". (Q3)
- Paths of length zero are removed from paths_finished.tsv as they are invalid for the scenario. (Q6)
- Due to errors in article names in paths_unfinished.tsv, I have renamed the article names for which a possible valid article name was present and others are assigned to category subject as no suitable mapping exists for them. (Q10)

A total of 4,604 articles exist in our experiment and these articles belongs to 146 different categories. There are 51,306 valid finished paths and 24,875 valid unfinished paths in the analysis domain dataset.

2 Observations & Analysis

2.1 Wikipedia Articles and Categories

We are analyzing a total of 4604 Wikipedia Articles. These articles are assigned to 146 different categories. These categories are sub-categorized level wise:

Number of Categories at level 1	1
Number of Categories at level 2	15
Number of Categories at level 3	130

Table 1: Number of sub-categories at each level

After assigning categories to each article, I found that there are 6 such articles which do not belong to any category. So to use these articles in the analysis, "subject" category is assigned to these 6 categories. Table 2 shows that there are only 8 articles that belongs to exactly 3 categories. So we can infer that our sample size include articles which are restricted to few categories only.

Articles belonging to at least 1 category	4598
Articles belonging to at least 2 categories	598
Articles belonging to at exactly 3 categories	8

Table 2: Number of categories to which an article belongs to

2.2 Connectivity of articles in Wikipedia

Nodes	Edges	Diameter
4589	106534	5
3	3	1
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0

Table 3: Connected components of graph made on articles

In our sample size of 4604 articles, we found that there are 14 connected components in which the largest component has 4589 articles. This shows that the articles are very

well connected to each other. The diameter of the graph drawn of these 4589 connected articles is 5, which means that an user will require to click on at most 5 links to go from any article to any other article among these 4589 articles.

There are 12 isoloated articles which are not connected to any other article. So if user lands to any of these articles it cannot go to any other article through the wikipedia's web interface. He/She needs to go back to search engine to go to another article.

2.3 Reachability of articles in Wikipedia

In the game, user is asked to go from one source article to a destination article through the interface of Wikipedia by clicking on the links present on the articles. Path that user followed to reach the destination is referred as "Human Path" in the entire report.

Figure 1 is a Pi chart which shows the comparision of human path lengths and their corresponding shortest path length in terms of percentage. This results are made from 51,306 test cases.

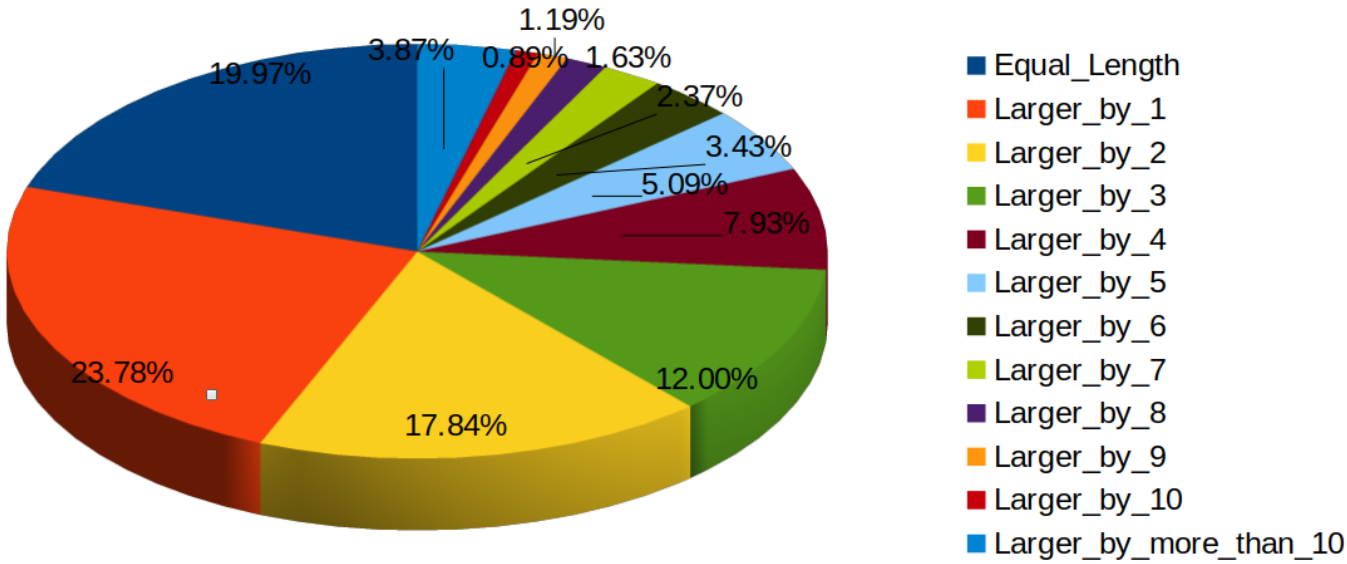

Figure 1: Comparision of Human path to its corresponding shortest path length

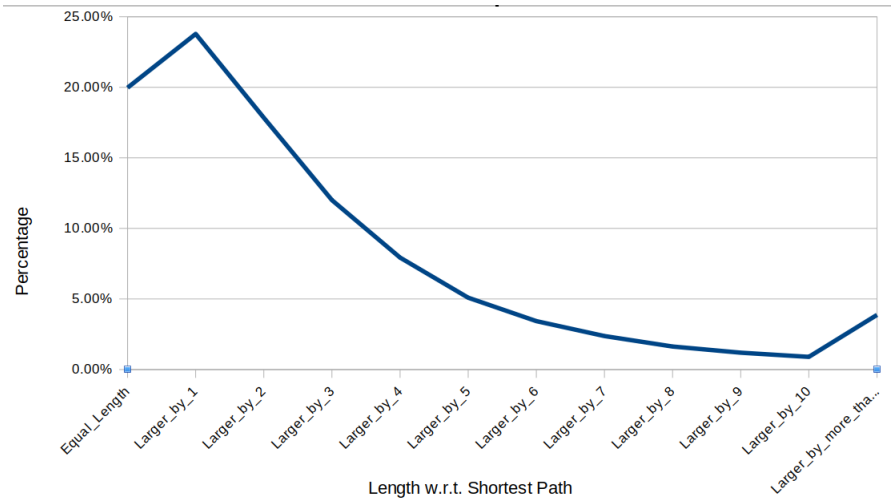

Figure 2

In Figure 2, we can see that around 20% of human paths are equal to their corresponding shortest path. There are 23.78% paths for which human path length is one more than its corresponding shortest length. This shows that Wikipedia article interface is quite readable and interpretable that users are able to move accross the articles very easily.

2.3.1 With Backlinks

There is a exponential fall in the percentage as we move accross the x-axis and increase the human path length.
Since around 60% of test cases have reasonable human path length, we can conclude that Wikipedia pages are quite reachable and user-friendly.

2.3.2 Without Backlinks

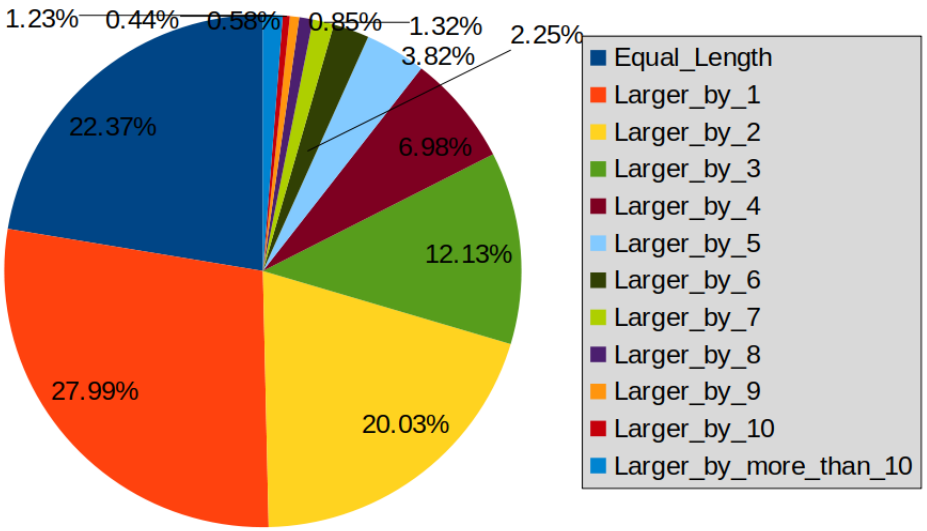

Figure 3: Comparision of Human path to its corresponding shortest path length

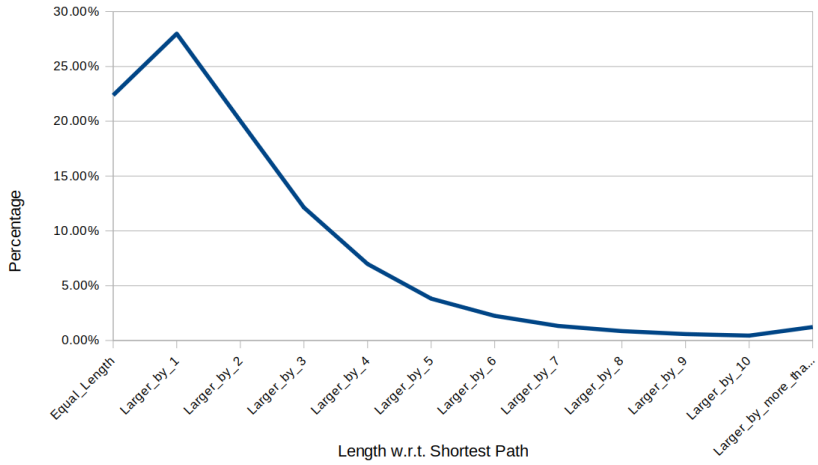

Figure 4

Back clicks: User clicking on the back button on the brower to go back to the last article. Here we have ignore the back clicks made by the user and we found that now 70% of the test cases are of reasonable length. There is an increase of 10%. Since this increase of 10% is occuring in the human path length greater than one, this shows that user is able to identify that he/she has traversed to a wrong article just after going to that wrong article. So we can concur that Wikipedia articles are quite interpretable that user is able to identify his/her mistake.

2.4 Category Analysis

2.4.1 Popularity of Categories

2.4.2 Bottom-Up Analysis

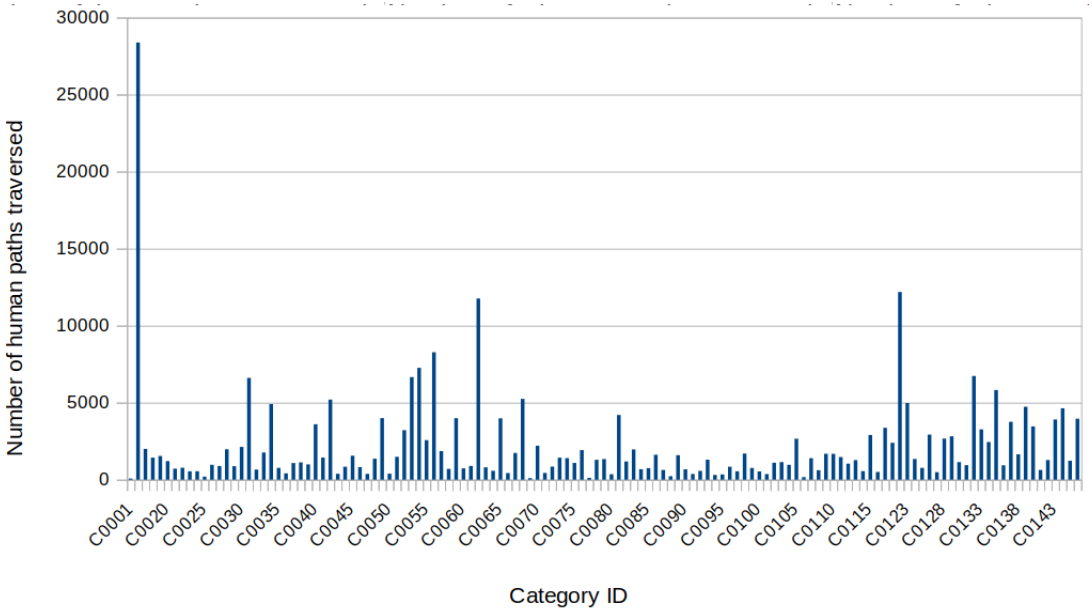

Figure 5: Bottom-Level Category ID vs Number of Human Paths they occur

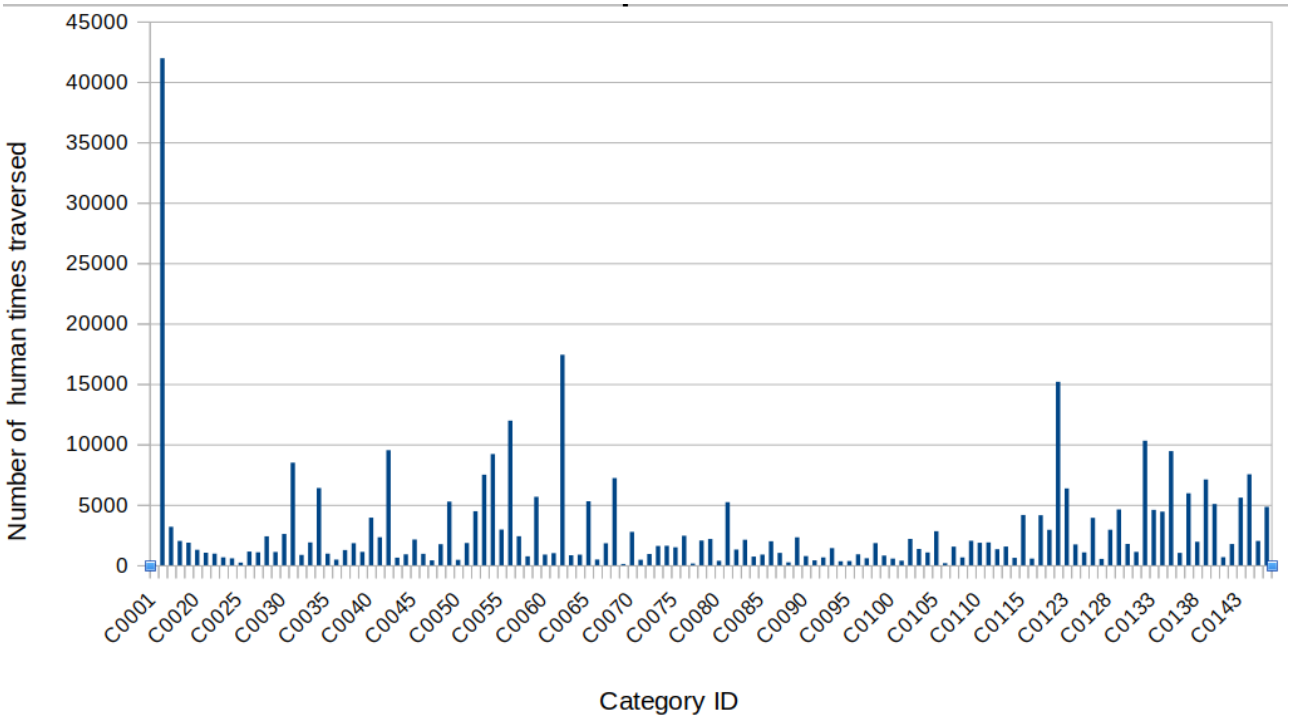

Figure 6: Category ID vs Number of times they occur in Human Paths

Figure 5 shows the number of paths in which a category is occured whereas Figure 6 shows the number of times a category is occured in the whole test case. Both are the results for the category at bottom most level, here we are focusing on the broadest granularity. So out of these 130 categories top 10 most popular categories are :

Category ID
C0005
C0122
C0062
C0056
C0054
C0132
C0053
C0031
C0135
C0068

Table 4: Top 10 Popular Categories

These are the categories which occured in most of the human paths. Category C0005 occured in 28,375 times alone.

2.4.3 Top-Down (Sub-Categories) Analysis

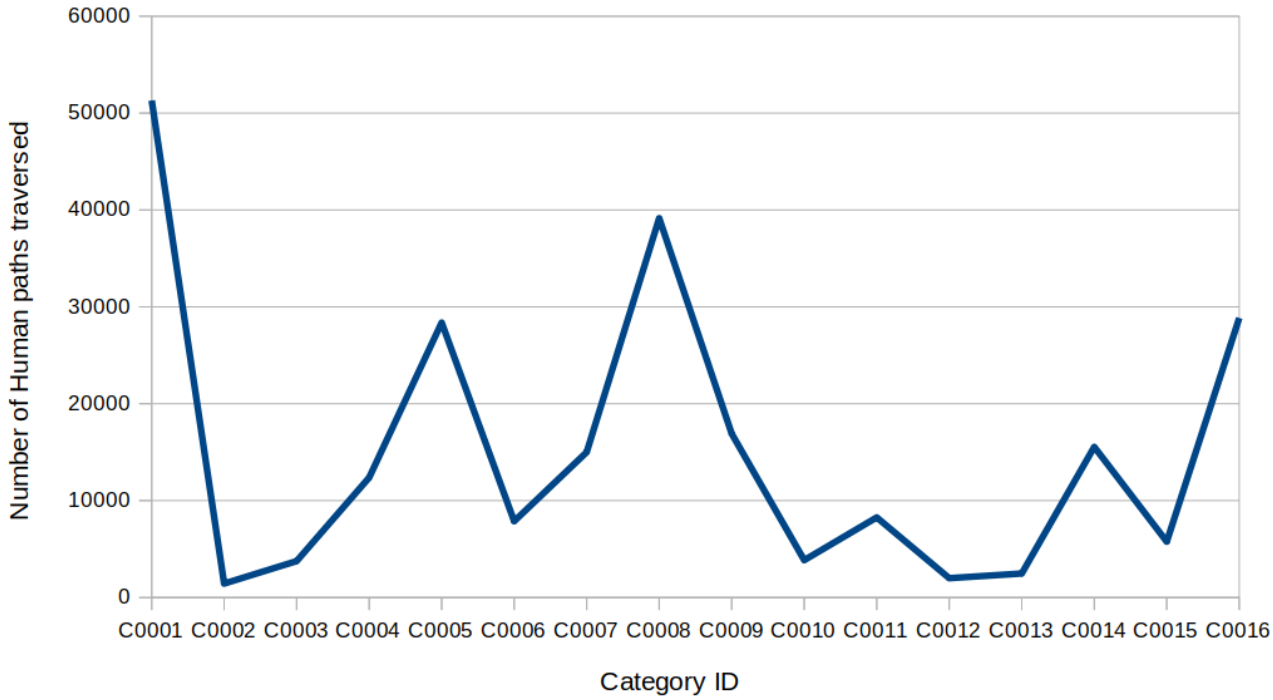

Figure 7: Top-level Category ID vs Number of Human Paths they occur

Here focus on 16 categories at level 1 and 2. These are parent to the 130 categories categories at level 3. Figure 7 shows the number of human paths in which these categories have occured. Top 5 popular categories are: Since C0001 is parent to all the categories,

Category_ID
C0001
C0008
C0016
C0005
C0117

Table 5: Top 5 populuar categories

it is obvious for it to occur on top. After it C008, C0016 and C0005 are the most popular categories that are occuring in the search pattern of the users.

2.5 Shortest Path Analysis

For each source and destination article pair that user was asked to traverse, we found the shortest path of articles. Following is an analysis on the categories that occured in the shortest path corresponding to each test case.

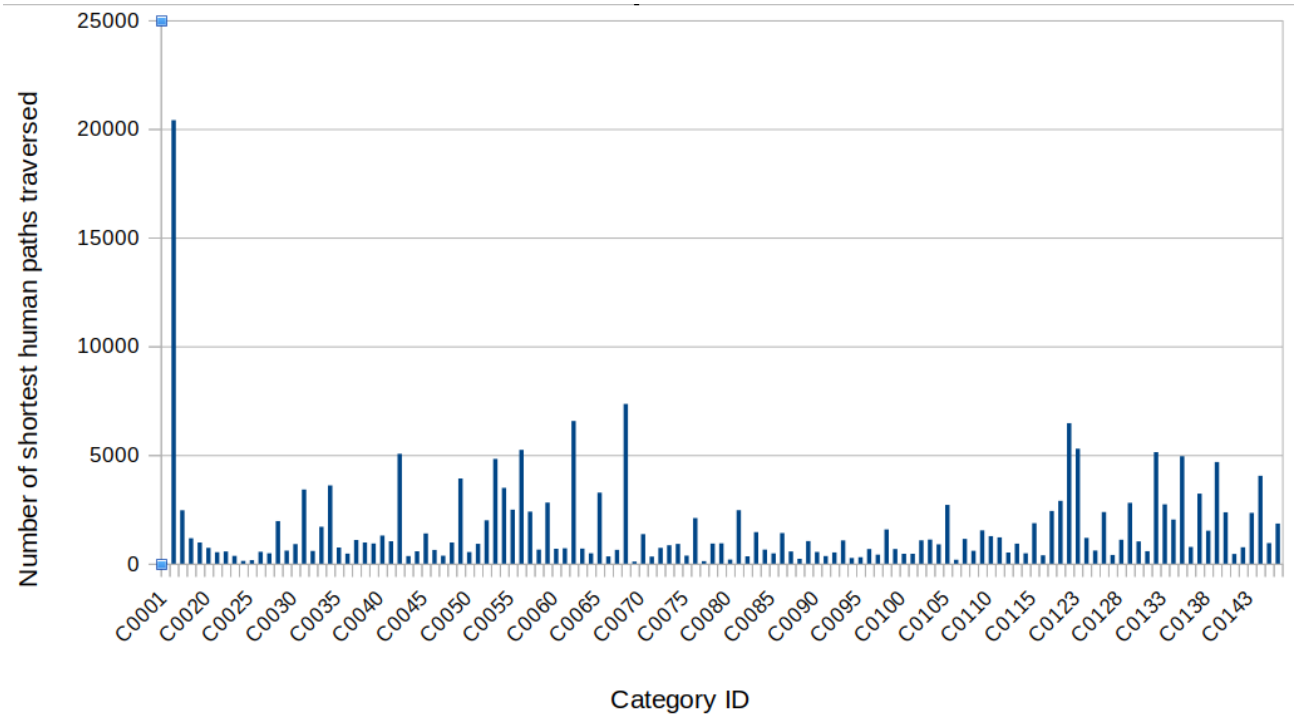

Figure 8: Category ID vs Number of times they occur in shortest path

Figure 8 shows the categories that occur in the shortest path for each test case of source and destination article. From the above graph we can tell the categories which are not only most popular but are also most reachable. These categories come in the shortest paths shows that they are very much related to other categories.

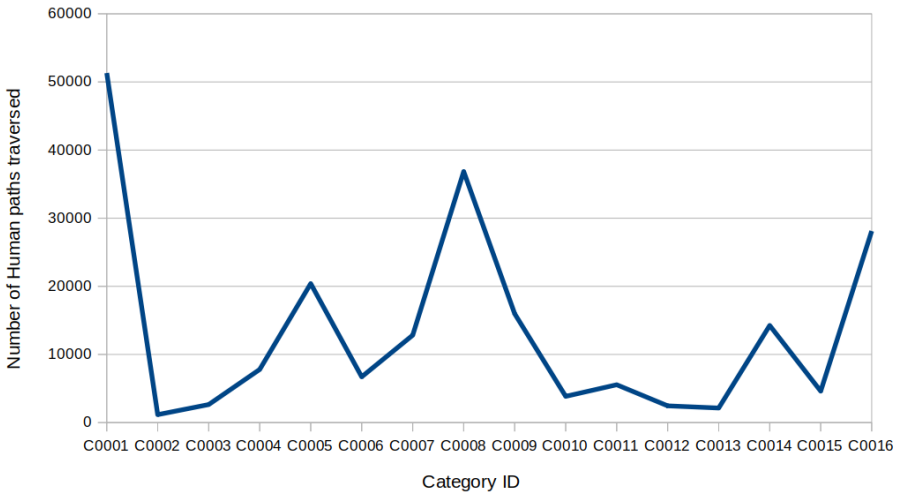

Figure 9: Top-level Category ID vs Number of shortest paths they occur

Figure 9 shows the categories that occur in the shortest path for each test case but here we only considered the categories at level 1 and 2.
Category IDs C0005, C0068, C0062, C0122, and C0123 occured most times in bottom-up analysis, whereas C0001, C0008, C0016, C0005, and C0117 occured most times in top-down analysis. This shows that these categories are most similar to other categories.

2.6 Category-Pair Analysis

Observation and results are as follows:

- Total of 17,196 category pairs exists for complete test case.
- Total of 15,356 category pairs exist for the finished paths.
- Total of 13,754 category pairs exist for the unfinished paths.
- There are 1,840 category pairs which are not connected, i.e. user can't traverse from one category to another through links.
- There are 1,419 category pairs have equal probablity to be completed by the user or left by the user in middle.
- There are 1,648 category pairs for which no finished test case exist, suggesting that these category pairs are unreachable.

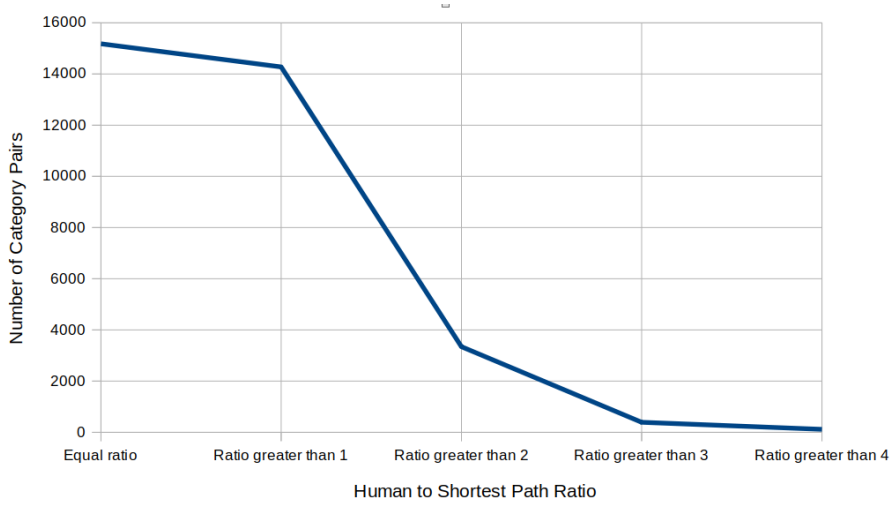

Figure 10: Number of category pairs in a Human/Shortest path ratio segment

Results from Figure 10:

- There are 1,081 category pairs whose ratio is 1.
- There are 14,275 category pairs whose ratio is more than 1.
- Ther are only 3,345 category pairs whose ratio is more than 2.
- Only 391 category pairs whose ratio is more than 3.

- Only 117 whose ratio is equal or more than 4.

There are 94.04% category pairs whose human to shortest path length is less than 2. This shows that Wikipedia's interface is very user-friendly such that user can traverse from one category's article to another category's article without any hassle.

3 Conclusion

- The articles are very well connected to each other. The user will require to click on at most 5 links to go from any article to any other article among most of the articles. Maximum path size is 9 links, but there are very few such cases.
- Wikipedia article interface is quite readable and interpretable that users are able to move accross the articles very easily using forward and backward links.
- There are 94.04% category pairs whose human to shortest path length is less than 2. This shows that Wikipedia's interface is very user-friendly such that user can traverse from one category's article to another category's article without any hassle.