Wikispeedia Analysis
CS685A- Assignment 2

Snehal Raj

December 6, 2020

1 INTRODUCTION

Navigating information spaces is an essential part of our everyday lives, and in order to design efficient and user-friendly information systems, it is important to understand how humans navigate and find the information they are looking for. We perform a large-scale study of human wayfinding, in which, given a network of links between the concepts of Wikipedia, people play a game of finding a short path from a given start to a given target concept by following hyperlinks. What distinguishes our setup from other studies of human Web-browsing behavior is that in our case people navigate a graph of connections between concepts, and that the exact goal of the navigation is known ahead of time. We study more than 30,000 goal-directed human search paths and identify strategies people use when navigating information spaces. We find that human wayfinding, while mostly very efficient, differs from shortest paths in characteristic ways.

2 OBJECTIVES OF THE STUDY

From an analytic perspective, it is important to understand what strategies and clues people use to find paths in the Wikipedia information network. In particular, as humans are navigating information networks, they might switch between various strategies. The interplay between the topical relatedness of concepts and the underlying network structure could give us important insights about the methods used by efficient information seekers. Also, the latter often face trade-offs: there may be wayfinding strategies that are safe but also inefficient; on the other hand, by trying to find only the shortest paths, the searcher might get lost more easily.

3 DATA AND METHODOLOGY

The data was taken from the stanford website. All the necessary files have been included along with the submission.

The main correlation between the study and the analysis could be to answer real world questions like 1) How people find their way through social networks?

2) How people find information on the Web, Wikipedia?

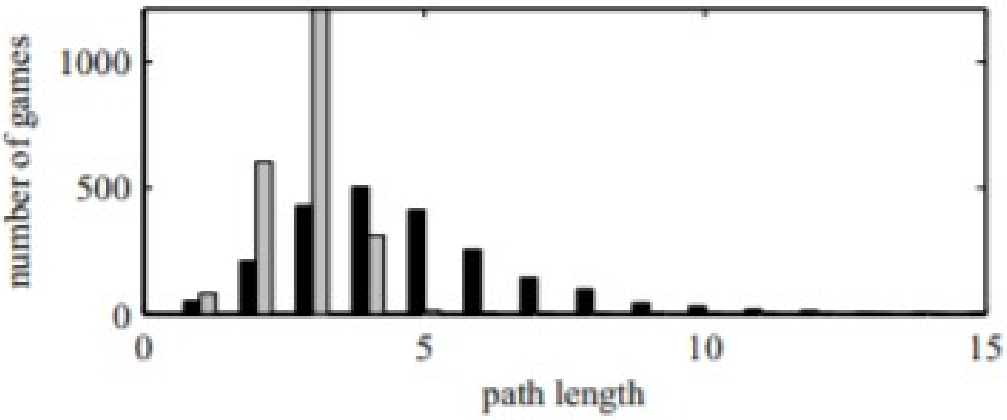

Figure 1: histogram of lengths of 1,694 games; the tail continues up to 30. Gray: histogram of shortest-path solutions to the same games.

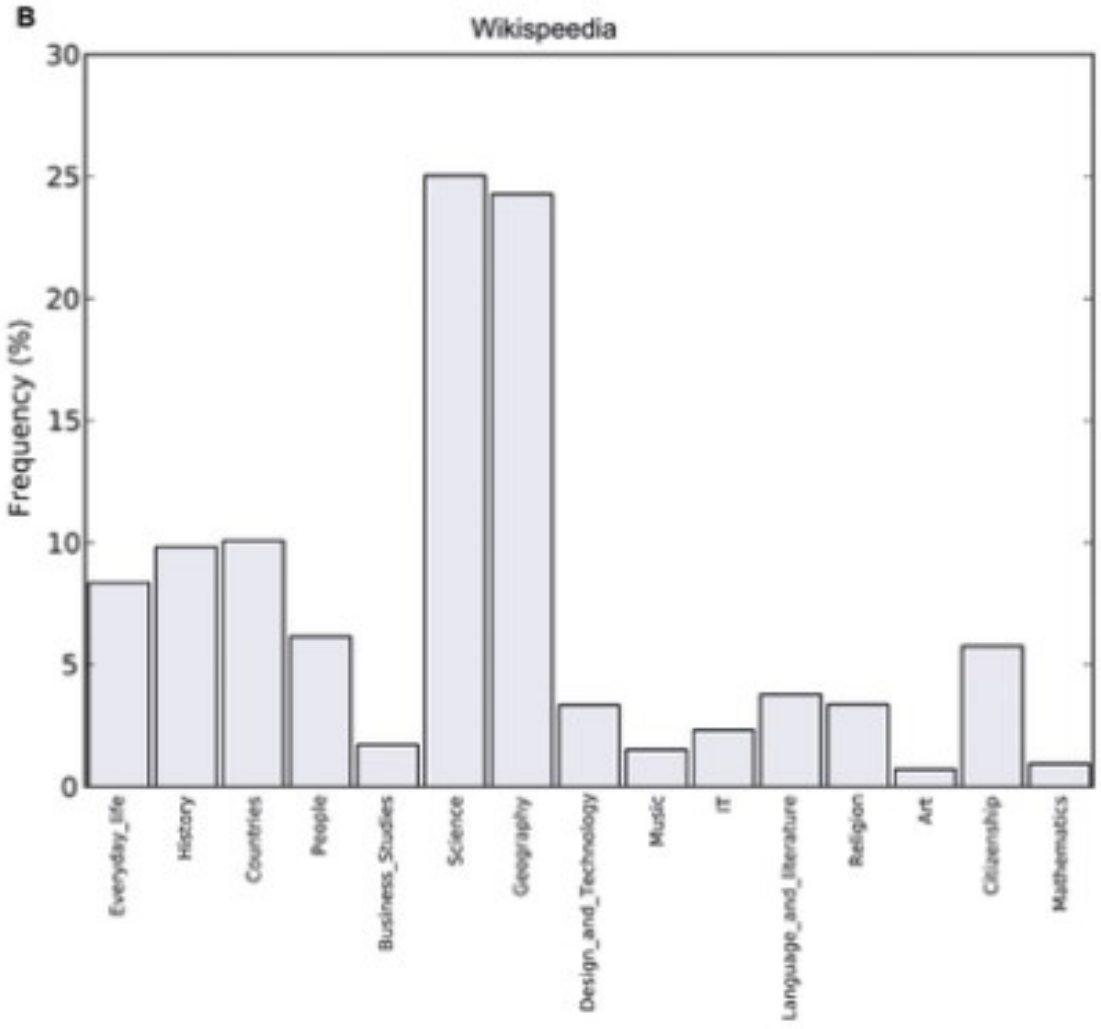

Figure 2: Frequency distribution of topics