Project Proposal - The Wikipedia Game

Mark Kong, Benjamin Rafetto, and Claire Stolz

November 9, 2018

1 Introduction

The Wikipedia game [?] is a path-finding challenge where a player is presented with two disparate topics and is expected to find a path between the two using links between Wikipedia pages. Due to the interconnectedness of various topics and an extremely large branching factor, there are a large variety of possible paths, and the user is rewarded for finding as short a path as possible. Our goal is to solve this game in as efficient a manner as possible.

2 Background and Related Work

There has been a small amount of theoretical work done on this subject. The Wikipedia game itself is free to play from www.thewikigame.com, while there are websites online that will solve for crude solutions, such as [?].

There is also evidence that approaches such as combining DFS for several levels and then using BFS might be an effective solution (cite www.cs.princeton.edu/ rs/talks/PathsInGraphs07.pdf). We intend to build on this previous work and explore innovative new approaches using clustering algorithms and natural language processing.

3 Problem Specification

Specifically, given two Wikipedia entries the problem is to find a path between the two, with rewards determined by the distance of the path. We will attempt to find close to optimal paths with minimal work.

4 Approach

We intend to to use various search algorithms, as well as clustering, classification, and potentially learning to identify optimal weightings for various heuristics and find solutions as efficiently as possible.

Initially, we will need to scrape and download as large a subset of Wikipedia as possible, and then a simple Breadth-First Search algorithm will be employed to find the optimal paths between various topics. Our goal will then be to refine a new algorithm that performs as efficiently as possible while sacrificing a minimal amount of optimality. To this end, we will try to explore and expand the efficient frontier between nodes expanded and loss relative to optimal solutions.

We will explore various algorithms to do so, but our research will primarily be focused towards greedy and A* search, and especially on coming up with efficient heuristics for distances between various topics. All heuristics will likely require us to consider a measure of distance between topics, while also trying to explore in the direction of topics with larger numbers of branching nodes. Topics with larger branching factors should allow us to find shorter solutions, though at the expense of additional node expansion if we move in an inefficient location (we can potentially use learning to come up with optimal weightings between opting for closer topics versus ones with more branches).

A significant challenge will be to come up with these measure of distance between various pages. We will potentially explore clustering based on pages with large numbers of inter-connections or similar connections, hierarchical sub-clustering, and NLP topic-modeling to come up with a variety of measures of distance, which we plan to evaluate relative to each other or in combination to find efficient algorithms.

A System Description

Appendix 1 We plan to download a large portion of the Wikipedia database, then setup a python tool which can be run with either randomly chosen topics, either uniformly or chosen to have specific properties (sufficiently far, sufficiently close, etc), or specified by the user. Our tool will then run a variety of approaches to find the optimal distance between the two, and then return the found paths to the user with information as to how much work was required before the path was found.

B Group Makeup and Distribution of Labor

Appendix 2 Our team consists of Mark Kong, Benjamin Rafetto, and Claire Stolz. Mark Kong and Benjamin Rafetto will focus primarily on various clustering approaches to find efficient heuristics, while Claire Stolz will work on the NLP topic-modeling heuristics.