CS685: Data Mining- Assignment 2

Debanjan Chatterjee (20111016)

20th, November 2020

Abstract

The aim of the assignment was to analyze user data from an online word association game called "Wikispeedia". The dataset used- 'wikispeedia paths-and-graph' is very robust in nature. The assignment helps us grasp the basic fundamentals of graph mining or network analysis, and how online navigation games can have underlying real world research objectives.

1 Introduction

About the game: "Wikispeedia" is a word association game that is played on Wikipedia.com. In this game, human players are asked to travel between two Wikipedia pages. The player must start at one page and reach the second one (the target page) exclusively by following hyperlinks found in the encountered pages. The goal is to minimize the number of intermediate pages to get to the target page. Step-by-step backtracking is also allowed, and is considered a free move that does not count towards the total.[WPP09]
In the next section we will look at the analysis methods that were used and the conclusions drawn from the results.

2 Analysis

The analysis has been carried out in step by step manner as followed.

2.1 Articles

The number of articles observed is 4604, which shows that a condensed version of Wikipedia is used.

2.2 Categories

Each article can be classified under some category. From the dataset, we observe, there are a total of 146 categories. Now, these categories have been arranged in a hierarchical or a tree-based structure, with the category 'subject' appearing at the root of the tree. The use of this tree based ordering, will help to find further correlations between categories, as parent-child relationships between the categories and articles can be established from them. The articles are also mapped with their corresponding categories.

2.3 Article Graphs

The articles are arranged in a graph data structure, with the articles as vertices and the edges as links between the articles. There are a total of 119772 directed edges between the articled. A connected component analysis has also been performed, assuming the graph to be undirected, to get an idea

of number and size of components. There are a total of 14 components, out of which 12 articles are isolated, thus no other articles can be reached from these. The other two components have 3 and 4589 articles respectively.

2.4 Human finished paths analysis

The human finished paths have been explored and the length of the human path have been compared to the shortest path. While considering the human or user paths, two cases have been considered one including back-clicks and one that does not include back-clicks. An interesting observation has been made, around twenty percent of the human paths are exactly same as the shortest path, and only a small one (or three) percent of the human paths exceed the shortest path by a length of 10. The result is particularly interesting as the even though the users don't have any idea of the underlying graph structure of the articles, or can't apply shortest path algorithms such as Dijkstra's, twenty percent of the times they are still able to match the shortest path distance. This shows a significance of the semantic understanding of the articles by users and the ability to find intuitive relationships between the articles to reach the target article (or common-sense knowledge).

2.5 Categories in path analysis

Across all finished human paths, the number of paths and times each category appears, has been found out and compared with the number of paths and times each category would have appeared in the corresponding shortest paths. These values can be further be used to find various comparative and correlation based analysis.

Also, from the articles, their categories have been derived, and accordingly, source and destination category pairs have been derived from finished and unfinished human paths files. For each such source-destination category pair, percentage of finished human paths and unfinished human paths have been computed. These values are useful to do further analysis to find for which category pairs, the users have struggled to finish the game.

Also, for each source-destination category pair in finished human path, the ratio of human path to the shortest paths have been determined. These values are useful for estimating, for which category pairs, the users have found the game relatively easier.

3 Conclusion

Computing the semantic distance between real-world concepts is crucial for many artificial intelligence applications. The way that a human player goes about this game is strikingly different from the way by which a computer would play. A computer would be able to find the shortest path between the start and goal page through a simple algorithm. This is not possible for a human, however, as they do not have all of the necessary information to know what the shortest path would be. Thus, a human player relies on semantic background knowledge to get to the target page. A common strategy for players is to first attempt to reach a general concept whose page has many outgoing hyperlinks. After this phase of getting away from the first page, a player can then begin to hone in on the target page.

In web navigation, the users goal and whether she reached it, is typically unknown. This makes navigation games particularly interesting to researchers, since they capture human navigation towards a known goal and allow performing analysis and building datasets, as illustrated in this assignment, which can used by machine learning models with the purpose of inferring semantic relationships among articles and their corresponding categories.

Thus, games like "Wikispeedia" have an underlying research objective of making computers learn common-sense knowledge.

References

[WPP09] Robert West, Joelle Pineau, and Doina Precup. "Wikispeedia: An online game for inferring semantic distances between concepts". In: *Twenty-First International Joint Conference on Artificial Intelligence*. 2009.