Week 4: Analyzing Centrality in Wikipedia Navigation Networks

Amish Rasheed

2025-02-22

Introduction

This analysis explores human navigation paths on Wikipedia using the **Wikispeedia** dataset. The dataset consists of a condensed version of Wikipedia, where users navigate from a source to a target article by clicking Wikipedia links. The objective is to compare centrality measures across different categorical groups of Wikipedia articles.

Wikispeedia

Data Sources

The dataset includes the following files:

- **`articles.tsv`**: Contains a list of articles with their unique IDs and titles.
- **`categories.tsv`**: Maps articles to categorical labels (e.g., Geography, Science, Politics).
- **`links.tsv`**: Defines directed hyperlinks between articles, forming a network.
- **`paths_finished.tsv`**: Captures human navigation paths on Wikipedia, including timestamps and ratings.

High-Level Plan

1. Data Loading

Load and preprocess the dataset using `readr` and `tidyverse` packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
articles <- read_tsv("wikispeedia_paths-and-graph/articles.tsv", col_names = c("ArticleID", "ArticleTitl
```

```
## Rows: 4615 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): ArticleID
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
categories <- read_tsv("wikispeedia_paths-and-graph/categories.tsv", col_names = c("ArticleTitle", "Cate
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 5216 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): ArticleTitle
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
links <- read_tsv("wikispeedia_paths-and-graph/links.tsv", col_names = c("SourceID", "TargetID"))
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 119893 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): SourceID
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
paths_finished <- read_tsv("wikispeedia_paths-and-graph/paths_finished.tsv", comment = "#", col_names =
```

```
## Rows: 51318 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): hashedIpAddress, path, rating
## dbl (2): timestamp, durationInSec
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

2. Graph Construction

3. Compute Centrality Measures

Key centrality metrics to be computed: - **Degree Centrality**: Number of links to/from an article. - **Betweenness Centrality**: How often an article appears in shortest paths. - **Closeness Centrality**: How easily an article can reach others.

4. Categorization

Merge the **centrality measures** with **categories.tsv** to compare centrality across different article groups.

5. Analysis: Comparing Centrality Across Categories

Hypothesis: Certain categories (e.g., **Geography** or **Politics**) will have **higher betweenness centrality**, serving as transition points in human navigation.

Hypothetical Outcome

I expect categories such as **Geography** and **Politics** to have **higher betweenness centrality**, indicating they serve as **key transition points** in navigation. Conversely, categories like **Science.Chemistry** may have lower centrality since they are more specialized and less commonly traversed.