RailGraph: Visualizing Indian Railways Network and Train Flows

B. Tech Project II (CS47006)

Bachelors of Technology
in
Computer Science and Engineering

by

Sourav Kumar Jena
(17CS10052)

Under the Supervision of

Prof. Saptarshi Ghosh

Department of Computer Science and Engineering

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Kharagpur - 721302, India

Apr, 2024

Acknowledgements

I am thankful to my bachelor's supervisor **Prof. Saptarshi Ghosh** for his continuous support, conviction, encouragement, and invaluable advice in bachelor's project work. I also like to thank **Ms. Koyena Chowdhury** for helping me throughout the project review and presentation preparation process.

Sourav Kumar Jena
(17CS10052)

S. Ghosh

Prof. Satarshi Ghosh

Abstract

This project aims to gather comprehensive information on train schedules within the Indian Railways network. Utilizing data from the Cleartrip website, which lists detailed schedules of numerous express trains across multiple pages, a systematic approach is proposed to collect and analyze this data. Initially, the HTML source of five pages containing details of 3,754 trains is collected. Train numbers are extracted from these pages, serving as unique identifiers for subsequent data retrieval. A crawler is then designed to automatically access individual train pages, download their details, and extract the list of stations in chronological order for each train. Unique station names are identified from the source and destination pairs. A graph representation is constructed with stations as nodes and edges representing connections between stations based on the presence of trains. The weight of each edge corresponds to the total number of trains traversing between two stations. Visualizations of the graph are generated with varying plot sizes, colors, and node sizes to enhance visibility and interpret-ability. Furthermore, the distributions of incoming and outgoing trains for each station are analyzed, providing insights into station connectivity and traffic flow within the railway network.

Contents

Chapter 1

Introduction

1.1 Overview

This project aims to gather comprehensive information on train schedules within the Indian Railways network. Utilizing data from the Cleartrip website, which lists detailed schedules of numerous express trains across multiple pages, a systematic approach is proposed to collect and analyze this data. Initially, the HTML source of five pages containing details of 3,754 trains is collected. Train numbers are extracted from these pages, serving as unique identifiers for subsequent data retrieval. A crawler is then designed to automatically access individual train pages, download their details, and extract the list of stations in chronological order for each train. Unique station names are identified from the source and destination pairs. A graph representation is constructed with stations as nodes and edges representing connections between stations based on the presence of trains. The weight of each edge corresponds to the total number of trains traversing between two stations. Visualizations of the graph are generated with varying plot sizes, colors, and node sizes to enhance visibility and interpret-ability. Furthermore, the distributions of incoming and outgoing trains for each station are analyzed, providing insights into station connectivity and traffic flow within the railway network.

1.2 Problem Statement

The Indian Railways system encompasses a vast network of train routes, serving millions of passengers daily. However, accessing comprehensive information on train schedules remains a challenge. The absence of a centralized database or platform providing detailed schedules hampers efficient planning and decision-making for both travelers and railway authorities.

To address this issue, the project aims to develop a solution for collecting, analyzing, and visualizing detailed train schedule data within the Indian Railways network, with a focus on express trains. Leveraging available online resources such as the Cleartrip website, which provides schedules for thousands of trains, the project seeks to automate the process of data extraction and analysis.

1.2.1 Problem Part I

Develop a solution to collect and organize detailed train schedule information for all trains within the Indian Railways network. Utilizing a suggested data source, consisting of five web pages containing details of 3,754 trains from Cleartrip, the solution aims to automate the process of data extraction and analysis.

1.2.2 Problem Part II

Develop a solution to analyze station connectivity and traffic flow within the Indian Railways network, focusing on unique station names, incoming and outgoing trains, and edge attributes such as train frequencies. The solution will involve constructing a graph representation of the railway network and generating visualizations to aid in data interpretation and decision-making.

1.3 Need of the Study

The study addresses the pressing need for comprehensive train schedule data and network analysis within the Indian Railways system. By collecting detailed information on express train schedules, analyzing station connectivity, and developing visualization techniques, the study aims to enhance travel planning for passengers, optimize network efficiency, and support informed decision-making for infrastructure planning and investment. Improved data accessibility and visualization tools can empower stakeholders to better understand complex network structures, identify optimization opportunities, and ultimately enhance the overall passenger experience and operational efficiency of the Indian Railways network.

1.4 Practical Implications

The study has several practical implications for various stakeholders within the Indian Railways system. Firstly, for passengers, access to comprehensive and accurate train schedule information, especially for express trains, will greatly improve travel planning, enabling them to make informed decisions about routes, timings, and connections. This could lead to reduced travel times, minimized disruptions, and enhanced overall travel experiences. Additionally, for railway authorities and operators, the insights gained from analyzing station connectivity and traffic flow can inform strategic decisions regarding infrastructure investments, route planning, and service optimizations. By identifying key hubs, bottlenecks, and optimization opportunities within the network, authorities can allocate resources more efficiently and improve the reliability and efficiency of train services. Furthermore, the development of automated data collection and visualization techniques can streamline data management processes and enhance decision-making capabilities, ultimately contributing to the modernization and improvement of the Indian Railways system as a whole.

1.5 Study Objectives

- Systematically collect detailed information on express train schedules from a suggested data source.
- Design and implement a crawler capable of accessing individual train pages, retrieving schedule details, and storing them for further analysis.
- Extract station information from the collected data, organizing it into a structured dataset for ease of access and manipulation.
- Construct a graph representation of the railway network, with stations as nodes and train routes as edges, facilitating network analysis and visualization.
- Analyze station connectivity, traffic flow, and train frequencies within the railway network, providing insights to stakeholders for optimization and decision-making purposes.

1.6 Scope of the Study

- Collecting detailed information on express train schedules from the Cleartrip website.
- Developing a crawler to access and parse train schedule data from multiple pages efficiently.
- Analyzing the collected data to extract station information and construct a graph representation of the railway network.
- Exploring various visualization techniques to represent the network graph and analyze station connectivity.

Chapter 2

Methodology

To collect detailed information about train schedules in Indian Railways, particularly for all express trains listed on the Cleartrip website, the following steps can be followed for data collection:

2.1 Data Collection

Collect HTML Source: Use web scraping techniques to fetch the HTML source of the provided pages containing the details of express trains on Cleartrip. This can be done using libraries like BeautifulSoup in Python.

2.2 Data Extraction

Parse the HTML source to extract all the train numbers listed on these pages. This can be achieved by identifying the HTML elements that contain the train numbers, Train Name, Source, and Destinations and extracting their text content.

2.3 Design Crawler

Develop a crawler or scraper script that iterates through the list of train numbers obtained in the previous step. For each train number, construct the URL of the corresponding Cleartrip pages and fetch the HTML source of that page.

2.4 Download and Store Train Pages

Download the HTML content of each train page and store it locally for further processing. Ensure that the data is organized in a structured format for easy retrieval and analysis.

2.5 Extract Station Information

Parse the HTML content of each train page to extract the list of stations along the train route, including their order of appearance. This information can be located within specific HTML elements or tags that contain the station names and their respective positions in the schedule.

2.6 Graph Construction

To construct a graph representation of the railway network from the collected train schedule data, we'll first identify each station as a node and establish connections between stations as edges. The station information extracted from the data serves as the basis for node creation, while direct train services between stations determine edge existence. Using Python and the NetworkX library, we implement functions to construct the graph, check for direct train services, calculate edge weights based on factors like train frequency, and visualize the resulting graph. This approach facilitates the visualization of station connectivity and traffic flow, enabling insights into the structure and dynamics of the railway network for optimization and decision-making purposes.

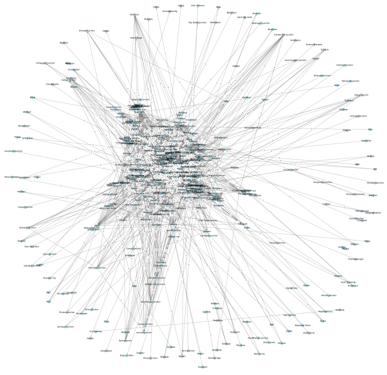

Figure 2.1: Directed Graph with Edge Weights

2.7 Summary

The methodology outlined involves a systematic approach to collecting detailed train schedule information from the Cleartrip website, particularly focusing on express trains within the Indian Railways network. By initially scraping the HTML source of relevant web pages, train numbers are extracted, serving as unique identifiers for subsequent data retrieval. A custom crawler is designed to access individual train pages, download their details, and extract station information. Utilizing Python and the NetworkX library, a graph representation of the railway network is constructed, with stations as nodes and connections as edges. This methodology facilitates the visualization of station connectivity and traffic flow, offering insights crucial for optimization and decision-making within the Indian Railways system.

Chapter 3

Conclusions

3.1 Conclusions

A systematic and comprehensive approach has been established for collecting, analyzing, and visualizing detailed train schedule data within the Indian Railways network. By leveraging web scraping techniques to gather information from the Cleartrip website, a large dataset comprising express train schedules has been obtained, laying the foundation for further analysis. The design and implementation of a custom crawler enable the automatic retrieval of train details, ensuring efficiency and scalability in data collection. Additionally, the extraction of station information and the subsequent construction of a graph representation of the railway network provide valuable insights into station connectivity and traffic flow dynamics. These insights have practical implications for stakeholders, including passengers, railway authorities, and operators, by enhancing travel planning, optimizing network efficiency, and supporting informed decision-making. Overall, the methodology presented offers a robust framework for understanding and improving the Indian Railways system's operational performance and passenger experience.

3.2 Scope for Future Work

Integration of Machine Learning Algorithms:

- Incorporate machine learning models for predictive analysis.
- Forecast train delays and optimize route schedules using historical data.
- Improve overall network efficiency by leveraging predictive analytics.

Real-Time Data Integration:

- Integrate real-time data feeds from multiple sources.
- Provide up-to-date information on train statuses and schedule changes.
- Enable dynamic analysis and decision-making based on real-time insights.

Collaboration with Stakeholders:

- Collaborate with railway authorities and stakeholders.
- Implement data-driven decision-making processes.
- Address practical challenges and improve service delivery in the Indian Railways system.