



INFORMATICS
INSTITUTE OF
TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Automatic summary generation of sunburst charts by data extraction

A dissertation by

Venushi de Silva (20191028/W1790291)

Supervised by

Ms. Sulochana Rupasinghe

Abstract

Sunburst charts are a popular data visualization tool used to represent hierarchical data structures. They allow users to view the data in a circular format where the outer rings represent higher-level categories, and the inner rings represent lower-level sub-categories.

However, reading sunburst charts can be difficult to read due to their circular layout and nested hierarchy. The viewer must decipher the size and colour of each wedge and navigate the different levels of categories. Additionally, too many categories or unevenly distributed data can result in clutter and confusion.

By extracting summary data from a sunburst chart, it is possible to quickly identify key trends and patterns in the data, such as which categories are the most significant or which subcategories are the most popular. This can be especially useful for business analysts and marketers who need to quickly analyze large amounts of data and make informed decisions based on the insights they uncover. In addition, generating a summary of a sunburst chart by extracting data can help to save time and improve efficiency. Rather than having to manually sift through a large amount of data to identify patterns and trends, analysts can use automated tools to extract the most relevant information and focus their analysis on the areas that are most important.

According to the research papers, there is no existing tool for extracting data and generating summaries of sunburst charts. As a result, the author opted to implement a system that utilizes image processing and optical character recognition (OCR) to extract the data and generate the summary for these charts. Although there were multiple methods available for developing such a project, the author chose to use image processing and OCR.

ACRONYMS

IDE Integrated Development Environment.

ML Machine Learning.

NLP Natural Language Processing.

OCR Optical Character Recognition