Synopsis

Title:

Detection of Anomalous Navigation Patterns in Wikipedia Clickstream.

Member details:

1) Vikasan S Nayak, 230962012, AIML-A1, Roll No.: 07
2) Dube Prajwal Manoj, 230962024, AIML-A1, Roll No.: 11
3) Dhruva Goyal, 230962071, AIML-A1, Roll No.:19

Abstract:

The increasing reliance on Wikipedia as a primary source of information has generated enormous volumes of navigation data, recorded in the Wikipedia Clickstream dataset. Analysing this data is critical for understanding unusual navigation patterns that could indicate abnormal user behaviour or unexpected traffic flows. This project utilises Big Data Analytics using Apache Spark to process and analyse the Clickstream dataset in a distributed environment. Anomalies are defined in three complementary ways: Statistical Outliers, where navigation pairs exhibit sudden spikes in click frequency compared to historical averages; Topological Anomalies, where new or unexpected navigation paths emerge; and Behavioural Anomalies, where clustering techniques such as K-means are used to detect unusual shifts in article click-out. The system demonstrates the scalability and effectiveness of Spark-based analysis for large-scale anomaly detection, while also providing insights into how collective information-seeking behaviour can shift in response to external events. The proposed system demonstrates the scalability and effectiveness of Spark-based machine learning techniques for large-scale anomaly detection in user navigation data.

Dataset Characteristics:

The Wikipedia Clickstream dataset contains counts of (referrer, resource) pairs extracted from the request logs of Wikipedia. A referrer is an HTTP header field that identifies the address of the webpage that is linked to the resource being requested. The data shows how people get to a Wikipedia article and what links they click on.
Each row in the dataset represents how users arrived at a specific article from another source (either within or outside Wikipedia). The data includes the following four fields:

- *prev*: The referrer, or the page the user came from. This can be an internal Wikipedia article, an external search engine, or another external site.
- *curr*: The resource, which is the Wikipedia article the user is currently on.

- *type*: The navigation type, which describes the relationship between the referrer and the resource. For example, a "link" if the referrer and resource are both articles and the referrer links to the resource, or "other" for different types of navigation.
- *n*: The number of occurrences of the specific (referrer, resource) pair. This represents the click frequency.

Objectives:

1) To load and preprocess the Wikipedia Clickstream dataset using PySpark RDD and DataFrame operations.
2) To define and detect anomalies through multiple approaches:
    a) Statistical outliers in navigation pair click frequencies.
    b) Topological anomalies arising from the sudden appearance of new navigation paths.
    c) Behavioural anomalies detected through K-means clustering on the article click-out.
3) To extract relevant features and metrics for clustering.
4) To analyze and interpret anomalous clusters to identify bot-like behaviors or suspicious navigation flows.

References/Sources:

1. https://dumps.wikimedia.org/other/clickstream/
2. https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream