Abstract

In recent years, digital streaming platforms are widely being consumed by the users as the information related to movies and entertainment are available on their fingertips. However, it is up to the users to make the selection but to make the selection form vast categories and overwhelming chooses it becomes difficult for users to know that what they are going to watch next. That's where movie recommendation using content-based filtering comes in handy.

Content based movie recommendation system relies on the attributes like genres, actors, directors, description of plot like keywords and tagline. To measure the similarity between the movies Jaccard algorithm is used, which finds the common features in movies, converts them into set vectors and provided the recommendation as needed. Along with that K-nearest neighbour algorithm is a machine learning algorithm which finds the K nearest movies which are similar to each other based on the different factors and features provided. It also shows that big data technology when integrated with machine learning like k-mean clustering can provide better result. To implement the system the data is collected which has large number of attributes, which is then clean and pre-processed for further process, which involves performing data analysis using Hive with Hive query language and to represent the data graphically the visualization tool called Power BI is used. As a recommendation after calculation and identification of attributes recommendation is performed using above-described algorithms, which provides the user with similar movies as required.

The strength of this movie recommendation system lies on its contents, through which the personalized recommendation can be made for an individual, it is simple and can handle sparse data making it a better system even if it contains missing data. On other hand it has its own weakness as it may lack diversity in the recommendation and it may struggle with complexity of data. Overall, it is the system which can provide good recommendation to the user and has huge potential, which can be achieved through further research and exploration on the field of user's preferences, context and different approaches.

Keywords – Content-based filtering, Movie Recommendation system, Hive, Power BI, K-nearest neighbour, Jaccard Similarity, K-mean Clustering, Big data technology