Chapter 6 – Conclusion and Discussion

In the field of entertainment, the popularity of recommendation is growing rapidly. Personalizing the content, engaging the viewer, providing good content is vital thing. In case of movies the systems are being designed to provide the users with good movies based on their preferences and their past history. In this project, content-based filtering has been implemented for suggesting movies to the viewer with the use of Jaccard similarity and K-nearest neighbour algorithm. And shows that big data technology like(pySpark) can be used for creating content-based movie recommendation system. As the system relies most on the content of movies, the system has been built on movies dataset where it contains the features like title, genres, cast, crew and other plots of movies.

The data has been cleaned, pre-processed and relevant features has been extracted. The feature vector has been created for each movie and after fitting the model to the system, Jaccard similarity has been used to calculate the similarity between the movies by their similarity scores, Jaccard metric is used to measure the similarity of two sets with the size of their intersection to the size of union providing common sets to recommend the movies based on the given features. With k-nearest neighbour the k most similar movies have been derived and provided as the recommendation to the user. And as a result if the system is asked to provide the recommendation for the similar movies based on different genres like drama, action, horror, thriller and so on, it will provide the list of movies on those specific genres, similar for the cast where if the actors names are searched their movies are provided, If the user wants to search the movies on directors name the movies made by the specific director will be displayed and finally if the title name of the movies is provided then the similar kind of movies are generated and listed which will be based on different features like keywords included in those movies, taglines of the movies which provide the unique identity to the movies and other features related to plot of movie. Not just for analysis big data technology proves that it can be used as a good option for recommendation system when used properly with machine learning algorithm and libraries. It is able to harness the power of big data technology and versatile capability of machine learning algorithm able to provide fast and better movie recommendation system.

For the analysis of data, big data technology Hive with Hive query language has been used to provide the insight of data. Revealing the important things and aspects related to data which is not provided with recommendation system. As it can handle large amount of data more features have been taken and important information has been extracted. And to make those analysis more appealing to the eyes, the best visualization tool for big data has been used named Power BI where the data is shaped and provided in the form where it is easy to understand the data even for general user. Overall representation of data in charts, graphs and so on has been carried out, which basically helps different organizations in decision making process.

This approach has some advantage over other system like collaborative filtering where it is hard to handle sparse data and focuses more on similarity of user rather than content. But saying that, content-based filtering system has also limitation in itself such as it limited diversity in recommendations, as it focuses more on content, it may lack the ability to capture user preferences, it may not be able to consider the context or situation of user. And the limitation of algorithm may not be able to handle large amount of data, scalability may decrease the performance of system. There is always room for improvement where user and content can be brought together by removing the limitations for each approach.

To obtain the aim of the project, there were some objectives that needed to be achieved. This leads to the research questions which has be answered throughout the whole project. Let's discuss some answers that has been achieved in this project

RQ1: What are the impacts of different kind data in movie recommendation system?

Going through the introduction and literature review part of this projects, there are several things about importance of data and how the types and their variation can have huge impact on project, not every data is useable for every project. The use of data may depend on the type pf project, in my case as the system is based on content of movies only the related data like plots of movies, actors, crew, genres and some relevant contents are taken into considerations

RQ2: What techniques can be used to reduce and identify the best content for user according to their interest?

As discussed in methodological section of this project, there are different techniques that are being used in this project starting with filtering algorithm, machine learning and big data technology. Use of filtering algorithm the process was bit slow but was able to provide better result, machine learning was able to process data quickly and had required clean data and big data technology was able to handle huge dataset and process result quickly.

RQ3: How can recommendation system handle incomplete and sparse data and can make prediction even with absence of data?

During the implementation phase Jaccard similarity algorithm was able to handle sparse data or it can be said that in the absence some data, it was able to provide far better results. As it may have taken some time to get the result but was able to provide good recommendation.

RQ4: How to provide recommendation based on content-based filtering with the large number of contents?

As the results obtained from analysis and results part of this project, the conclusion can be derived that there are different technologies have their own strength, some are good at handling large number of data, some can process it better and some can provide better results than other, these all techniques and technologies used show that proper recommendation can be made based on content of movies and are able to process and handle huge data.

Overall, Movie recommendation system based on content-based filtering has huge potential and has some limitations which may affect its performance in certain scenarios, so it is essential to explore its potential and make use of new technologies and develop different techniques to make the system more comprehensive and robust.