METHODS FOR TACKLING COLD START PROBLEM IN RECOMMENDER SYSTEMS

Supervisor: Ceni Babaoglu

Course: Big Data Analytics

Code: CIND 820

Prepared By: Abdur Rahman Mahbub

Student ID: 501142642

Submitted On: 26th October, 2023

Table of Contents

Tables & Charts

Abstract

The recommender systems help users discover preferred items in modern e-commerce platform, making it an essential part of the digital economy. It is a significant machine learning model that learns from historical data and auxiliary source to provide personalized user recommendations, increasing user engagement and resulting in higher satisfaction. It plays a pivotal role in almost all kinds of e commerce platform However, due to some systematic constraints, these systems often face difficulties in delivering reliable results over all hampering its primary objective. Users need to search for preferred items from a wide range of items and products and the system needs to execute that preference from different subcategories to satisfy that need. When it comes to a highly dynamic platform, many users typically join and adding new items makes it difficult for the system to predict appropriate outcomes. Cold Start Problem directly impacts the recommender system to perform accurately and deliberately force it to be unreliable.

Introduction

Despite the advantages of recommender system, recommender systems often grapple with the cold start problem. This issue arises when the system struggles to provide accurate recommendations for new users or items with limited historical data, primarily due to a lack of user interaction data. This problem typically arises in the following two context.

1. **User Cold Start:** When a new user interacts or appears on the platform without any history or context, the system fails to understand the preferences of the user due to lack of information and suggest irrelevant items.

2. **Item Cold Start:** When a new item is introduced into the vast product catalog, there is a lack of historical user data associated with this item. The system confuses about how to categorize or response to that particular item as they don't have any prior history.

this project proposal aims to propose a pertinent technique to tackle cold start problem in recommender systems, a challenge that holds significant implications for customer satisfaction and business success in the e-commerce platform. By employing the desired dataset and a blend of techniques and tools, this project aspires to offer practical solutions to improve the precision and effectiveness of recommendations, catering to both new and existing users.

Problem Statement

Optimization of product recommendations is a vital issue we seek to address here. To solve problems like handling new product and user introductions, improving the accuracy of recommendations, this project aims to propose a pertinent technique to tackle cold start problem in recommender systems which can give users more personalized and up-to-date recommendations.

Research Questions

The following research questions have been identified to guide this project:

1. How demographic and behavioral data can be leveraged to provide user specific recommendations mitigating cold start problem?

2. How can we use novel strategies when the user interaction is minimal to none or new items are entering the market to enhance the accuracy and personalized nature of product or item recommendations?

3. How can we assess the performance of the advanced recommender systems comprehensively and ensure that the provided recommendations are accurate for a diverse user base?

Dataset Selection

For this project, we have primarily selected the UC Irvine Machine Learning Repository's Online Retail Data Set. After reviewing and analyzing rigorously, MovieLens dataset seemed to be more aligned with the project preference and its applicability for mitigating cold start problem. Public access to the dataset is available and can be viewed or accessed using the following link:

https://grouplens.org/datasets/movielens/

There are several dataset available with the title “MovieLens”. For this project, "MovieLens Latest Small" dataset is selected due to its relevance, richness of data, availability, scalability, and applicability. It is the most suitable dataset, as it demonstrates real scenarios that give an idea of how possible mitigation techniques could be adopted to address the cold start problem.

Tools

In the areas of data preparation, model development and evaluation, we will use industry-standard tools and libraries to efficiently deal with cold start problems in recommender system. For data analysis, modelling and evaluation, Python will be the main programming language. Data manipulation and machine learning, as well as collaborative filtering will be facilitated through libraries and frameworks such as pandas, Scikitlearn or Surprise. The efficient handling of large quantities of data will benefit from cloud computing resources.

Literature Review

Recommender systems plays a vital role in digital landscapes, from e-commerce to movie streaming. It allows customers and viewers to receive personalized recommendations based on their demand and requirements. Nevertheless, there are several underlying problems with the system, "Cold Start Problem" is one of the prominent ones which is well acknowledged. This has become a prevalent concern among researchers and practitioners especially on e-commerce platforms. The Common Understanding of the “Cold Start Problem" in recommendation systems has been explored in this literature review. In real-world scenarios where new products and users are constantly brought forward and data changes regularly, this project offers practical solutions to improve the precision and effectiveness of recommendations, catering to both new and existing users in ecommerce platform. To justify the importance of this issue, existing research papers and journals are critically analyzed in the light of previous studies are evaluated using "MovieLens Latest Small" dataset as the core data source.

Common Knowledge

Recommender systems (RS) is a specific type of intelligent systems, which exploits historical user ratings on items and/or auxiliary information to make recommendations on items to the users. It plays a critical role in a wide range of online shopping, e-commercial services and social networking applications (Jian Wei, 2017).

RS are systems that are used to generate meaningful recommendation of items or products. The amount of information generated and available today is just overwhelmingly huge and its is becoming difficult to access relevant information from the severally available sources. It is illogical and if not impossible to go through all items to decide whether they are useful or not and thus finding relevant items has become more and more challenging. RS were conceived and created to solve this problem (Odumuyiwa & Oloba, 2021).

Despite the traditional RS funnelling and presenting recommendations to their users in a preferable way, design limitations make its overall outcome unreliable. This predicament arises when new users or items join the platform, leaving the system with scant data to make meaningful recommendations. The difficulty of providing an appropriate recommendation in cases with insufficient or no historical data on new users or items is known as Cold Start Problem (CSP) (Herce-Zelaya, Porcel, Bernabé-Moren, Tejeda-Lorente, & Herrera-Viedma, 2020).

In the case of group recommendations, however, the data sparsity may be particularly challenging since the preferences and interactions of a group of users are not individually assessed but rather combined (Yin, Wang, Zheng, Li, & Zhou, 2020). This may result in situations where it is not possible to provide the necessary and appropriate advice, with limited or no sharing of experience between users, creating an incorrect scenario.

Critical Analysis of Common Knowledge

CSP in RS is highly complex and requires an in-depth and intensive analysis. Though factors disrupting RS are well acknowledged, critical valuation and fundamental research is the only feasible way to develop more efficient and user-friendly solutions that will benefit customers and businesses.

Traditional recommendation algorithms must be complemented by an extensive analysis of new strategies to increase accuracy and personalization of recommendations, while they struggle with cold start users. The question of how different recommendation techniques deal with the challenge of providing personalized recommendations on both popular and niche products should also be examined critically (Zhang Z.-K. , Liu, Zhang, & Zhou, 2010). While personalization is necessary to satisfy each user's specific preferences, many products are likely to be overlooked in favour of an over-personalized recommendation. It is essential to critically analyze how to reach an appropriate balance between providing relevant products for users and introducing diversity to keep the user experience exciting and stimulating.

Standard metrics are sometimes used in traditional recommendation systems for evaluation. These matrices are not appropriate indicators to assess the quality of recommendations in retail settings, particularly when a Cold Start scenario is involved (Zangerle & Bauer, 2022).Therefore, A critical analysis is needed to define and adopt comprehensive evaluation metrics consistent with the objectives of RS.

Similar Projects

After carefully analyzing few journals and research papers online, I have found several research efforts addressing the CSP in RS. Articles like “Dealing with the new user cold-start problem in recommender systems: A comparative review” (Son, 2016) “Collaborative Filtering and Deep Learning Based Recommendation System For Cold Start Items” (Jian Wei, 2017), & “Facing the cold start problem in recommender systems” (Lika, Kolomvatsos, & Hadjiefthymiades, 2014) has similar approaches. Although most of these studies have tackled CSP and have addressed it, the specific feature of this study is its application to combine techniques in one platform to mitigate CSP in one go.

While this report seeks to evaluate the effectiveness of the methods we are going to use for mitigating CSP, existing studies have already looked at similar topics in different contexts or dataset and sometimes focused only on single or may be two methods. We will also implicate a broader picture comparing the traditional system and the upgraded system using the new methods. The particular strategies and techniques which will be applied to this report may differ slightly from previous research in terms of their purpose, but they all share a common objective to combat CSP in RS.

Related Works

Indeed, CSP in RS addressed by a great number of research papers with innovative approaches which has been elaboratively described. Collaborative Filtering (CF) is the most classical techniques that's used for the rating user predictions (Jian Wei, 2017). Among other popular methods, Matrix Factorization (MF) is widely used as it enhances overall recommendation accuracy and coverage of listed items (Zhao, Sun, Han, & Peng, 2016). Another popular method is Content-based Filtering (CBF) technique that recommends items to users according to the interests of the users (Zhang S. , Liu, Yu, Feng, & Ou, 2022).

However, the use of these methods in a specific dataset and the challenges posed by RS may vary, which makes it necessary to assess their effectiveness in this context.

Project Significance

This work relates to a wider scope of research about the CSP in RS. However, its unique contribution lies in applying the mentioned methods. This specific research plan will provide practical insights that might contribute to the development of more accurate and robust RS rather than elaborating techniques on their own.

Aligning with the work titled “Hybrid recommendation system combining collaborative filtering and content-based recommendation with keyword extraction”, this project is also designed to implicate the impact of the cold-start and the sparsity problems through applying Hybrid Recommendation System and also compared the performance in terms of precision, recall, novelty, and diversity (Zhang S. , Liu, Yu, Feng, & Ou, 2022).

Another purpose of this project is to analyze the effectivity of the popular techniques that’s used to mitigate CSP in RS. Project titled “Dealing with the new user cold-start problem in recommender systems: A comparative review” has proposed a similar idea where the author tried to analyze methods and provide recommendations how we can produce more effective techniques to resolve the matter (Son, 2016).

Likewise, the article “Addressing Cold Start Problem in Collaborative Filtering using Demographic Data with Entropy-based Methodology”, this project also aims to demonstrate how demographic data can be combined with CF can reduce overall uncertainty in produced recommendations (Odumuyiwa & Oloba, 2021).

Project Merit

The solution to the CSP using a different database is becoming increasingly important given dynamics and increased reliance on e-commerce platforms. Practical solutions to the CSP may substantially improve user experience and stimulate sales due to changes in consumer behaviour, development of product lines or continuous demand for personalization.

While Prior efforts have already been made in that respect, the unique nature of e-commerce and its subdomains calls for constant research to keep pace with changing trends. The project will also add an additional dimension aligning with prior research efforts.

The “MovieLens Small” dataset is one of the most extensive and detailed datasets we have used in this project for evaluating the performance of RS. It provides an exceptional environment for solving a cold start problem due to its wide range of user interactions, movie metadata and high temporal dynamic.

To evaluate the effectiveness of individual techniques and to see their synergy effects, the project will attempt a combination of several techniques to deal with CSP. The results of this comprehensive approach will provide insight on how combinations of methods can best work under different scenarios that could lead to a more robust RS to mitigate the CSP.

Moreover, this project will evaluate the performance of current RS with the new improved RS to depict a clear picture of the significance of research requirement. An evolving trend can be detected by comparing data from different time periods. The results are likely to be justified or undisputed by the initial findings if the approach is not modified. The methodology will provide practical implications that can be adopted in the e-commerce platform eventually ensuring personalized experiences for new users or items.

Reference Review

As part of the literature review, we summarized the articles that has directly dealt with CSP in RS and have proposed multiple scenarios resolving that issue.

The article "Collaborative Filtering and Deep Learning Based Recommendation System for Cold Start Items" proposed a hybrid recommendation system that combines CF with Deep Learning (DL) Techniques to provide precise recommendations for cold start items. The authors tried to illustrate accurate recommendations using a combination of methods, even when historical interactions with items are not available or may be limited (Jian Wei, 2017).

In the article “Hybrid Matrix Factorization For Recommender Systems In Social Networks”, the authors tried to introduce a novel and efficient probabilistic matrix factorization technique that can link ratings with unique user and item preferences. The result demonstrated how this method significantly improved overall accuracy and coverage of the recommended items (Zhao, Sun, Han, & Peng, 2016).

For handling CSP more efficiently, authors in the article “Facing the cold start problem in recommender systems” have implemented a novel strategy through implementing CF and CB filtering together. The paper suggest that the new model can easily blend this two methods and does not require previous probabilistic models to be adopted. The method can easily avoid complex calculations in generating realistic and accurate recommendations (Lika, Kolomvatsos, & Hadjiefthymiades, 2014).

The article titled “New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests” is based on a recommendation approach prediction model that has been designed using behavioural information. These information is extracted from sources like social media to categorize users based on their behavioural profiles. The authors have introduced a novel system where users are not bound to share personal information and still can receive accurate suggestion that alleviates CSP (Herce-Zelaya, Porcel, Bernabé-Moren, Tejeda-Lorente, & Herrera-Viedma, 2020).

In the article “Solving the cold-start problem in recommender systems with social tags”, the authors have proposed a strategy where social tags are considered as a connecting tool that links users and objects through social tags. The results are evidently visible as the method assists in producing improved RS. Users can efficiently find relevant objects using tags in modified RS (Zhang Z.-K. , Liu, Zhang, & Zhou, 2010).

In the article "RBPR: A hybrid model for the new user cold start problem in recommender systems", authors have proposed a model that makes proper use of the users' feedback information of rating data and implicit data. The proposed model aims to improve the recommendation quality by extracting common latent features of users and items making it clear and realistic. (Feng, Xia, Feng, & Peng, 2021).

Approach

Initially the notable solution that will be used in this project to mitigate CSP is by leveraging of demographic and behavioral data to offer personalized suggestions. Demographic data, including age, gender, and location has the characteristic that contributes directly to the improvement of recommendation accuracy (Odumuyiwa & Oloba, 2021).

The project will attempt to collect data based on minimal user interaction behavior and content attributes. This method is especially useful for new items or user entering the e-commerce platform. These collected information will be clustered initially and a hybrid approach will be taken that combines both CF and CB filtering (Zhang S. , Liu, Yu, Feng, & Ou, 2022).

The project will also employ Matrix Factorization techniques to improve the accuracy and dependency of the output results. This method can mitigate CSP and data sparsity simultaneously. The project will assess effectiveness of the implemented method and will introduce a multi-faceted evaluation framework that will compare traditional and upgraded RS for measuring dependability.

Data Description

"MovieLens Latest Small" dataset presents an ideal scenario for researching CSP gaps in RS. It consists of comprehensive and diverse set of ratings, combined with the associated user and item metadata.

- **Diversity:** The dataset involves a wide range of users with varied demographic profiles which is essential for understanding and modeling the behavior of both mainstream and niche users.

- **Richness:** "MovieLens Latest Small" dataset offers metadata about movies, such as genres, which can be utilized for delivering content-based recommendations. The genre and other metadata can offer meaningful insights into tackling cold start problem.

- **Scalability**: The MovieLens dataset is available in different sizes, ranging from a small dataset to the full dataset containing millions of ratings. This scalability makes it ideal for researchers to test algorithms under different scales. This project have used "MovieLens Latest Small" as its organized and ideal for small scale projects.

- **Structured:** The "MovieLens Latest Small" dataset is well-structured and clean. This allows researchers to focus on the core problem rather than spending excessive time on data preprocessing and cleaning.

Dataset Statistics

The "MovieLens Latest Small" dataset contains a comprehensive view of user movie interactions. The dataset contains four separate data points about links, movies, tags and their ratings. Descriptive statistics of each dataset is summarized below:

Links Dataset

The “links” dataset contains list of movies and information about external sources like IMDB, TMDB etc. The dataset tried to identify movies with unit numbers and showed a link of those movies with IMDB and TMDB. Note below the descriptive statistics of “links” dataset.

index	movieId	imdbId	tmdbId
count	9742.000000	9.742000e+03	9734.000000
mean	42200.353623	6.771839e+05	55162.123793
std	52160.494854	1.107228e+06	93653.481487
min	1.000000	4.170000e+02	2.000000
25%	3248.250000	9.518075e+04	9665.500000
50%	7300.000000	1.672605e+05	16529.000000
75%	76232.000000	8.055685e+05	44205.750000
max	193609.000000	8.391976e+06	525662.000000

Table 1: Descriptive Statistics of “links” Dataset

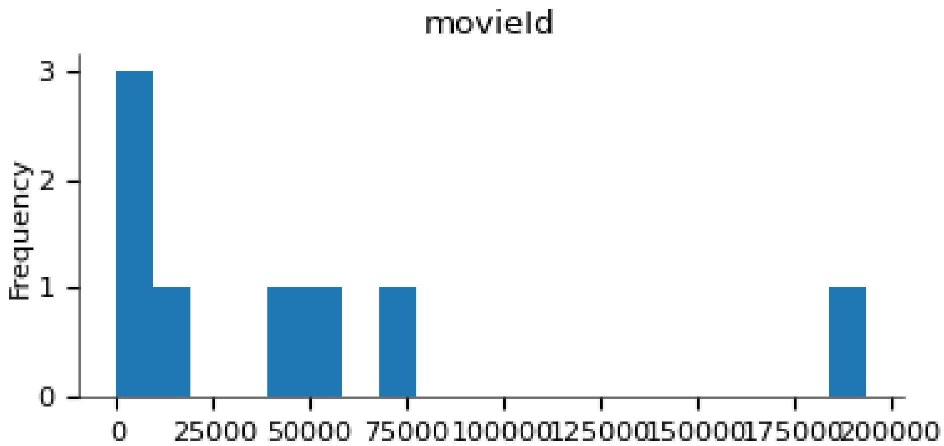

Chart 1(A): Frequency Distribution of “links” Dataset (movieid column)

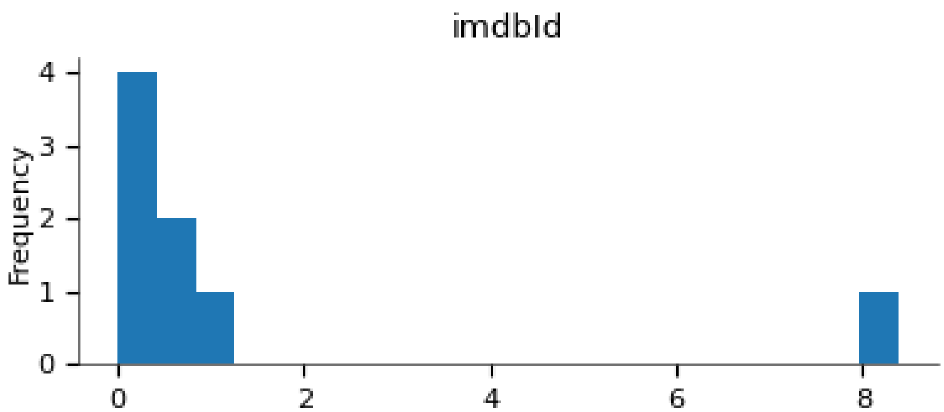

Chart 1(B): Frequency Distribution of “links” Dataset (imdbid column)

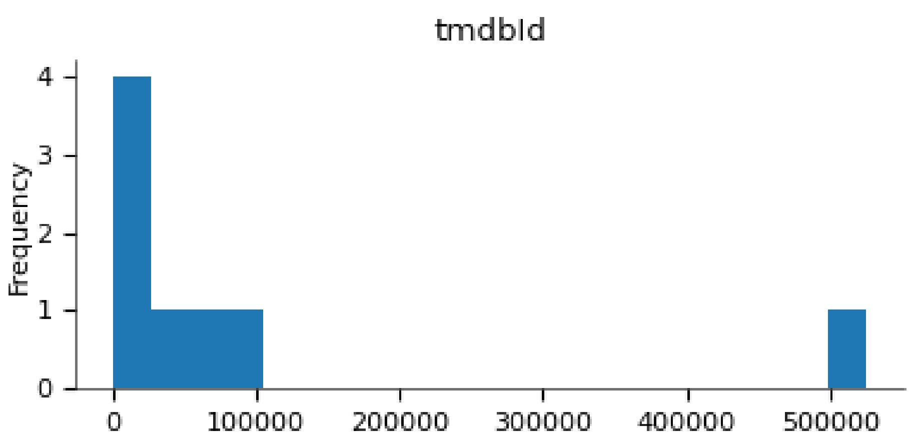

Chart 1(C): Frequency Distribution of “links” Dataset (tmdbid column)

Movies Dataset

The “movies” dataset depicts the details about all the movies involved. It contains the list of similar movies linked with their title and for categorization ease, the genres have been mentioned for each movie for easier recommendation purposes. This will assist in generalizing movie preferences for users on e-commerce platforms.

index	movieId
count	9742.0
mean	42200.353623485935
std	52160.49485443833
min	1.0
25%	3248.25
50%	7300.0
75%	76232.0
max	193609.0

Table 2: Descriptive Statistics of “movies” Dataset

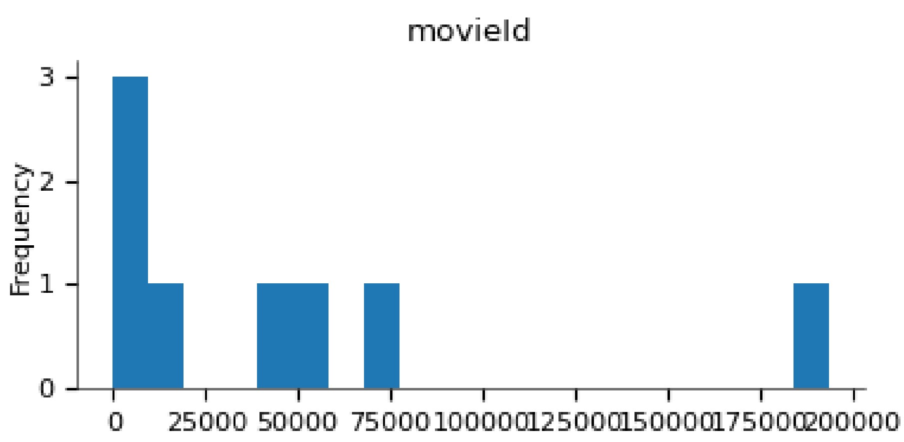

Chart 2: Frequency Distribution of “movies” Dataset

Ratings Dataset

This dataset captures ratings given by individual users. These user are identified spcific ID's and the connection is recognised through ID specific ratings to given item. The ratings are generated within a scale from 1-5 represented by the timing it was given. This dataset is vital for finding users behavioral pattern and their personal preferences.

index	userId	movieId	rating	timestamp
count	100836.0	100836.0	100836.0	100836.0
mean	326.12756356856676	19435.2957177992	3.501556983616962	1205946087.3684695
std	182.61849146349994	35530.98719870018	1.042529239060635	216261035.99513158
min	1.0	1.0	0.5	828124615.0
25%	177.0	1199.0	3.0	1019123866.0
50%	325.0	2991.0	3.5	1186086662.0
75%	477.0	8122.0	4.0	1435994144.5
max	610.0	193609.0	5.0	1537799250.0

Table 3: Descriptive Statistics of "ratings" Dataset

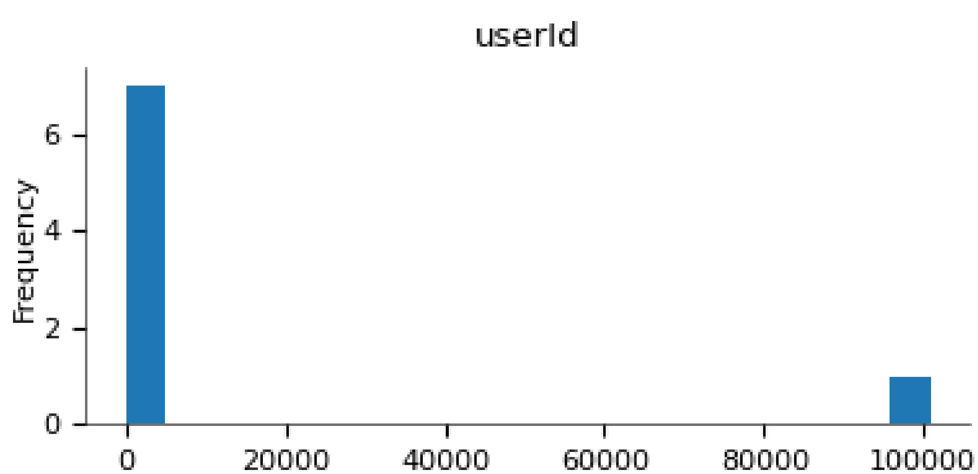

Chart 3(A): Frequency Distribution of "ratings" Dataset (userid column)

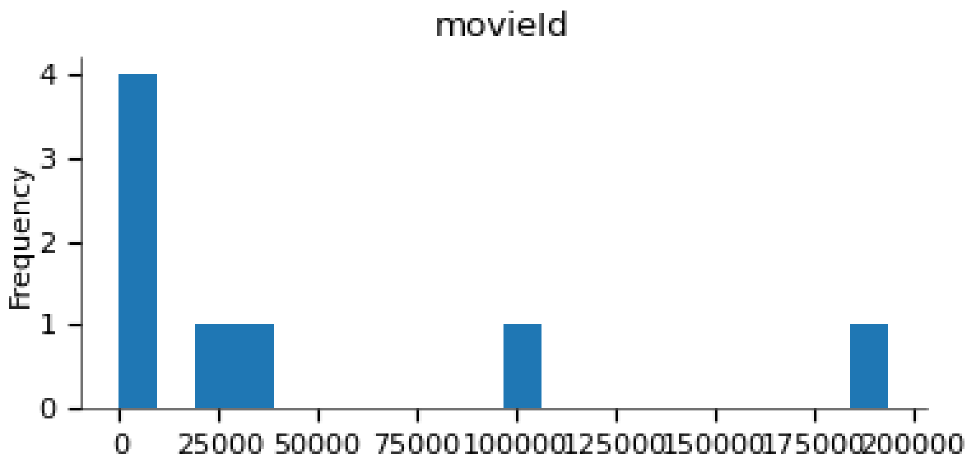

Chart 3(B): Frequency Distribution of “ratings” Dataset (movieid column)

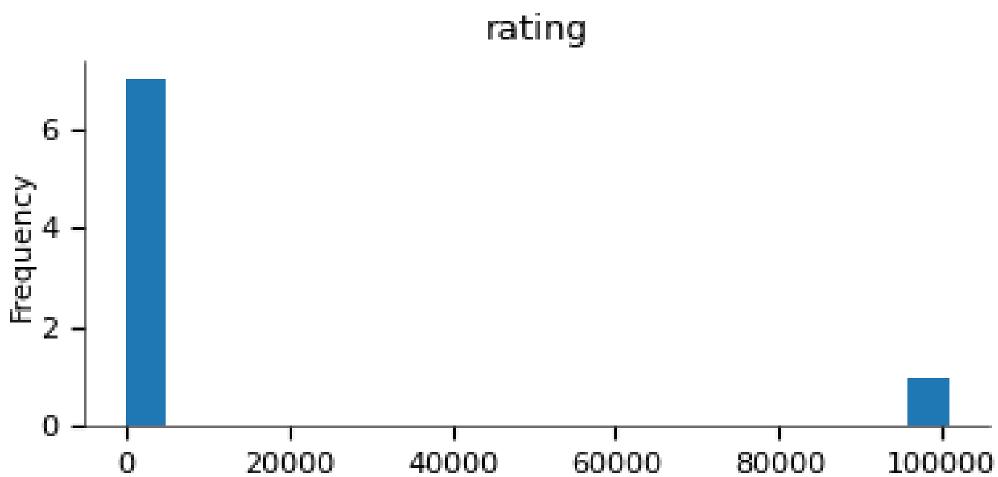

Chart 3(C): Frequency Distribution of “ratings” Dataset (rating column)

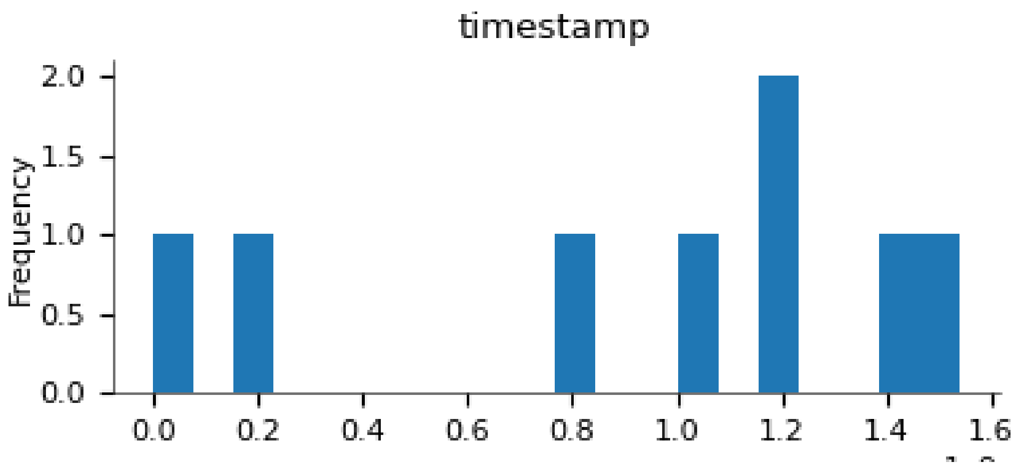

Chart 3(D): Frequency Distribution of “ratings” Dataset (timestamp column)

Tags Dataset

It is an ideal dataset for segregating user profiles based on their content. The database contains tags which is the external content that’s used for defining or categorizing a particular movie. The tag is then linked with the user who provided the tag for a specific movie id.

index	userId	movieId	timestamp
count	3683.0	3683.0	3683.0
mean	431.1493347814282	27252.01357588922	1320031966.823785
std	158.47255348483486	43490.55880276778	172102450.43712643
min	2.0	1.0	1137179352.0
25%	424.0	1262.5	1137521216.0
50%	474.0	4454.0	1269832564.0
75%	477.0	39263.0	1498456765.5
max	610.0	193565.0	1537098603.0

Table 4: Descriptive Statistics of “tags” Dataset

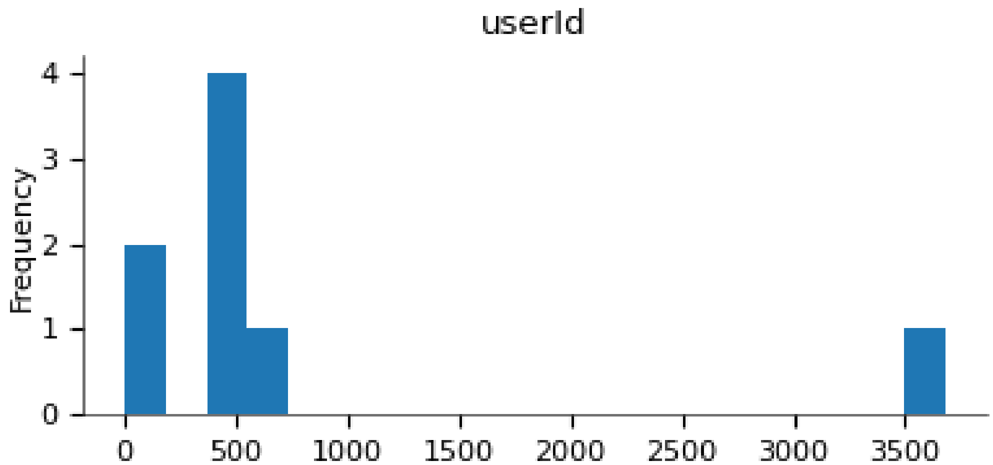

Chart 4(A): Frequency Distribution of “tags” Dataset (userid column)

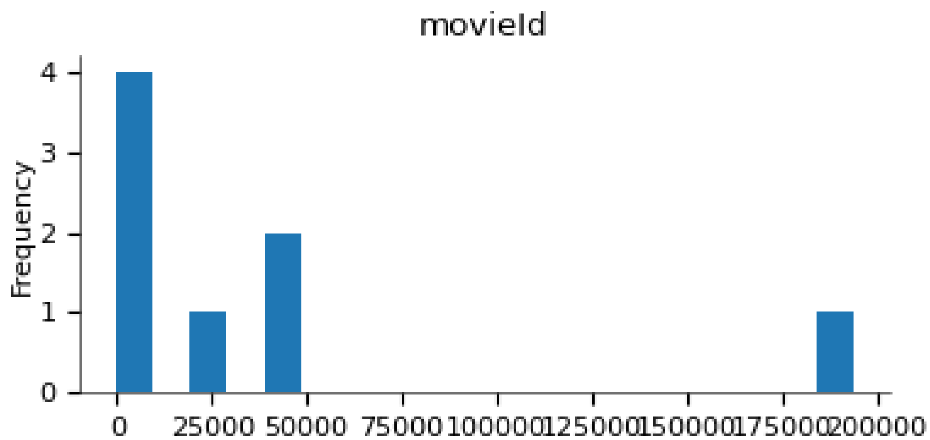

Chart 4(B): Frequency Distribution of Tags Dataset (movieid column)

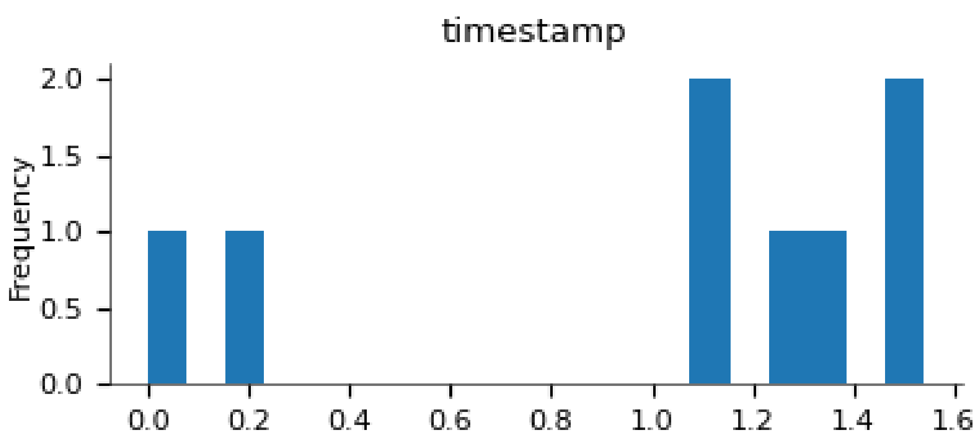

Chart 4(C): Frequency Distribution of “tags” Dataset (timestamp column)

Proposed Methodology

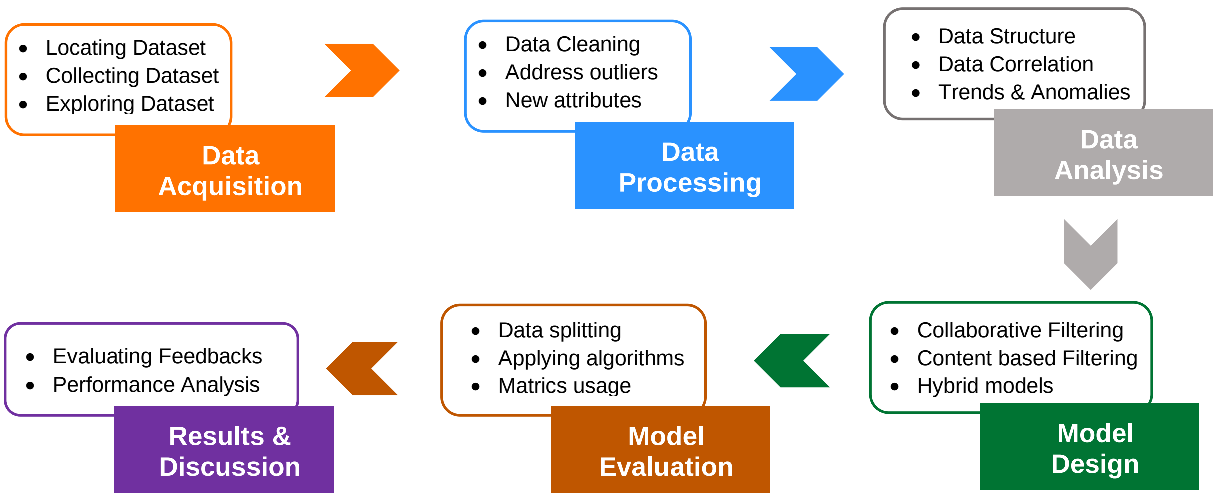

Chart 5: Proposed Methodology (Step by step)

GitHub Repository Link

The Following Github link will contain all the related files of this project. The repository will be updated periodically based on project requirements.

https://github.com/abdurmahbub123/Big_Data_Analytics_Project

References

Duong, T. T., Duong, H. H., & Nguyen, H. S. (2014). A Hybrid Framework for Enhancing Correlation to Solve Cold-Start Problem In Recommender Systems. *the 2014 Seventh IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*. doi:10.1109/CISDA.2014.7035626

Feng, J., Xia, Z., Feng, X., & Peng, J. (2021). RBPR: A hybrid model for the new user cold start problem in recommender systems. *214*. doi:https://doi.org/10.1016/j.knosys.2020.106732

Ferdaous, H., Bouchra, F., Imad-eddine, O. B., & Asmaa, B. (2017). Recommendation using a clustering algorithm based on a hybrid features selection method. *Springer*, 183–186.

Han, H., Huang, M., Zhang, Y., & Bhatti, U. A. (2018). An Extended-Tag-Induced Matrix Factorization Technique for Recommender Systems. *9*(6), 143. doi:https://doi.org/10.3390/info9060143

Herce-Zelaya, J., Porcel, C., Bernabé-Moren, J., Tejeda-Lorente, A., & Herrera-Viedma, E. (2020). New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests. *536*, 156-170. doi:https://doi.org/10.1016/j.ins.2020.05.071

Jian Wei, J. H. (2017). Collaborative Filtering and Deep Learning Based Recommendation. *69*, 29-39. doi:10.1016/j.eswa.2016.09.040

Kim, D., Park, C., Oh, J., & Yu, H. (2017). Deep Hybrid Recommender Systems Via Exploiting Document Context And Statistics Of Items. *417*, 72-87. doi:http://dx.doi.org/10.1016/j.ins.2017.06.026

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing The Cold Start Problem In Recommender Systems. *41*(4), 2065-2073. doi:https://doi.org/10.1016/j.eswa.2013.09.005

Odumuyiwa, V. T., & Oloba, O. P. (2021). Addressing Cold Start Problem in Collaborative Filtering using Demographic Data with Entropy-based Methodology. *6*(4). doi:http://dx.doi.org/10.46792/fuoyejet.v6i4.704

Son, L. H. (2016). Dealing with the new user cold-start problem in recommender systems: A comparative review. *58*, 87-104. doi:https://doi.org/10.1016/j.is.2014.10.001

Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications - Volume 69*, 29-39.

Yin, H., Wang, Q., Zheng, K., Li, Z., & Zhou, X. (2020). Overcoming Data Sparsity in Group. *34*(7), 3447-3460. Retrieved from https://ieeexplore-ieee-org.ezproxy.lib. torontomu.ca/document/9195784/authors#authors

Yu, Y., Wang, C., Wang, H., & Gao, Y. (2017). Attributes coupling based matrix factorization for item. *46*, 521–533. doi:https://doi.org/10.1007/s10489-016-0841-8

Zangerle, E., & Bauer, C. (2022). *Evaluating Recommender Systems: Survey and Framework* (Vol. 55). ACM Computer Survey. doi:https://doi.org/10.1145/3556536

Zhang, S., Liu, K., Yu, Z., Feng, B., & Ou, Z. (2022). Hybrid recommendation system combining collaborative filtering and content-based recommendation with keyword extractions. *The 4th International Conference on Computing and Data Science*. doi:https://doi.org/10.54254/2755-2721/2/20220579

Zhang, Z.-K., Liu, C., Zhang, Y.-C., & Zhou. (2010). Solving The Cold-Start Problem In Recommender Systems. *92*(2), 28002. doi:https://doi-org.ezproxy.lib.torontomu.ca/10.1209/0295-5075/92/28002

Zhao, C., Sun, S., Han, L., & Peng, Q. (2016). Hybrid Matrix Factorization For Recommender Systems In Social Networks. *Neural Network World, 26*(6), 559-569. doi:10.14311/NNW.2016.26.032