**NATIONAL RESEARCH UNIVERSITY**

**HIGHER SCHOOL OF ECONOMICS**

**GRADUATE SCHOOL OF BUSINESS**

**PROJECT PROPOSAL**

# **GRAPH-SEARCH BASED RECOMMENDATION SYSTEMS: PROTOTYPE FOR A MOVIES STREAMING SERVICE**

Goreva A., Grosheva A., Listarov G.,

BIDP 221

**MOSCOW**

**2023**

**Abstract**

A large amount of constantly accumulating data can help companies set up more personalized services to increase user engagement in their products. Currently, a large amount of resources is spent searching for the most optimized and effective methods of using a large amount of user data to improve the accuracy of recommendation systems.

***The aim of this project*** is to propose a graph-based model for a recommendation system that deals with the data of users of streaming service “Netflix” of TV shows and movies for modeling users’ preferences and setting up content recommendations.

The goal is to create a prototype of a recommender system with help of a dataset from a popular streaming service. For this purpose, this project will look at the possibilities to create a recommender system - this includes a review of recent theoretical literature, a study of already existing classes of methods, such as Content-based filtering, Collaborative filtering, Hybrid filtering while evaluating the advantages and shortcomings of each method and offering a prototype of a graph-based model.

**Introduction**

With the increasingly rapid development and coverage of the Internet, modern companies accumulate a huge continuous amount of data from their users. While this may require decision makers to allocate huge resources to analyze this data and invest in the technology component, information about user actions may serve as a good opportunity to effectively improve their services and applications. Moreover, the modern user is less and less willing to put up with applications aimed at widespread mass use - on the contrary, he wants a personalized experience that will recognize his preferences and adjust the content in accordance with his tastes. Recommender systems can solve this problem by searching through lots of dynamically generated data and providing the user with needed content. Recommender systems can be defined as tools aimed at filtering data and finding patterns to offer relevant information for specific customers using different algorithms. While it may seem that such tools are only useful for the consumer, they are a powerful advantage for companies that increase their revenue due to greater user engagement and satisfaction.

This approach is especially relevant when it comes to streaming services, which are now filled with completely different types of content. It can be extremely difficult to navigate a huge number of genres and types of TV shows, movies and TV series, which are released literally every day. Thus, in this project, a model of a recommendation system for such a service will be proposed, based on the data of the “Netflix” (it was chosen as the most popular streaming service by quantity of unique users in the world [5]).

Firstly, to gain an understanding of how specifically recommender systems work, a review of the main recommendation techniques should be provided. Three main techniques are widely used by companies today, content-based filtering, collaborative filtering, knowledge-based systems. There also exists a combination of multiple techniques which is called "hybrid filtering".

While it can be observed that there is a sufficient number of methods that are used in the sphere of AI- and ML-powered recommendation systems and should be considered, recent advances in graph-based learning approaches have demonstrated their effectiveness in modeling preferences of a service user and items’ characteristics for various apps and services. So it can be an effective instrument for streaming services too.

In order to provide a relatively deep explanation of the graph-search based recommendation systems, its advantages and shortcomings, a literature review will be carried out in the next section, that will cover some of the background research that has been executed on the topic.

**Literature Review**

To achieve a systematic structure of existing research on graph-based recommendation, this project was performed with the use of a bibliographic review. One of the articles was published by the University of Amsterdam and is called “Graph Convolutional Matrix Completion' [1]'. The authors considered matrix completion for recommender systems from the point of view of link prediction on graphs. They proposed to represent interaction data such as movie ratings by a bipartite user-item graph with labeled edges denoting observed ratings. Then they proposed a graph auto-encoder framework based on differentiable message passing on the bipartite interaction graph. Their model shows competitive performance on standard collaborative filtering benchmarks. Another work “Graph Convolutional Neural Networks for Web-Scale Recommender Systems” was presented by Rex Ying and Ruining He [3]. They described a large-scale deep recommendation engine that they developed and deployed at Pinterest. They developed a data-efficient Graph Convolutional Network (GCN) algorithm PinSage, which combines efficient random walks and graph convolutions to generate embeddings of nodes that incorporate both graph structure as well as node feature information. According to offline metrics, user studies and A/B tests, PinSage generates higher-quality recommendations than comparable deep learning and graph-based alternatives. One more paper “Dynamic Graph Neural Networks for Sequential Recommendation” was written by Mengqi Zhang and Shu Wu [4]. They proposed a new method named Dynamic Graph Neural Network for Sequential Recommendation(DGSR), which connects different user sequences through a dynamic graph structure, exploring the interactive behavior of users and items with time and order information. Furthermore, they designed a Dynamic Graph Recommendation Network to extract user’s preferences from the dynamic graph. Consequently, the next-item prediction task in sequential recommendation is converted into a link prediction between the user node and the item node in a dynamic graph. Extensive experiments on three public benchmarks show that DGSR outperforms several state-of-the-art methods.

**Methods**

One of the main methods used to achieve the goal of the project was creating a prototype of a recommender system based on the dataset of a service from the targeted sphere. To this aim a structured database that can be searched with different queries should be designed, from which the data can be transformed into bipartite graph representation.

Some social network analysis methods could be used in building a recommendation system on graphs for a movie streaming service “Netflix”. The dataset, including more than 17k movies reviewed by 480K users, will be used in this project [2].

One of them are node centrality measures, which are used to identify the most important nodes in the network based on their degree of centrality. Some examples of centrality measures include degree centrality (the number of connections a node has), betweenness centrality (the degree to which a node lies on the shortest paths between other nodes), and eigenvector centrality (the influence of a node based on the centrality of its neighbors). Community detection algorithms will be implemented to identify groups of nodes that are densely connected within themselves but sparsely connected to other groups. Communities can provide insights into the structure and dynamics of the network, as well as opportunities for targeted marketing or product recommendations based on shared interests or preferences. Furthermore, the link prediction models, such as common neighbors, will predict the likelihood of a connection between two nodes that are not currently linked in the network. This can be useful for identifying potential new connections between users or items that could lead to increased engagement or views on a streaming service. Meanwhile, a k-shell decomposition will be used to identify influential users or items in the network, based on their ability to affect the behavior or opinions of other users. This can be useful for understanding the dynamics of the network.

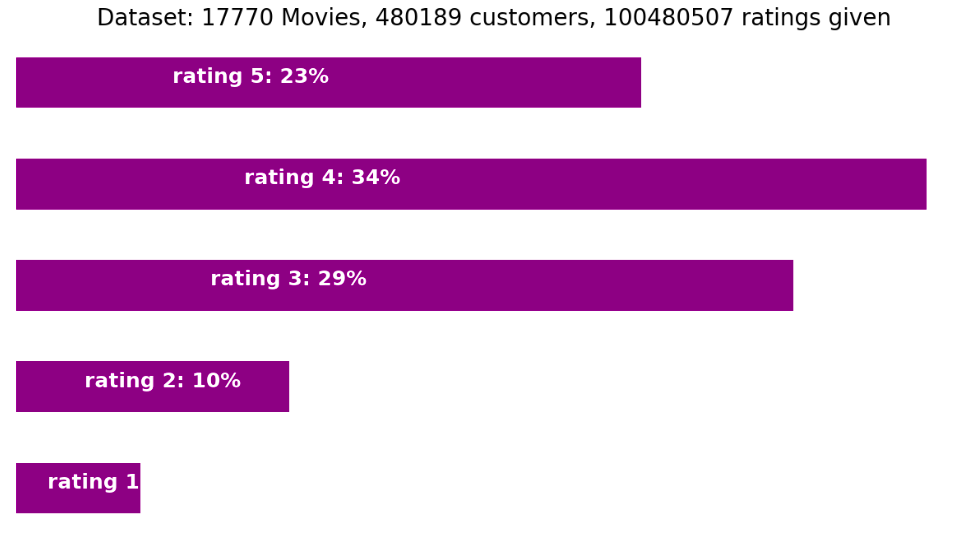
Overall, above-described methods, while it is worth noting that perhaps not all of them will be used in the work, might help to identify the most important nodes, communities, and connections in the network, as well as predict future behavior and influence for a movie streaming service recommendation system.

**Expected Result**

The expected result of the project lies in the development of a recommender system based on users’ preferences of movie genres with the help of a graph database structure. By structuring the initial database in graphs, the streaming service can use the latest technology for recommender systems.

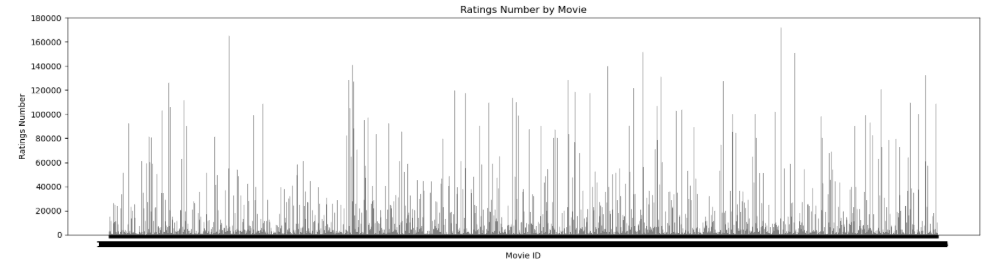
The current project is partially grounded on the existing research in the field, but also brings its own expertise that can serve as a foundation of future research on the optimization and effectiveness of recommendation systems. For example, other techniques that could be observed in finding recommendation patterns are systems based on artificial intelligence and machine learning (ML).

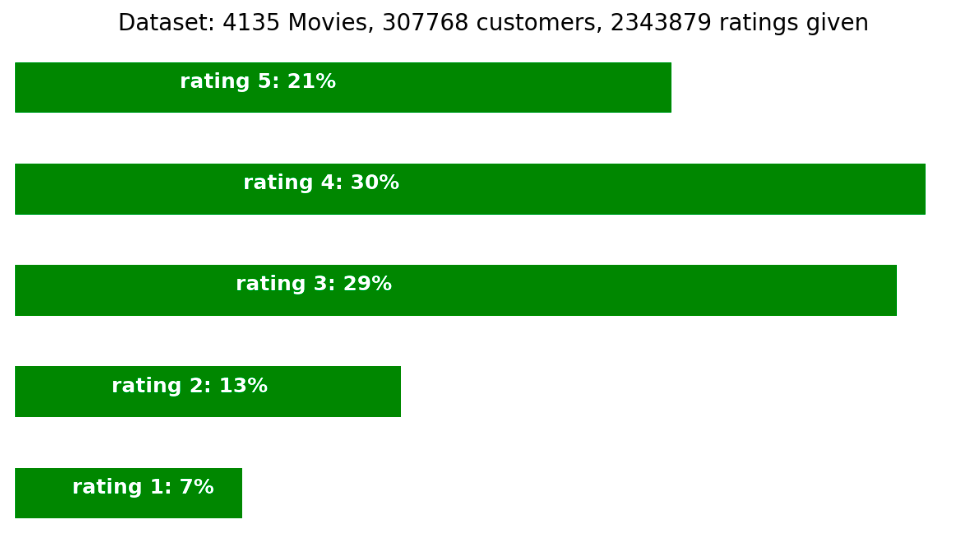
**Technical Report**

First, several datasets with films and their ratings from the Netflix streaming service were taken and combined. The data obtained were visualized and analyzed using a graph.

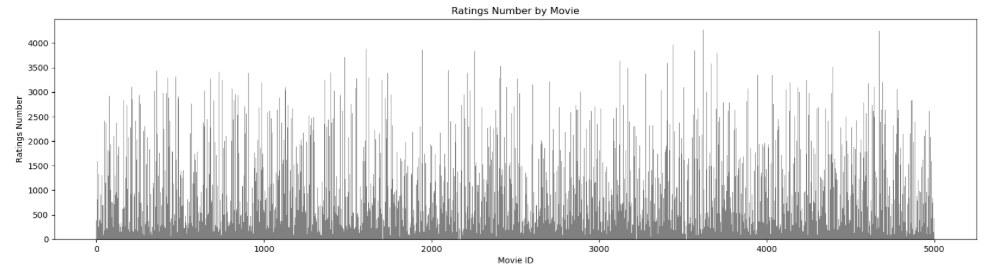
Thus, we learned that the data in the dataset does not have a normal distribution: there are significant outliers in it. For example, the largest share of ratings is 3-4 points. Probably, users do not want to put negative ratings (1-2) on a disliked movie and instead simply stop watching them.

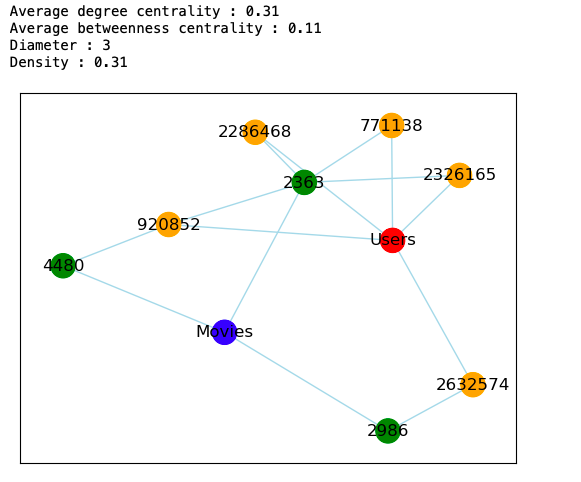
Also, before further analysis of the data, the dataset was examined for how many films contain a small number of ratings. Based on the graph below, it is clear that the distribution is also not normal: there are extremely many films with a small number of ratings in the dataset. Such objects cannot be taken for further analysis. That is why data filtering was carried out at the next stage of the work.



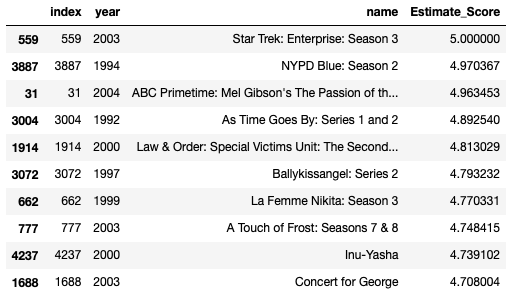
To prepare the dataset for further analysis, all films with less than 1000 ratings were excluded from the sample (since it will not be possible to build a high-quality recommendation system based on them), as well as films with more than 5000 ratings (they may become outliers, which will also negatively affect the quality of the recommendation system). Also, users with less than 200 ratings were excluded from the dataset, as well as films with an ID of less than 5000: thanks to such a solution, we get rid of inactive users whose preferences are not clear due to the limited number of data.

Thus, by filtering the data, we get a better sample without a critical amount of outliers. based on the graph below, it can be seen that the sample includes only films with a sufficient number of ratings.



Then graphs were built on the filtered data. Since the sample for analysis turned out to be extremely voluminous, graphs were built on 5, and then on 10 films (the data obtained were extrapolated to the entire sample). With the help of graphs, it was possible to understand how the analyzed data are interconnected: thus, we see that disparate users have an intersection point in the form of the same film, which was rated. Presumably, such films that have been rated by a large number of users are in the top of the rating and are extremely popular.

Next, a recommendation system was developed, which is based on the ratings that the user gave to various films. First, we created a training sample and trained our recommendation model. More detailed code can be found in the repository. Below is a piece of a table with a recommendation system: based on the user's preferences, we show him those films that will satisfy his interests (depending on the ratings that he gave the films earlier).



**References**

1. Berg R., Kipf T. N., Welling M. Graph convolutional matrix completion //arXiv preprint arXiv:1706.02263. – 2017.
2. Netflix Prize Data // Kaggle - 2020. - Retrieved from: <https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>
3. Ying R. et al. Graph convolutional neural networks for web-scale recommender systems //Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. – 2018. – С. 974-983.
4. Zhang M. et al. Dynamic graph neural networks for sequential recommendation //IEEE Transactions on Knowledge and Data Engineering. – 2022.
5. Онлайн-видео (мировой рынок) // Tadviser - 2023. - Retrieved from: <https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%9E%D0%BD%D0%BB%D0%B0%D0%B9%D0%BD-%D0%B2%D0%B8%D0%B4%D0%B5%D0%BE_(%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%BE%D0%B9_%D1%80%D1%8B%D0%BD%D0%BE%D0%BA)?ysclid=lh5981cj6q218806500>