A Paper Review on ”The Netflix Recommender System: Algorithms, Business Value, and Innovation”[1]

Anthony Geglio*, Md Nehal Salik*, Alexander Putman*
College of Computing,
Michigan Technological University
Houghton, MI 49931, USA

ajgeglio@mtu.edu, msalik@mtu.edu, amputman@mtu.edu
* Denotes equal contribution

***Abstract*—Recommender system is a subclass of information retrieval system and is a technique widely used by the e-commerce, streaming service and other companies to suggest goods and services to their users. Mathematically its a function that predicts how a user will like an item. In this paper the authors have explained how Netflix Recommendation system drives their main business value and prepositions. Netflix core product is their recommendation system. Authors have explained different algorithms which makes the Netflix Recommendation System and how they influence their subscription model, user retention and attrition. Authors have also explained steps taken to improve the algorithms, A/B testing and future open problems. They have also explained how they want to build a global algorithms for Netflix that could be universal to all the language domains.**

***Index Terms*—Recommender systems**

I. Introduction

Netflix is a media service provider which is born in the USA and is entertaining the world. We may think of Netflix as an entertainment company but, it is a pure technology company. Research and Innovation are the core values of how Netflix operates. It makes use of the Technology to maximize its revenue and entertain its user in the best possible way. Among all the subscription-based entertainment providers, Netflix has the least attrition rate and as per the latest estimate, it has more than 165 million worldwide subscribers and controls more than 55 percent of the world's subscription-based market share. Netflix is a company that considers itself at the crossroads of the internet and storytelling. This article discusses the various algorithms that make the Netflix Recommendation System. What all business value each algorithm derives. What all steps are taken to improve the algorithms and key open problems and challenges. To sum up, the entire homepage of Netflix can be compared to a 100 by 75 matrix(roughly), where each element is a recommendation derived from a certain Algorithm. Each row consists of a recommendation with the best one on the left and the least recommended one on the right end. Also, the recommendations go top to down, with the highest recommended one at the top and the least recommended one at the bottom.

II. Background and Motivation

The authors introduce the paper with the theme of storytelling. The emphasis is that storytelling is intrinsic to human nature, and that technology such as language, writing, and the printing press have historically been harnessed to as a tool to more widely distribute stories. Now, applications like Netflix lie at the intersection of storytelling and the way the internet is utilized for such. The authors also emphasise that the recommender system and the innovation is the pillar of the Netflix tool.

The motivation of this paper is to introduce the algorithms that make up the recommender, it's business value, the process to improve the algorithms, and some open problems that remain.

The Netflix Recommender system is composed of a variety of algorithms, the output of which is presented to the user on the Netflix Homepage. The output is presented in a matrix layout where the rows represent different themes, and each entry of a given row is the recommendations in that theme. A typical homepage has around 40 rows, and approximately 75 video recommendations per row.

The videos in a given row typically come from a single algorithm. For example, the genre rows are driven by the personalized video ranker (PVR).

III. Main Contribution

The paper was written in 2015 by Carlos A. Gomez-Uribe and Neil Hunt both executive level machine learning specialists at Netflix. At the time

Gomez-Uribe was the Vice President of Product Innovation and Penalization Algorithms, and Niel Hunt was a Chief Product Officer.

The contribution of the paper is to bring a high level understanding of the Netflix Recommender system to the public and academic domain.

IV. Recommender algorithms

The recommender system relies on the combination of many machine learning techniques. Some examples which are implemented in the NRS are dimensionality reduction through clustering or compression, topic modeling, supervised and unsupervised classifi cation, regression, and matrix factorization. While the authors do not explicitly describe the detail of each of their algorithms, they do include a list of papers that have adaptations of topic models specific to the recommender systems domain. One which is particularly around matrix factorization[2].

Netflix Recommender system consists of the below mentioned algorithms.

- Personalized Video Ranker (PVR)
- Top n video ranker (TVR)
- Trending Now(TN)
- Continue Watching(CW)
- Video-video similarity(VVS)
- Page Generation: Row selection and Ranking(PG)
- Search

PVR orders the entire catalog of videos for each member in a personalised way. PVR is used widely and is good at general purpose relative ranking throughout the catalog.Basically, for each row it creates unique rank of recommendation for the unique users. TVR combines the personalization with popularity and produces ranking in the top picks section. TVR looks at the head of the each catalog ranking and extract the head element to create the top picks section.

Figure 1: Top Ranks generated through TVR Algorithm

The trending Now algorithm generates the trending now ranker on the home screen. It generates recommendations based on the shorter-term temporal trends. Trending now generates two types of trends nicely. They are the one those repeats every several months, for example recommending romantic movies during valentine's day week. Another one is those based on one-off short-term events, for example, recommending war movies during the current Ukrainian war.

Figure 2: Trending Now row generated through the TN Algorithm

Continue Watching ranker is based on the users recent views and displays the contents in the continue watching section. Based on what user has watched previously and left it somewhere, it ranks the best possible elements out of those to the users in the continue watching section.

Figure 3: Continue Watching row generated through the CW Algorithm

Video-Video similarity also called as the SIMS algorithm in the paper is content based and is the only depersonalized algorithm in the Netflix Recommendation System. It creates a best possible image of every video the user have watched, rank them and show the contents in the because you watched section.

Page Generation algorithms is a sum up of all the algorithms and based on all the inputs received from the output of all the above algorithms it generates a personalised recommendation page for each unique user. All the algorithms mentioned till now makes the Netflix Recommendation System(NRS). Eighty percent of the stream which goes through the Netflix

Figure 4: Because You Watched row generated through the SIMS Algorithm

is through their NRS and rest of the twenty percent is through their search algorithm.

V. Success evaluation criteria

Another main contribution of the paper is the discussion around the difficult problem of evaluation. Evaluating the success of the recommender system algorithms is not straightforward and does not come with labels like a supervised learning problem (although some supervised techniques are implemented in offline testing). Rather, improving algorithms and observing their performance is full of nuance and is not based on the intuition of the developers or the collective. For example, intuitively a good business implements what customers ask for, which in this case is as much choice as possible and comprehensive search tools to access and navigate those choices. What actually works in practice is a few compelling choices simply presented.

The Primary performance metric that Netflix uses is the retention rate of the members.

Ultimately, the success of Netflix is driven by the number of subscribers, and Netflix is continuously updating the algorithms to add value to the product. The success or failure of the changes is based on the Δ retention rate of the members in reaction to those changes. The basic hypothesis is that by creating better personalized recommendations, those who are on the fence will stay longer. One consideration is that retention rate is already high and therefore it takes a very meaningful change in the algorithms to make a difference.

A secondary performance metric is the engagement time. Engagement time is strongly correlated with improving retention and therefore it can be used to validate the statistical significance between the changes and the Δ retention.

The official way in which Netflix implements and observes the impact of changes is through what is called ”A/B testing”. A/B testing is the live testing of their new algorithms on the users, and is continuously carried out by Netflix. A typical test lasts 2-6 months before evaluation. Members are randomly assigned to different experience called cells, in which there are ”test cells” which implement the new changes and ”control cells” which contain the current production algorithm. The cell experience is consistent for the respective users during this 2-6 month time.

During evaluation, the data can indicate several things about the changes. After evaluation, they are able to observe whether the test cell users watch the similars more, whether test cell engagement time has increased with the actual change (local engagement), and/or does the change increase overall engagement time with the Netflix product, and finally, are the Δ retention rates higher in the test cell relative to the control cell.

In order to validate whether the changes observed are actually due to the algorithm changes implemented by Netflix (rather than random variations), the authors introduce a somewhat novel use of a probability model for another layer of validation. This model is used to predict how much they expect the metric to vary if the test was performed repeatedly. The smaller the percentage of hypothetical experiments that yield negative results, the higher confidence that the test cell increased retention. The probability model is also used to determine the sample size needed to measure an increase of a given magnitude.

The last novel approach discussed by this paper in terms of improving algorithms is the use of offline testing. Offline testing uses historical data on user choices. This database allows them to test sometimes 100 algorithm variants with slightly different parameters against historical user choices. This process can be implemented quickly and prune the candidate variants that actually end up in the A/B testing. One drawback is the unrealistic model that a user's historical choices are repeated, and therefore when very different recommendations are generated they behave poorly in the tests.

VI. Open Problems

The Netflix Recommender system, while impressive for its size, does contain problems that are not adequately addressed. Certain bias is introduced using a recommender; for Netflix, a positive feedback loop exists based on recommended content. When a unit of media has a high level of engagement, the recommender in turn reciprocates that media to more users, who then have engagement. This leads to a loop of extreme engagement of media that is not directly accounted for in the algorithm. In these cases, popularity is overwriting personalization in the recommender.

Because personalization is so important to the Netflix model, new members are often ostracized from the full experience offered by Netflix. The recommender improves as it has more training information, and without adequate training the model does not perform to its full potential. Members who recently joined or are using a free trail have not had the time to develop their preferences within the recommender, and therefor are not essentially offered the same product as those who have been using it for years. This problem is not unique to Netflix, but rather to most media recommender models.

More uniquely to Netflix is the personalization across multiple people within a single account. The recommender currently does not have a solid method of distinguishing between multiple users across a single Netflix subscription. To attempt to combat this, Netflix implemented a profile system upon launch of the application. The app prompts “Who’s watching?”, where different members of the account can select their individual system. Even still, Netflix estimates that most of the users in a household share a single named profile within one subscription. Other open problems come from the essence of A/B testing as opposed specifically to the recommender system. Interpretation of A/B testing, while supported by the resulting data developed by Netflix, remains somewhat of an art rather than a science. The system has recognized multiple trials where changes in A/B testing provided better or worse results that are not reproducible. There is a certain amount of random variation that is assumed when working within A/B testing.

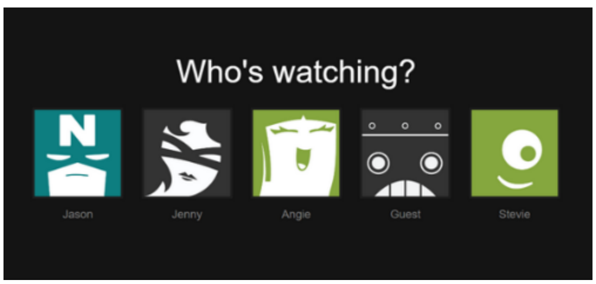

Figure 5: Who’s Watching prompt upon launching the Netflix application

VII. Future Work

Continued improvement of the Netflix Recommender model is the path for business success. Netflix aims broadly to keep memberships alive; that is, the main goal is not aimed at creating new membership, but rather increasing membership retention. It is difficult to measure why membership subscriptions are deactivated, but many of them are estimated to originate from payment failures rather than deliberate cancellations. Many of these previous members may reactivate their accounts later. Creating a possibility to measure these rates of membership can improve the performance metrics used to evaluate the model.

Addressing the key open issues with the recommender system will be critical to the continuous improvement of the Netflix model. Account sharing research is underway to understand how to credit recommendations to different users under a single account. Moreover, understanding the differences in recommendation software amongst different age groups, particularly children, poses an even greater problem that is left for more research. Understanding how to collect more data on new users and make the existing recommender system work for them is another potential future work; Netflix has already implemented an introductory survey to capture the perceived interests of a new subscriber based on their own judgement. Capitalizing on this information and developing new ways to work in newer members will be critical to new or returning engagement.

An overarching conclusion to their future work will be to choose the correct metrics for evaluation of their models. When basing the model development in A/B testing, understanding the results is critical to making improvements over the testing information. The concentration of a global company is another strong consideration, as the same model must be adapted to work without implicit cultural bias to one country or region versus another. Understanding the differences in metrics and how they apply across their offering is important to enhancing the model. The future of Netflix as a product is heavily dependent on the recommender system that Netflix is built upon. Facing these challenges and problems will tell the level of success recognized by the product.

References

[1] C. A. Gomez-Uribe and N. Hunt, “The netflix recommender system: Algorithms, business value, and innovation,” *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, dec 2016. [Online]. Available: https://doi.org/10.1145/2843948

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.