Netflix - big data

Mario Kjurchievski

What is big data?

I would like to start my paper with defining the term big data. Big data is a term that describes the large volume of data, both structured and unstructured. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves, but also we are witnessing big number of data leaks and wrongfully using, which has damaged a lot of people.

Big data is defined with five V's:

- **Volume** - knowing that the name contains the word *big*, we can assume that the volume is one of the five V's.
- **Velocity** - in addition to managing data, companies need that information to flow quickly, as close to real-time as possible.
- **Variety** - a company can obtain data from many different sources.
- **Veracity** - this context is equivalent to quality.
- **Value** - this refers to the ability to transform all the data into a business.

Now that we learned how big data is defined, I can move on to my assignment.

Netflix recommendations: Beyond the 5 stars (part 1 and part 2)

I will try to sum up all the sources I had to read, regarding Netflix and their big data. I will write for each source individually. The first two sources have the same topic and are as one divided in two parts. The links are:

- https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429
- https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-2-d9b96aa399f5

The Netflix Prize and the beginning of Netflix streaming platform

In these topics, the authors start their story by telling us about the Netflix Prize. So, what is that? The Netflix Prize was a contest which was started in 2006. The previous years Netflix was a company which rented DVDs. When they decided to launch a streaming platform, they still haven't developed the most efficient recommendation algorithms, which would help their users get recommendations on movies and tv shows which would potentially be in their likings.

Up until this point, they had about 8 to 9 years' worth of data, which they collected in their DVD rental places. Through the years they documented all the ratings given by the users, the date which the rating was submitted, unique ID numbers of the subscribers and some other information regarding to their users.

When they announced the Netflix Prize, they gave the competitors two big datasets which included 100 million subscribers. The goal was to improve the already existing algorithms by 10%. The prize was big, 1 million dollars for the most successful one. With this big of a prize, there was a lot of competition. The winning team reportedly worked 2000 hours and used a combination of 107 algorithms to achieve an improvement of 8.43%. Using that recommendation system, Netflix launched their streaming platform in 2007.

Everything is a recommendation

In order to keep their users happy, Netflix has to personalize recommendations, as much as possible. They tell us how they develop everything around that thought. Starting from the physical layout of their page and application, to their movies and tv shows selections, everything is a recommendation. They mention a few elements for which they tell us why they are important. Some of those elements are:

- **Household** - It is important to keep in mind that Netflix' personalization is intended to handle a household that is likely to have different people with different tastes.
- **Diversity** - to appeal to the ranges of interests and moods, beside accuracy, they need diversity.
- **Awareness** and **explanation** - they want their subscribers to be aware of how they are adapting to their tastes, allowing the user to gain trust but also, encourage will to give feedback that will result in better recommendations.
- **Social** - by connecting Netflix to your Facebook account, the members are getting recommendations based on the likings of their friends.
- **Similarity** - a very important source of personalization, coming in many forms.
- **Ranking** - one of the most crucial elements, which allows Netflix to order their content to every individual.

Data and models

The data sources can be divided in two groups, internal and external sources. In the internal group, Netflix has a lot of information for every user, here are some examples:

- Several billion ratings from members.
- Location of the users.
- Directly entered search terms.
- Rich metadata, including favorite actors, directors and genre.
- They observe the user's interaction with the recommended content, including scrolls, mouse-overs, clicks or the time spent on a given page.

These are just a few chunks of internal data. For the external sources they take in consideration the box office performance or critics reviews.

At Netflix, they use many different modeling approaches for building personalization engines. Some of the methods they use are: Linear regression, Logistic regression, Elastic nets, Singular Value Decomposition, Restricted Boltzmann Machines, Markov Chains, Latent Dirichlet Allocation, Association Rules, Matrix factorization, Gradient Boosted Decision Trees, Random Forests, and Clustering techniques from the simple k-means to graphical approaches such as Affinity Propagation.

More data usually beats better algorithm

Next up I will talk about a professor at Stanford, who challenged his students to participate in the Netflix Prize contest (link to source). For easier understanding of his point, he takes two groups of students for comparison. The first group, let's call it Team A, came up with a very complex algorithm using the Netflix data. The second group, Team B, used a very simple algorithm, but besides Netflix's data, they used an additional information about movie genres (source IMDB).

The results came rather surprising, Team B got a bigger improvement of the Netflix existing recommendation system. As a conclusion the professor suggests that if you have limited sources, add more data rather than focusing on a complex algorithm. But of course, we must be aware of what data are we adding to, if it's not right, we will probably get worse results ("aha" moment).

How to use data to make a hit TV show

This source is a TED Talk. The speaker is a data scientist called Sebastian Wernicke. He makes a comparison between an Amazon Prime and a Netflix TV show which are about a same thing (both directed thanks to data analysis) but the one on Netflix has a much bigger rating. What he wants to tell us with this, is that we can't always rely on data itself ("aha" moment).

He makes one very good point which refers to the previous sentence that I wrote. When you are solving a complex problem, you take apart every piece of that problem, you analyze them and then put them in the correct form to finally solve this problem. What he says is that data and data analysis are only good for taking a problem apart and understanding all the pieces, but it's not best at putting those pieces back together and solving the problem. For that second part, we should use another powerful tool, which is our brain ("aha" moment). So even though Netflix and Amazon Prime had similar information gathered from the data, Amazon Prime didn't make the right decision in using that data for creating the final product.

Netflix privacy lawsuit

This final source is a bit different than all the previous. Here I read about an in-the-closet homosexual mother, who was "exposed" because of Netflix data. In this paper, we can see that data use is not always ethical. There are many victims to data leaks and unlawful use of people's private information ("aha" moment).

As I mentioned previously, for the Netflix Prize, the competitors were given two bid datasets which included 100 million movie ratings, along with the date of the rating, a unique ID number for the subscriber, and the movie info. Within this dataset, was information for that mother. The thing is that Netflix released this information as anonymous, but just 2 to 3 weeks after the competition has started, two University of Texas researchers, identified several Netflix users. They did that by comparing subscriber's reviews in the Netflix data, to reviews posted on the IMDB website. This exposed those people's political leanings and sexual orientation.