3
NETFLIX

How Netflix Used Big Data To Give Us The Programmes We Want

Background

The streaming movie and TV service Netflix are said to account for one-third of peak-time Internet traffic in the US, and the service now have 65 million members in over 50 countries enjoying more than 100 million hours of TV shows and movies a day. Data from these millions of subscribers is collected and monitored in an attempt to understand our viewing habits. But Netflix's data isn't just "big" in the literal sense. It is the combination of this data with cutting-edge analytical techniques that makes Netflix a true Big Data company.

What Problem Is Big Data Helping To Solve?

Legendary Hollywood screenwriter William Goldman said: "Nobody, nobody – not now, not ever – knows the least goddam thing about what is or isn't going to work at the box office."

He was speaking before the arrival of the Internet and Big Data and, since then, Netflix have been determined to prove him wrong by building a business around predicting exactly what we'll enjoy watching.

How Is Big Data Used In Practice?

A quick glance at Netflix's jobs page is enough to give you an idea of how seriously data and analytics are taken. Specialists are recruited to join teams specifically skilled in applying analytical skills to particular business areas: personalization analytics, messaging analytics, content delivery analytics, device analytics … the list goes on. However, although Big Data is used across every aspect of the Netflix business, their holy grail has always been to predict what customers will enjoy watching. Big Data analytics is the fuel that fires the "recommendation engines" designed to serve this purpose.

Efforts here began back in 2006, when the company were still primarily a DVD-mailing business (streaming began a year later). They launched the Netflix Prize, offering $1 million to the group that could come up with the best algorithm for predicting how their customers would rate a movie based on their previous ratings. The winning entry was finally announced in 2009 and, although the algorithms are constantly revised and added to, the principles are still a key element of the recommendation engine.

At first, analysts were limited by the lack of information they had on their customers – only four data points (customer ID, movie ID, rating and the date the movie was watched) were available for analysis. As soon as streaming became the primary delivery method, many new data points on their customers became accessible. This new data enabled Netflix to build models to predict the perfect storm situation of customers consistently being served with movies they would enjoy. Happy customers, after all, are far more likely to continue their subscriptions.

Another central element to Netflix's attempt to give us films we will enjoy is tagging. The company pay people to watch movies and then tag them with elements the movies contain. They will then suggest you watch other productions that were tagged similarly to those you enjoyed. This is where the sometimes unusual (and slightly robotic-sounding) "suggestions" come from: "In the mood for wacky teen comedy featuring a strong female lead?" It's also the reason the service will sometimes (in fact, in my experience, often!) recommend I watch films that have been rated with only one or two stars. This may seem counterintuitive to their objective of showing me films I will enjoy. But what has happened is that the weighting of these ratings has been outweighed by the prediction that the content of the movie will appeal. In fact, Netflix have effectively defined nearly 80,000 new "micro-genres" of movie based on our viewing habits!

More recently, Netflix have moved towards positioning themselves as a content creator, not just a distribution method for movie studios and other networks. Their strategy here has also been firmly driven by their data – which showed that their subscribers had a voracious appetite for content directed by David Fincher and starring Kevin Spacey. After outbidding networks including HBO and ABC for the rights to *House of Cards*, they were so confident it fitted their predictive model for the "perfect TV show" that they bucked the convention of producing a pilot and immediately commissioned two seasons comprising 26 episodes. Every aspect of the production under the control of Netflix was informed by data – even the range of colours used on the cover image for the series was selected to draw viewers in.

The ultimate metric Netflix hope to improve is the number of hours customers spend using their service. You don't really need statistics to tell you that viewers who don't spend much time using the service are likely to feel they aren't getting value for money from their subscriptions, and so may cancel their subscriptions. To this end, the way various factors affect the "quality of experience" is closely monitored and models are built to explore how this affects user behaviour. By collecting end-user data on how the physical location of the content affects the viewer's experience, calculations about the placement of data can be made to ensure there is an optimal service to as many homes as possible.

What Were The Results?

Netflix's letter to shareholders in April 2015 shows their Big Data strategy was paying off. They added 4.9 million new subscribers in Q1 2015, compared to four million in the same period in 2014. Netflix put much of this success down to their "ever-improving content", including *House of Cards* and *Orange is the New Black*. This original content is driving new member acquisition and customer retention. In fact, 90% of Netflix members have engaged with this original content. Obviously, their ability to predict what viewers will enjoy is a large part of this success.

And what about their ultimate metric: how many hours customers spend using the service? Well, in Q1 2015 alone, Netflix members streamed 10 billion hours of content. If Netflix's Big Data strategy continues to evolve, that number is set to increase.

What Data Was Used?

The recommendation algorithms and content decisions are fed by data on what titles customers watch, what time of day movies are watched, time spent selecting movies, how often playback is stopped (either by the user or owing to network limitations) and ratings given. In order to analyse quality of experience, Netflix collect data on delays caused by buffering (rebuffer rate) and bitrate (which affects the picture quality), as well as customer location.

What Are The Technical Details?

Although their vast catalogue of movies and TV shows is hosted in the cloud on Amazon Web Services (AWS), it is also mirrored around the world by ISPs and other hosts. As well as improving user experience by reducing lag when streaming content around the globe, this reduces costs for the ISPs – saving them from the cost of downloading the data from the Netflix server before passing it on to the viewers at home.

In 2013, the size of their catalogue was said to exceed three petabytes. This humungous amount of data is accounted for by the need to hold many of their titles in up to 120 different video formats, owing to the number of different devices offering Netflix playback.

Originally, their systems used Oracle databases, but they switched to NoSQL and Cassandra to allow more complex, Big Data-driven analysis of unstructured data.

Speaking at the Strata + Hadoop World conference, Kurt Brown, who leads the Data Platform team at Netflix, explained how Netflix's data platform is constantly evolving. The Netflix data infrastructure includes Big Data technologies like Hadoop, Hive and Pig plus traditional business intelligence tools like Teradata and MicroStrategy. It also includes Netflix's own open-source applications and services Lipstick and Genie. And, like all of Netflix's core infrastructure, it all runs in the AWS cloud. Going forward, Netflix are exploring Spark for streaming, machine learning and analytic use cases, and they're continuing to develop new additions for their own open-source suite.

Any Challenges That Had To Be Overcome?

Although a lot of the metadata collected by Netflix – which actors a viewer likes to watch and what time of day they watch films or TV – is simple, easily quantified structured data, Netflix realized early on that a lot of valuable data is also stored in the messy, unstructured content of video and audio.

To make this data available for computer analysis and therefore unlock its value, it had to be quantified in some way. Netflix did this by paying teams of viewers, numbering in their thousands, to sit through hours of content, meticulously tagging elements they found in them.

After reading a 32-page handbook, these paid viewers marked up themes, issues and motifs that took place on screen, such as a hero experiencing a religious epiphany or a strong female character making a tough moral choice. From this data, Netflix have identified nearly 80,000 “micro-genres” such as “comedy films featuring talking animals” or “historical dramas with gay or lesbian themes”. Netflix can now identify what films you like watching far more accurately than simply seeing that you like horror films or spy films, and can use this to predict what you will want to watch. This gives the unstructured, messy data the outline of a structure that can be assessed quantitatively – one of the fundamental principles of Big Data.

Today, Netflix are said to have begun automating this process, by creating routines that can take a snapshot of the content in Jpeg format and analyse what is happening on screen using sophisticated technologies such as facial recognition and colour analysis. These snapshots can be taken either at scheduled intervals or when a user takes a particular action such as pausing or stopping playback. For example, if it knows a user fits the profile of tending to switch off after watching gory or sexual scenes, it can suggest more sedate alternatives next time they sit down to watch something.

What Are The Key Learning Points And Takeaways?

Predicting what viewers will want to watch next is big business for networks, distributors and producers (all roles that Netflix now fill in the media industry). Netflix have taken the lead but competing services such as Hulu and Amazon Instant Box Office and, soon, Apple, can also be counted on to be improving and refining their own analytics. Predictive content programing is a field in which we can expect to see continued innovation, driven by fierce competition, as time goes on.

Netflix have begun to build the foundations of “personalized TV”, where individual viewers will have their own schedule of entertainment to consume, based on analysis of their preferences. This idea has been talked about for a long time by TV networks but now we are beginning to see it become a reality in the age of Big Data.

REFERENCES AND FURTHER READING

For more on Netflix's Big Data adventure, check out:

http://techblog.netflix.com/

http://www.netflixprize.com/http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

http://www.wired.com/insights/2014/03/big-data-lessons-netflix/

http://files.shareholder.com/downloads/NFLX/47469957x0x821407/DB785B50-90FE-44DA-9F5B-37DBF0DCD0E1/Q1_15_Earnings_Letter_final_tables.pdf