Recommendation systems, from fancy ideas to real world practice

Tiit Sepp 01.04
Bio

- M.A. Theoretical physics
- PhD studies at Tartu Observatory
- Data Scientist @ STACC since 2018
  - Recommender systems
  - Data analytics
  - Custom projects
- Random personal quote: “Running a marathon does not require training”
Simple intro...

Estimate business case:

\[
\begin{align*}
k_1 &= h f(t_j, y_j) \\
k_2 &= h f\left(t_j + \frac{1}{4} h, y_j + \frac{1}{4} k_1\right) \\
k_3 &= h f\left(t_j + \frac{3}{8} h, y_j + \frac{3}{32} k_1 + \frac{9}{32} k_1\right) \\
k_4 &= h f\left(t_j + \frac{12}{13} h, y_j + \frac{1932}{2197} k_1 - \frac{7200}{2197} k_2 + \frac{7296}{2197} k_3\right) \\
k_5 &= h f\left(t_j + h, y_j + \frac{439}{216} k_1 - 8 k_2 + \frac{3680}{513} k_3 - \frac{845}{4104} k_4\right) \\
k_6 &= h f\left(t_j + \frac{1}{2} h, y_j - \frac{8}{27} k_1 + 2 k_2 - \frac{3544}{2565} k_3 + \frac{1859}{4104} k_4 - \frac{11}{40} k_5\right)
\end{align*}
\]
Dream of a perfect recommender system...

- Ideally a perfect recommender system would:
  - For customer - Add automatically all the desired products to the cart with minimal effort
  - For retailer - Maximize long and short term profit

Copyright: memegenerator.net, Dreamworks
From idea to Recsys

What is the business goal?

Independent
- Increase alternatives
- Decrease search time

Support to physical store
- Sell outlet products
- Increase physical store revenue
How one designs a recsys
How the process works

- Collect the ideas
- Collect the data
- Build the model
- Optimize the model
- Add business rules
- Turn on the recsys
- Everyone collects the money
Collecting ideas

GOAL: find the optimal solution between what is wanted, needed and possible
Collecting data

This is where the fun starts...
And it just might be that you spend up to 90% of your project time dealing with issues related to data
Build the model

● No events or low event count + not all events are equal
  ○ Feature based models
    ■ Word features
    ■ Image features
    ■ Categorial features

● Abundance of events
  ○ Event based models
    ■ Collaborative filtering
    ■ Session modeling
    ■ Intent detection modelling
  ○ Hybrid models

Credit: gtihub, 20th century fox
Few words on events
BOT TRAFFIC REPORT 2016

Bots once again comprise the majority of online traffic amid an increase in good bot activity.

Increase in good bot activity, which went up by 44 percent.

Bot activity is in an uptrend, after a three-year decline.

1.2% MONITORING BOTS
Health checkers that monitor website availability and the proper functioning of various online features.

2.9% COMMERCIAL CRAWLERS
Spiders used for authorized data extractions, usually on behalf of digital marketing tools.

6.6% SEARCH ENGINE BOTS
Bots that collect information for search engine algorithms, which they use to make ranking decisions.

12.2% FEED FETCHERS
Bots that fetch website content to mobile and web applications, which they then display to their users.

24.3% IMPERSONATORS
Bots that assume false identities to bypass security solutions. They are commonly used for Good assays.

1.7% SCUMPERS
Bots used for unauthorized data extraction and the reverse engineering of pricing models.

0.3% SPAMMERS
Poluters that inject spam links into discussions and comment sections.

2.5% HACKER TOOLS
Scammers that look for systems with vulnerabilities to exploit for data theft, malware injection, etc.

Credit:
Impterva Incapsula
What are the business goals..
What can models easily do...
What could possibly go wrong...

• You get the data and train content-based filtering model.
• If the data doesn’t include enough features to calculate the product similarity, the results will be bad!
What could possibly go wrong...

...the client knows you know too much
What could possibly go wrong...

- You get the data and train your collaborative filtering model.
- If people doesn’t buy often from the store, you’ll get weird results. It is because the patterns in the data are not statistically important!
What could possibly go wrong...

- If you find patterns based on historical data, you will be making wrong recommendations if person’s interests have changed!
What could possibly go wrong..

- You build your reinforcement learning based system.
- But it takes 7s to re-train and display the recommendations.
- The complex systems are making good recommendations but might be too slow to be used in real-time environment!
Add business rules

- The recommendation system that is trained on USA retailers' data could not be used in India!
- The behavior of one client-base is not the same as other client-base (geography, socio-cultural aspects, etc.)!
Business rules

- Let’s assume that person has only ever bought alcohol from our store and recommender thinks that this is the only thing person is interested in!
- Recommending only alcohol might be against company policy!? 
- Every business has specific rules and principles – the AI system must be aligned accordingly!
Now we finally turn on the recsys

But before that we evaluate the system.

As bare minimum...

- Model must outperform popularity
- Offline metrics must be reasonable
Once online...

Credit: New Line Cinema
Results validation
Result validation
SORRY I ANNOYED YOU

WITH MY LIFE SAVING RECOMMENDATION

Credit: Concordia university Texas