Article Recommendation System

By Team 404

Building a simple, yet sophisticated Article Recommender System

The User

The User

Analysing the user of the article recommender system.

- Definite Taste
- Taste not very volatile
- Few actions per day
- Can be categorised into different types

Model Requirements

Which model would best suite our user?

- A model capable of grouping similar types of items on the basis of their content.
- A model which can group people on the basis of their affinities.
- A model which is steadfast
- A model which works well with very few actions, however takes these actions into account before recommending a new set of items

The Data

- From CI&T's Internal Communication platform (DeskDrop).
- Contains a real sample of 12 months logs (Mar. 2016 - Feb. 2017)
- Contains about 73k logged users interactions on more than 3k public articles shared in the platform.
- The training set contains the first (as per timestamp) 80% of the data.
- The testing data contains the remaining 20%.

The Model

The Model

A bird eye view of the model.

- Different algorithms for new and old users.
- New users - clustering and most popular
- Regular users - Hybrid of interactive content based filtering and collaborative filtering with some inspiration from Reinforcement Learning.

Ratings

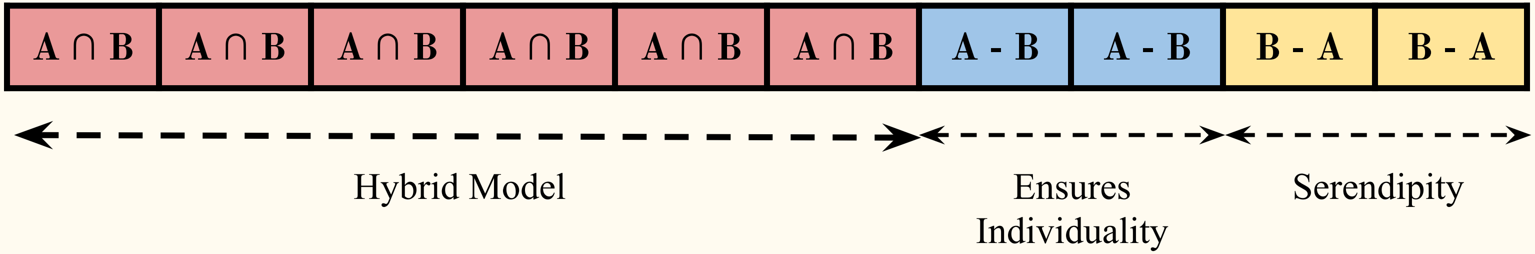
- Weights assigned to various parameters (**action strength**):
    - 'VIEW': 1.0
    - 'LIKE': 2.0
    - 'BOOKMARK': 3
    - 'FOLLOW': 3.0
    - 'COMMENT CREATED (if liked)': 3.5

- Weights of parameters = (100 - n*8)% of assigned weights
    - Where n = number of months ago a particular action was taken
    - min(n) = 12
    - **Time Strength** = (100 - n*8)

Event Strength = Action Strength * Time Strength

Regular Users

- Actions > 5
- Interactive content based filtering (A) - Items which are similar to user's preferences
- Collaborative Filtering (B) - Items which are preferred by users similar to our current user.



Content based

- Finding similarity using TF - IDF
- TF - IDF reflects how important a word is an article of a given corpus.

- TF(t) = (Number of times term t appears in a document) / (Total number of terms in the article)
- IDF(t) = log_e(Total number of articles/ Number of articles with term t in it)
- TF-IDF score = TF*IDF

Content based

	Word 1	Word 2	Word n	Word 5000
Article 1	3	2	1	5
Article 2	5	3	2	1
Article n+1	2	3	2	4

Rating
3
4
5

User Profile

Word 1	Word 2	Word n	Word 5000
8	6	3	6

New User Profile

Word 1	Word 2	Word n	Word 5000
7	6	3	7

Content based

- User profile is compared with the vector of every article and the most similar articles are found using Cosine Similarity.

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\,\|\vec{b}\|}$$

Collaborative filtering

- Matrix Factorization using SVD
- The unknown ratings are found using the Gradient Descent Algorithm
  - $\boldsymbol{\alpha} = 0.03$
  - $n = 10$

New Users

- 10 clusters of articles formed using centroid based based K-means clustering
- Most popular article displayed for each new user (with no actions)

- For every action ($<$ 5) n * 2 content based recommendations, and (10 - 2*n) recommendations on the basis of the above mentioned method.

Suggestions for New User

```
*************NEW USER**************


Registered User[Y/N]: N
Your user Id is : 1042
Top 10 Recommendations :


1. Analyst: Google's cloud business could cause the stock to soar to over $900
2. Fooling The Machine
3. Apple Invites Media to 'Hello Again' October 27th Mac-Centric Event
4. Prototipação: erre cedo para acertar cedo - Hipsters #28
5. Stackdriver Trace for App Engine is GA; app latency has nowhere to hide
6. SOA com microserviços - Sensedia
7. Probabilistic Programming
8. Braincast 207 - A Revolução das Máquinas Inteligentes
9. Making digital strategy a reality in insurance
10. Introducing online resizing of Google Cloud Persistent Disks without downtime


Select Article number (0 for Search) : 3


You can select one of the actions below:
1.Like
2.View
3.Comment
4.Bookmark
5.Follow


Enter your entry : 1
```

Train-Test Split

- Split by a reference date
- The train set is composed by all interactions before that date - 31284
- The test set are interactions after that date - 7822
- Split ratio- 4:1

Accuracy

Testing the accuracy of our model.

- Recall@n methodology used
- Percentage of the number of items actually viewed from n recommendations
- 52% - recall@10
- 40% - recall@5

Accuracy

Testing the accuracy of our model.

modelName	recall@10	recall@5
Popularity	0.341984	0.219637
Collaborative Filtering	0.415750	0.290335
Content-Based	0.510611	0.397852
Final	0.520583	0.394656

eventStrength	title
3.735522	6 reasons why I like KeystoneML
3.726831	Auto-scaling scikit-learn with Spark
3.475085	5 reasons your employees aren't sharing their ...
3.475085	At eBay, Machine Learning is Driving Innovativ...
3.356144	Algorithms and architecture for job recommenda...
3.339137	10 Stats About Artificial Intelligence That Wi...
3.269033	AI Is Here to Help You Write Emails People Wil...
3.195348	Deep Learning for Chatbots, Part 1 - Introduction
3.104337	Graph Capabilities with the Elastic Stack
3.007196	Being A Developer After 40 - Free Code Camp
2.887525	Building with Watson Technical Web Series
2.855990	Worldwide Ops in Minutes with DataStax & Cloud
2.782409	5 Unique Features Of Google Compute Engine Tha...
2.687061	How to choose algorithms for Microsoft Azure M...
2.632268	Bad Writing Is Destroying Your Company's Produ...
2.608809	Creative Applications of Deep Learning with Te...
2.594549	How Netflix does A/B Testing - uxdesign.cc - U...
2.536053	Machine Learning Is No Longer Just for Experts
2.464668	How Google is Remaking Itself as a "Machine Le...
2.454176	Text summarization with TensorFlow

User #1

Recommendations generated

contentId	title
3269302169678465882	The barbell effect of machine learning.
5092635400707338872	Power to the People: How One Unknown Group of ...
-7126520323752764957	How Google is Remaking Itself as a "Machine Le...
7395435905985567130	The AI business landscape
1415230502586719648	Machine Learning Is Redefining The Enterprise ...
-5756697018315640725	Being A Developer After 40 - Free Code Camp
7395435905985567130	The AI business landscape
-8085935119790093311	Graph Capabilities with the Elastic Stack
5250363310227021277	How Google is Remaking Itself as a "Machine Le...
638282658987724754	Machine Learning for Designers

Conclusion

The highlights of the model

- Model based on user behaviour
- Dynamic recommendations on timestamp
- Tackling the problem of ‘groupism’ born out of CF by actively preserving individual choice
- Serendipity and Exploration an integral part
- Problem of Cold Start effectively tackled

Thank You!

Questions?