News Article Recommender System

Abstract

This report is aimed to propose a recommender system for a start-up JhakasNewsVaala. This organisation is developing an application for users which recommends them news articles based on their interests and the articles liked by similar peer groups.

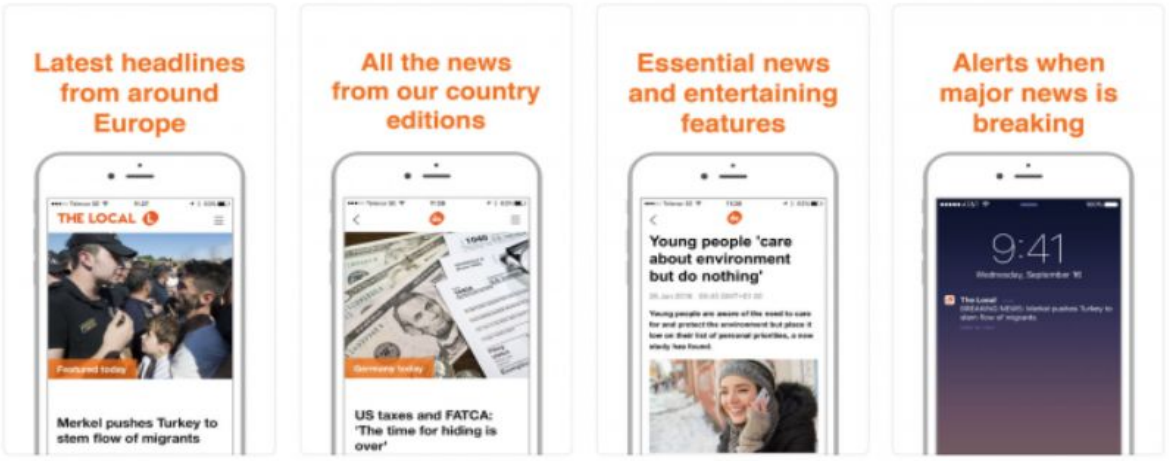
Introduction to Recommender Systems

THE OBJECTIVE of a RecSys is to recommend relevant items for users, based on their preference. Preference and relevance are subjective, and they are generally inferred by items users have consumed previously

Reading the news online has exploded as the web provides access to millions of news sources from around the world. The sheer volume of articles can be overwhelming to readers. Therefore, building a news recommendation system to help users find articles that are most interesting to them is a crucial task for every online news service.

Need

Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. Recommender Systems give an unique experience to their users which inturn will help in reach of the business and its growth. Also large volumes of data available makes it difficult to choose from. The personalised suggestions are always welcomed.



Usage of recommender Systems in Industry

ECommerce Platforms: Flipkart, Amazon(Recommended items)
Streaming Services: Netflix, hotstar (Items you may like)
Social Platforms- Instagram (Feed)

Research

A news recommendation system helps users find articles that are most interesting. A recommendation system can categorised broadly into 3 categories:

1. Content based
2. Collaborative
3. Hybrid

News recommendation systems must be able to handle the challenge of fresh content: breaking news that hasn't yet been viewed by many readers. Thus we need to leverage the data associated with the article content available at time of publishing — such as topics, categories, and tags — to build a content-based model, and match it to readers' interests learned from their reading histories. However, one drawback of content-based recommendations is that when there is insufficient user history, the coverage of the recommendations is very limited. This is referred to as the cold-start problem, common in recommender systems.

Data Collection

We scraped 20,000 articles from a popular news website www.indianexpress.com from their central news articles category and assigned an articleid to them with their timestamp, URL as shown in image.

article_id	headline	desc	date	url	articles
5	DNA sampling in rape cases: MP showcauses 1,25...	On July 18, 2019, the government submitted a r...	August 11, 2019 4:27:34 am	https://indianexpress.com/article/india/dna-sa...	The Madhya Pradesh government has served showc...
4	Sonia to Rahul to Sonia: Congress takes step b...	Sonia was appointed interim president until or...	August 11, 2019 9:44:49 am	https://indianexpress.com/article/india/sonia-...	Three months after Rahul Gandhi quit as party ...
3	Pay cash relief to flood-hit people: Congress ...	Sachin Sawant, general secretary and spokesper...	August 11, 2019 9:08:30 am	https://indianexpress.com/article/india/pay-ca...	SLAMMING THE state for its decision to deposit...

Our Main purpose of scraping news articles was to get a real mix of articles around which we can build our recommender system and have a variety of topics

Data Generation

Clickstream Dataset

We have generated clickstream data taking 25000 users into account.

Steps:

- We have used Geometric Distribution for a number of users in various sessions with probability of an individual success equal to 0.5.
- We have applied bernoulli distribution for number of clicks on 10 recommended items for every user in each session with a probability of clicking as 0.25.
- We generated time using mix Gaussian distribution using 3 means and deviations according to the article length and average time required to read them. By randomly assigning time we calculate % of articles read by a user by taking in account the ideal time to read that specific article based on the length of that article.
- We have also included a clickbait scenario. We have computed interest finally as a rating for our recommender system.
- We have computed article length in the form of characters. And assumed that the average reading speed of the user is 20-25ch/s where the average word in the document is of 5-6 characters.

user_id	session_id	article_id	click	click_value	article_length	timeonpage	read_percentage	updated_percentage	interest
12049	session-7047	10295	0	not_view	2615	0.000000	0.000000	0.000000	0.000000
12049	session-7047	15353	0	not_view	1637	0.000000	0.000000	0.000000	0.000000
12049	session-7047	751	1	view	1098	20.119364	45.809117	50.809117	0.508091
12049	session-7047	7493	1	view	3158	129.134206	102.227839	100.000000	1.000000
12049	session-7047	2999	0	not_view	3298	0.000000	0.000000	0.000000	0.000000

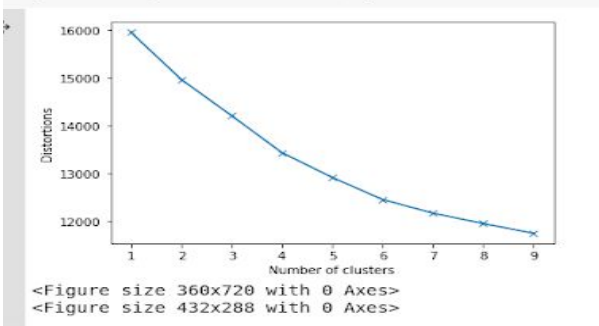
Content Based

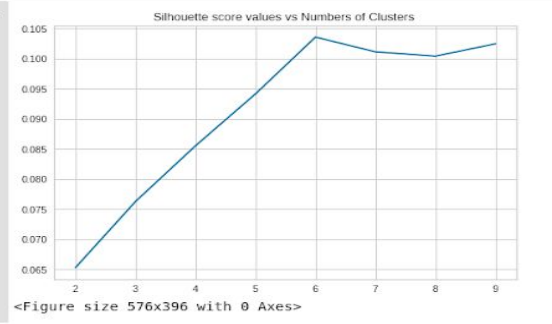
Clustering:

We used the TF-IDF to convert the article data into vectors. And used K-Means clustering to make the clusters of those vectors.

TFIDF score is used mainly because TF-IDF weighting negates the effect of high frequency words in determining the importance of an item. Log weighting is used to dampen the high frequency weight.

To analyse the optimal number of clusters, we used the following to methods:





1. Elbow curve:
2. Silhouette analysis

In both the curves, the optimal number of clusters are 6

Implementation:

- We will cluster the article dataset and for addressing cold start problems we will randomly choose 10 articles from these clusters.
- We used cosine similarity to find similar articles of interest(user profile vs all item profiles)
- We recommended them on the basis of decreasing strength of cosine similarity.

Collaborative Filtering

Collaborative methods for recommender systems are methods that are based solely on the past interactions recorded between users and items in order to produce new recommendations. The main idea that rules collaborative methods is that these past user-item interactions are sufficient to detect similar users and/or similar items and make predictions based on these estimated proximities.

Implementation

- Create a users vs item matrix with ratings as values.
- We factorise the matrix and apply SVD.
- We predict every user's rating for every item in the corpus .
- Based on these ratings wwe will recommend top 10 items rated for that user.

Hybrid Recommendation System

This type of recommendation system takes in account both user's item interest and reading history and tries to provide recommendations based on these.

We have not trained hybrid model for recommendation, instead we have assumed a ratio of articles from collaborative and content based recommendation of the same user and shuffled them. The ration was 7:3 (collaborative:content).

What did not worked:

Comparing accuracy using the same metric for all techniques. It is difficult to evaluate a recommender system.

Future Scope:

- Apply LSH(Locality Sensitive Hashing) to see the results.
- Improvising Clickstream Data (Increasing number of users)
- To use the case where the user has not clicked on articles in top 10 recommendations.

Peer Review:

Ratio for hybrid 7:3 was recommended by a team to us. They found results fascinating and wanted clickstream to be improvised

Results:

User Profile:

	token	relevance
4990	tmc	0.088936
4991	kolkata	0.097315
4992	idol	0.097854
4993	congress	0.099609
4994	police	0.106238
4995	woman	0.112096
4996	policemen	0.121565
4997	nrc	0.146956
4998	trinamool	0.210515
4999	trinamool congress	0.215602

Content Based:

article_id	headline	url
10970	Tamil Nadu cop killed in revenge attack for te...	https://indianexpress.com/article/india/tamil-...
16176	Across party lines, one demand: Need more aid ...	https://indianexpress.com/article/india/corona...
8081	Woman sits on solitary protest outside Parliam...	https://indianexpress.com/article/india/woman-...
13383	Kanhaiya's convoy attacked in Ara, two injured...	https://indianexpress.com/article/india/kanhai...
16943	Lockdown diary, Day 31: Yamraj chases lockdown...	https://indianexpress.com/article/india/lockdo...
4464	J&K won't be UT forever, says Amit Shah	https://indianexpress.com/article/india/jk-won...
11729	UP CAA protests: Firozabad FIRs same, accused ...	https://indianexpress.com/article/india/up-caa...
15209	'Cattle smuggler' killed in Mathura encounter:...	https://indianexpress.com/article/india/cattle...
13439	Firing near CAA protest site in Bihar, 2 detained	https://indianexpress.com/article/india/firing...
18593	Covid-19 cases in Maharashtra Police reaches 1...	https://indianexpress.com/article/india/covid-...

Collaborative Filtering:

article_id	headline	url
4970	Kerala state Lottery Today Results announced: ...	https://indianexpress.com/article/india/kerala...
8723	Citizenship Amendment Bill: SAD backs legislat...	https://indianexpress.com/article/india/citize...
3279	Pakistan continues ceasefire violation for thi...	https://indianexpress.com/article/india/pakist...
3247	Gandhian Leelatai Merchant dies	https://indianexpress.com/article/india/gandhi...
18476	Nirmala Sitharaman Announcements highlights: P...	https://indianexpress.com/article/india/fm-nir...
13925	Navy's MiG 29k aircraft crashes near Goa, pilo...	https://indianexpress.com/article/india/indian...
18122	UP: In Jhansi, 300 civil volunteers act as for...	https://indianexpress.com/article/india/up-in-...
4509	Dantewada: Police say Maoist killed in encount...	https://indianexpress.com/article/india/dantew...
4404	PMC Bank fraud: Ex-chairman sent to police cus...	https://indianexpress.com/article/india/pmc-ba...
841	Justice Sunil Gaur, who handled key cases, ret...	https://indianexpress.com/article/india/justic...

Hybrid:

article_id	headline	url
8723	Citizenship Amendment Bill: SAD backs legislat...	https://indianexpress.com/article/india/citize...
13439	Firing near CAA protest site in Bihar, 2 detained	https://indianexpress.com/article/india/firing...
4404	PMC Bank fraud: Ex-chairman sent to police cus...	https://indianexpress.com/article/india/pmc-ba...
13383	Kanhaiya's convoy attacked in Ara, two injured...	https://indianexpress.com/article/india/kanhai...
18122	UP: In Jhansi, 300 civil volunteers act as for...	https://indianexpress.com/article/india/up-in-...
13925	Navy's MiG 29k aircraft crashes near Goa, pilo...	https://indianexpress.com/article/india/indian...
4970	Kerala state Lottery Today Results announced: ...	https://indianexpress.com/article/india/kerala...
3279	Pakistan continues ceasefire violation for thi...	https://indianexpress.com/article/india/pakist...
10970	Tamil Nadu cop killed in revenge attack for te...	https://indianexpress.com/article/india/tamil-...
3247	Gandhian Leelatai Merchant dies	https://indianexpress.com/article/india/gandhi...