Recommender Systems for Instacart

A Data Mining Analysis

Team 7: Nakul Camasamudram, Rosy Parmar, Rahul Verma, Guiheng Zhou

Northeastern University

Content

Introduction

Background Information

Background:

- Instacart is an American company that operates as a same-day grocery delivery service.
- Customers select groceries through a web/mobile application from various retailers which are then delivered by a personal shopper.
- With data on customers product purchases, a recommender system can be designed using unsupervised learning methods.

Investigation

In this project we investigate if products suggested by our recommender systems are more indicative of user purchases on Instacart as compared to recommending the 10 most popular products to each user.

Experiment Design

Data Collection and Preprocessing

Data Source: The Instacart Online Grocery Shopping Dataset 2017 - around 3 million orders from more than 200,000 users.
Missing Values and Outliers: None

Data Transformations and Partitions

- Purchases of all users are split into prior and current.
- Prior purchases are used to build **utility matrix**, with products as rows, users as columns and entries as purchase frequencies.
- A random 20% of the current purchases was used as **test data**

Applying Methods

Methods Chosen: Collaborative Filtering

- Neighborhood based. Cosine similarity on a TF-IDF weighted matrix.
- Matrix Factorization(MF) using Singular Value Decomposition(SVD).
- MF for Implicit Feedback using Alternating Least Squares(ALS).

Data Mining Analysis

TF_IDF based Neighborhood

Why TF-IDF?

It's originally a method to evaluate the importance of a word

wWWWw

ANALOGY

Word	TERM
	DOCUMENT
	QUERY
	SIMILAR DOCUMENTS Related to **QUERY**
	PRODUCT
	PURCHASE HISTORY
	PURCHASE HISTORY of a **TARGET USER**
	PURCHASE HISTORIES OF USERS Similar to **TARGET USER**

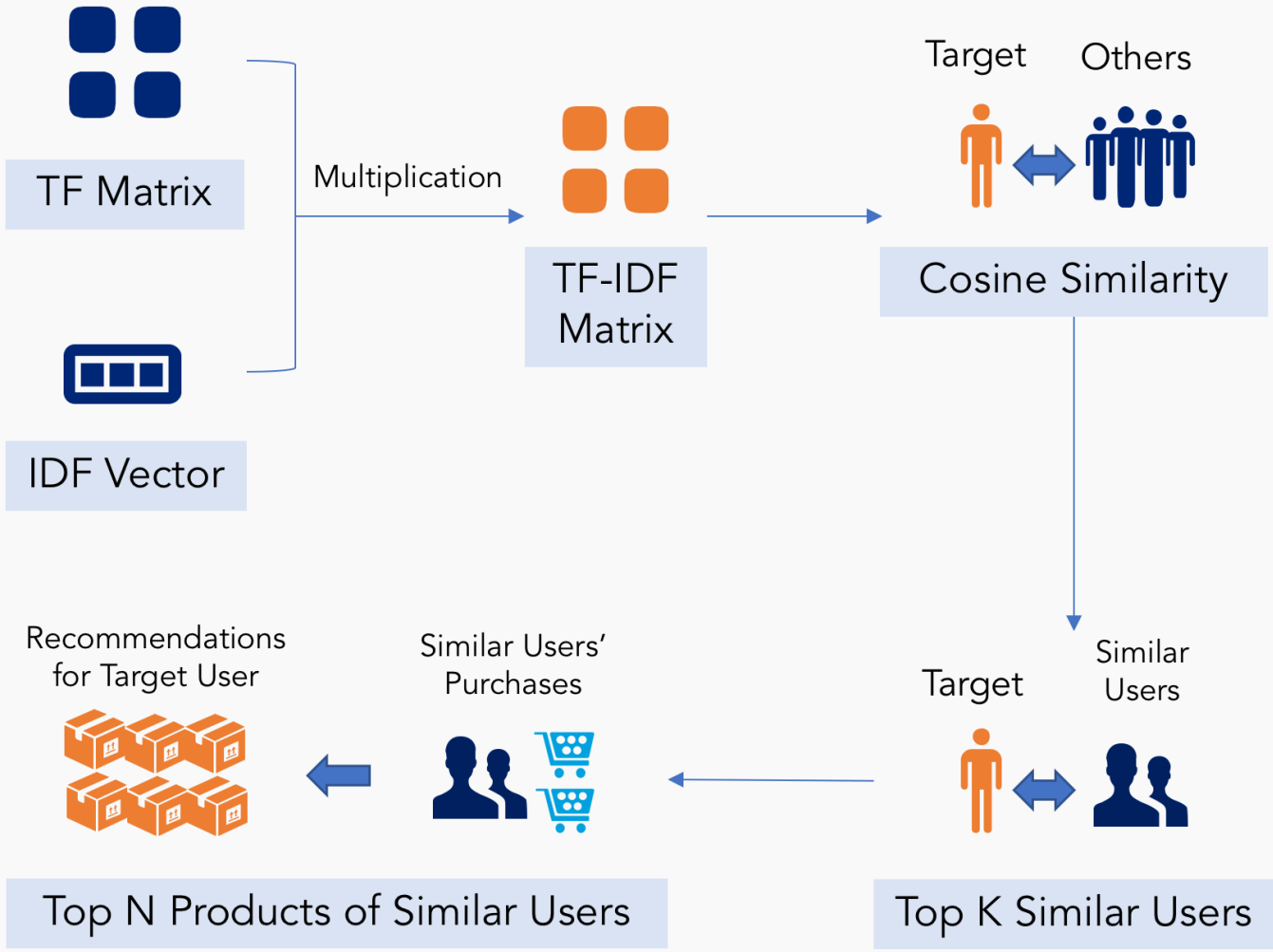
TF_IDF based Neighborhood

What is a TF-IDF matrix?

- Each column of the matrix is a term, more specifically a product users purchased;
- Each row of the matrix is a document, representing for the purchase history of a user, comprised of varied products.

TF_IDF based Neighborhood

How does TF-IDF work?



Singular Value Decomposition

Why?

- The utility matrix is 99.87% sparse. Will uncovering latent features through matrix factorization give better results than neighborhood-based models?
- Used SVD to generate a low rank approximation of the utility matrix

Singular Value Decomposition

How?

- Based on SVD formula below we find an approximate [product,user] matrix A where U and V^T are the left and right singular vectors and D is the singular value for largest singular values.

$$A = UDV^T$$

- Similarly, Product Vector for a user $i = U[i] * D * V^T$,the values in the vector represents the user preference for products. i.
- Then we find top K products preference values excluding products previously purchased by a user and recommend it.

MF for Implicit Feedback

Implicit vs Explicit:

- No explicit negative feedback.
- Implicit feedback is inherently noisy.

MF for Implicit Feedback

Implement method specifically designed for implicit feedback datasets. [1]

Features:

- Binary preferences p_{ui}.
- Confidence in observing p_{ui}. $c_{ui} = 1 + \alpha r_{ui}$.
- Optimize using ALS:

$$\min_{x,y} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2)$$

[1] Hu Y. & Koren Y. & Volinsky C. (2008) Collaborative Filtering for Implicit Feedback Datasets

MF for Implicit Feedback

- Transform utility matrix into a matrix of confidence values.

$$C_{ui} = 1 + \alpha R_{ui}$$

- $\alpha = 15$ worked best.
- Number of factors $= 100$.

Results

Evaluation

Recommendations are compared to the current order for every user in the test dataset.

Metric:

- No negative feedback. Hence, Precision-based metrics ruled out.
- We use recall averaged over every user in the test dataset.

$$\text{Mean Recall} = \sum_{\text{test users}} \frac{|\{\text{recommended}\}| \cap |\{\text{actual}\}|}{|\{\text{actual}\}|}$$

Results

Test dataset: around 26,000 users

Baseline: 10 most popular products to every user.

	Mean Recall (%)	Running time(minutes)
Baseline	2.62	-
TF-IDF Neighborhood	20.08	169
MF using SVD with 50 factors	2.84	12
Implicit MF using ALS	4.13	2

Discussion

Performance Useful?

Question: Will products suggested by our recommender systems result in more purchases on Instacart as compared to the baseline model?
Answer: Yes.

Potential Improvements?

- Cold Start problem. Implement hybrid recommender using Content-Based methods.
- More data .