République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene

Recommendation Systems Based On Covariance Distance Similarity Case Study YASSIR EXPRESS

Faculté de Mathématique
Département de Probabilités et de Statistiques
Domaine Mathématiques et Informatique

Mémoire

Pour l'obtention du Diplôme de Master
Probabilités et Statistiques Appliqués

Présente par :
Nihad Senhadji

Encadré par :
Mr. Medkour Tarek
Mr. Hocine Abdelouaheb
(Yassir Express)

Soutenu le **29 juin 2022**

Table Of Contents

Introduction To RSs

The enormous amount of digital information generated by the increasing number of users on the internet has created the challenge of information overload, It becomes challenging and time consuming for users to retrieve the exact information from the web.
This overwhelming size of data has shifted the focus of research community from simple information extraction to filtering of information this has increased the demand for recommendation systems,
which provide suggestions of items to users.

Objective

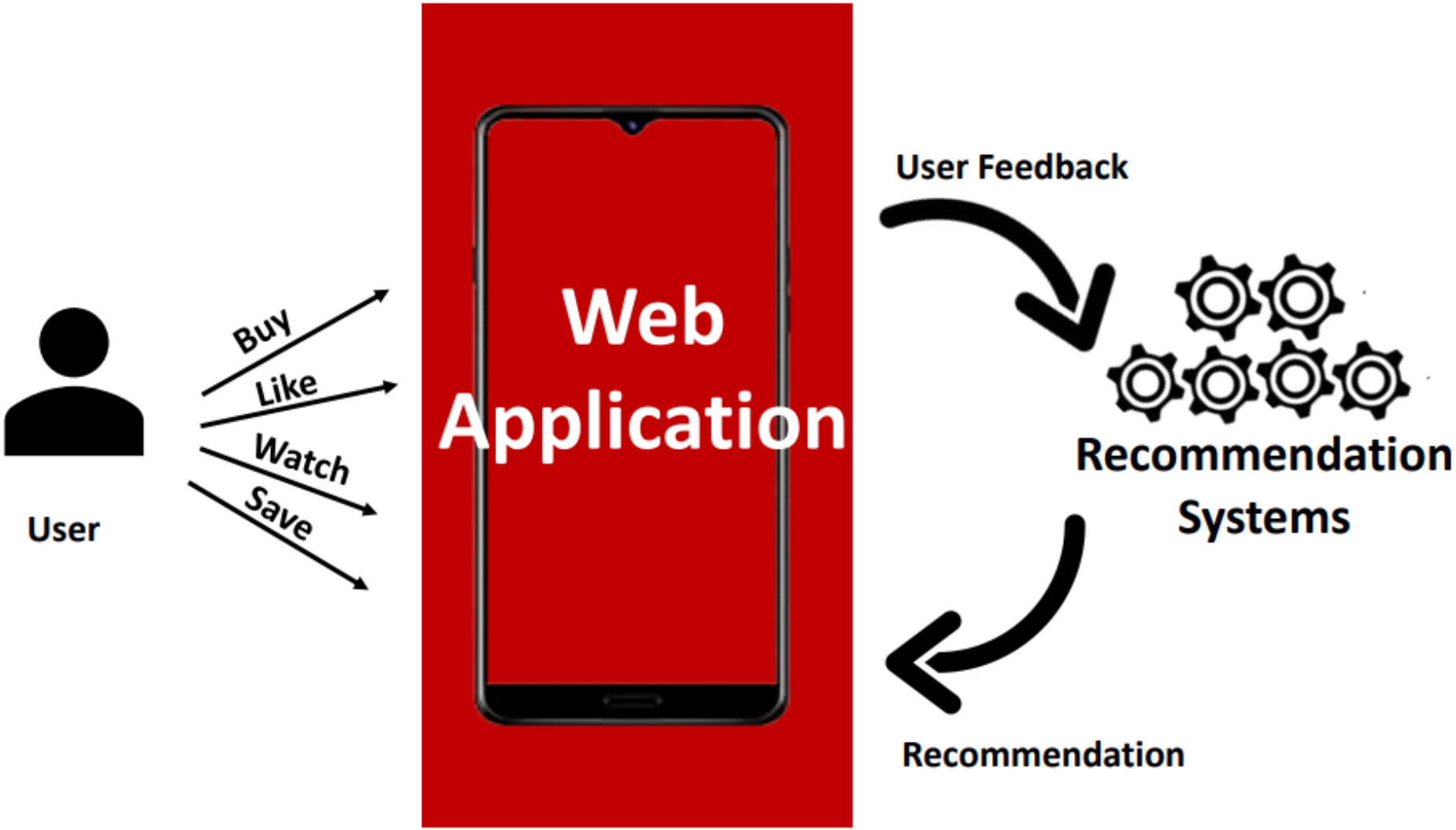
This presentation provides the results of a building recommendation systems project dealing with new similarity measure Distance Covariance in terms of their performances and effectiveness compared to several similarity measures that have been proposed. We implement these systems using the YASSIR EXPRESS data set to build a restaurant recommendation system as a case study.

What are Recommendation Systems ?

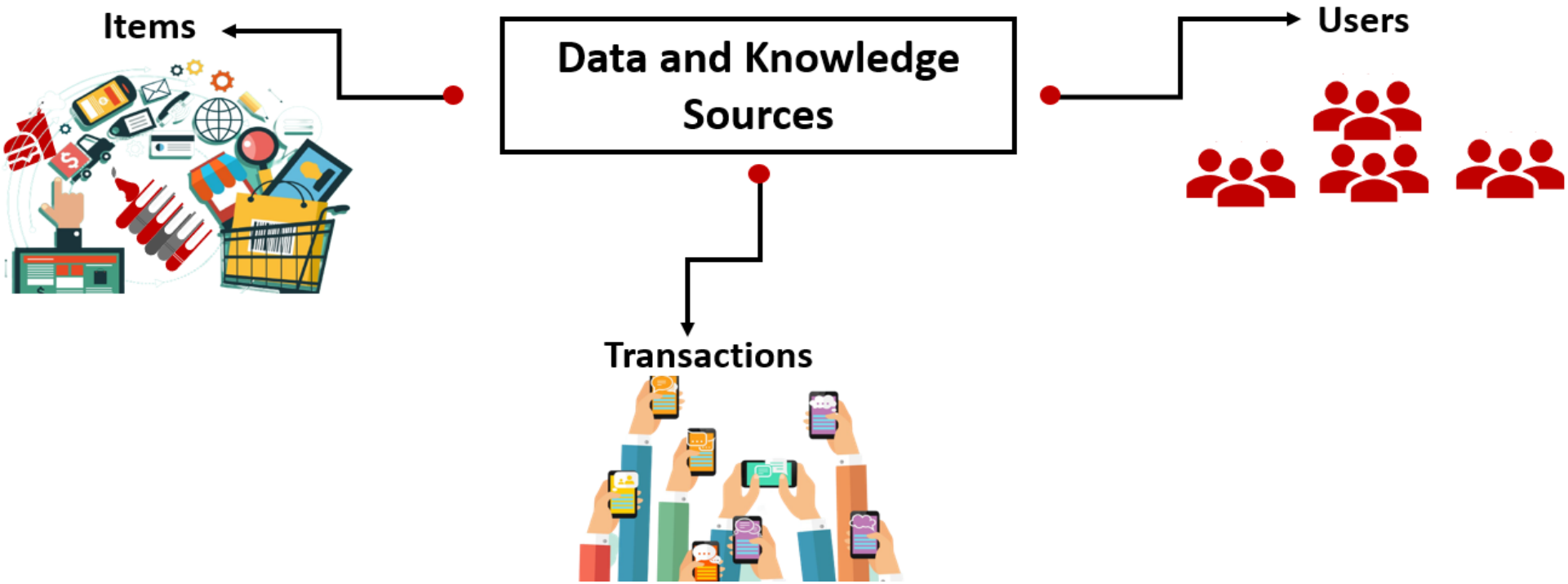
Recommendation Systems

Recommendation systems are an information filtering system that aims to provide users with relevant items,they were introduced in the early 1990s as tools that help them deal with overloaded data by generating personalized recommendations based on their preferences and interests.

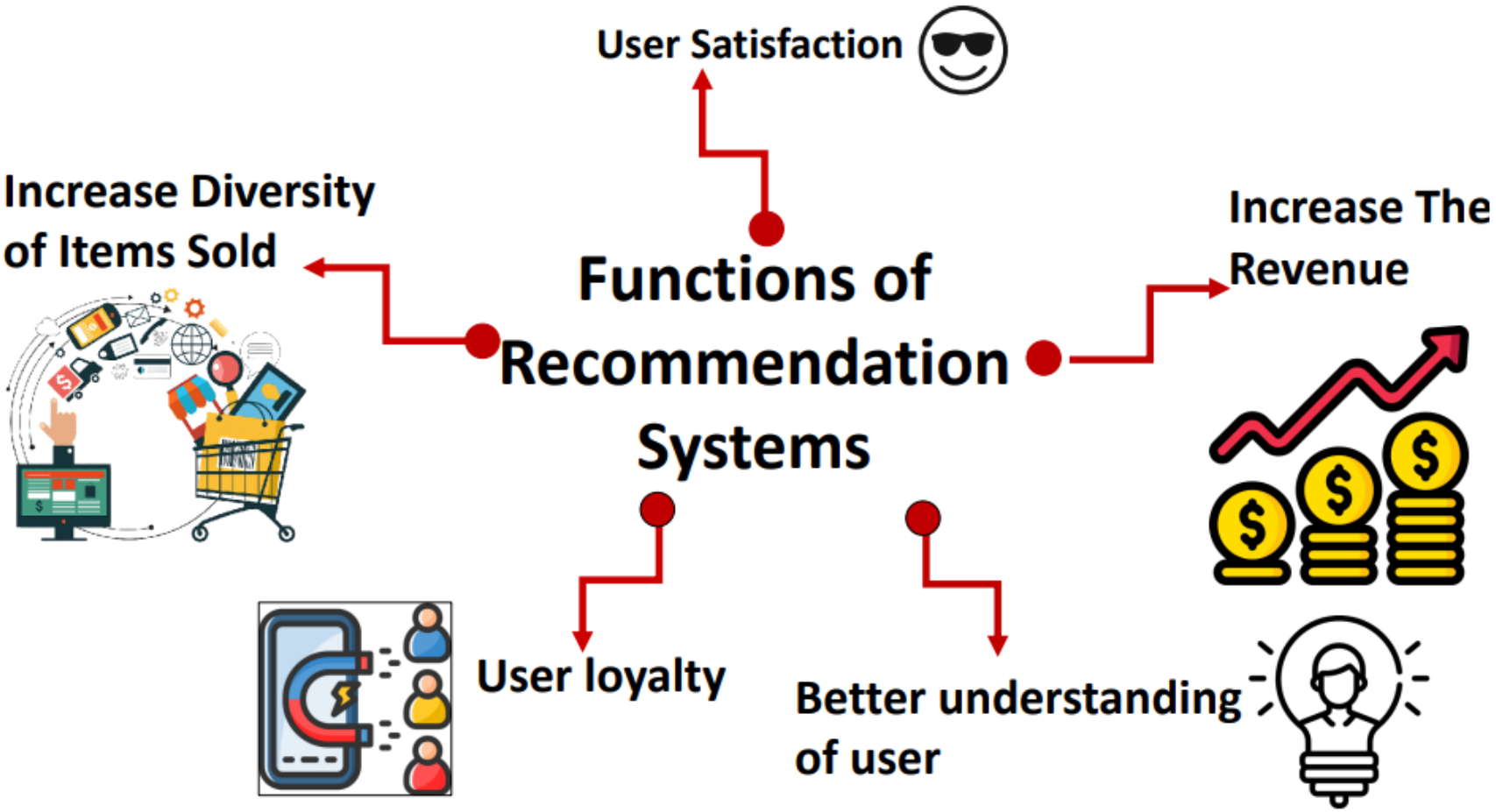
Recommendation Systems



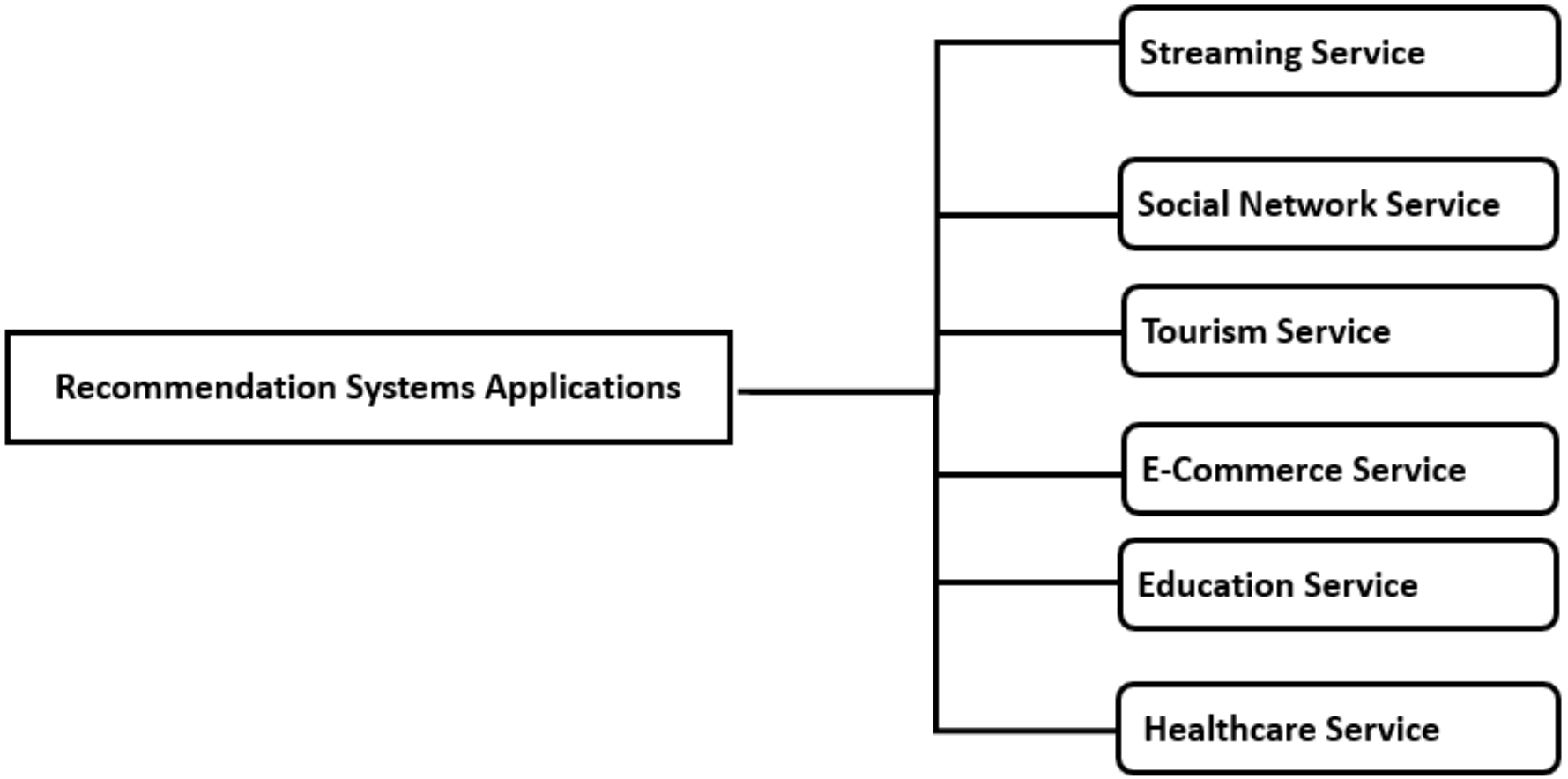
Data and Knowledge Sources



Functions of Recommendation Systems



Recommendation Systems Applications in The industry



Implementations Of Recommendation Systems In the Industry

amazon

Google News

facebook

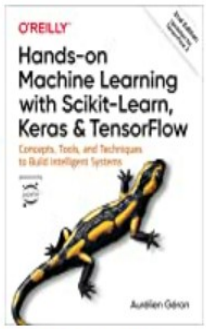
NETFLIX

Netflix Recommendation System

Amazon Recommendation System

Customers who viewed items in your browsing history also viewed

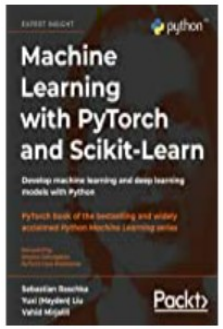
Page



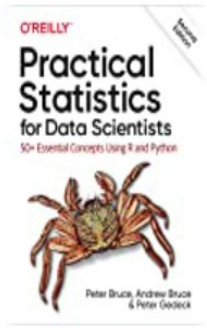
Hands-On Machine Learning with Scikit-Learn, Keras, and...
›Aurélien Géron
4.5 out of 5 stars 2,851
Paperback
#1 Best Seller in Computer Vision & Pattern Recognition
$49.69
$49.98 shipping

Deep Learning with Python, Second Edition
Francois Chollet
4.5 out of 5 stars 65
Paperback
$39.49
$49.98 shipping

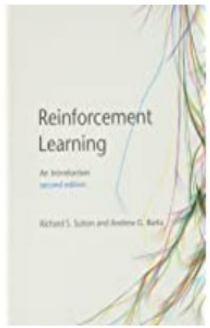
Recommender Systems Handbook
Francesco Ricci
4.5 out of 5 stars 3
Hardcover
$372.08
$49.98 shipping
Only 2 left in stock (more ...



Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep...
›Sebastian Raschka
4.5 out of 5 stars 77
Paperback
$44.99
$49.98 shipping

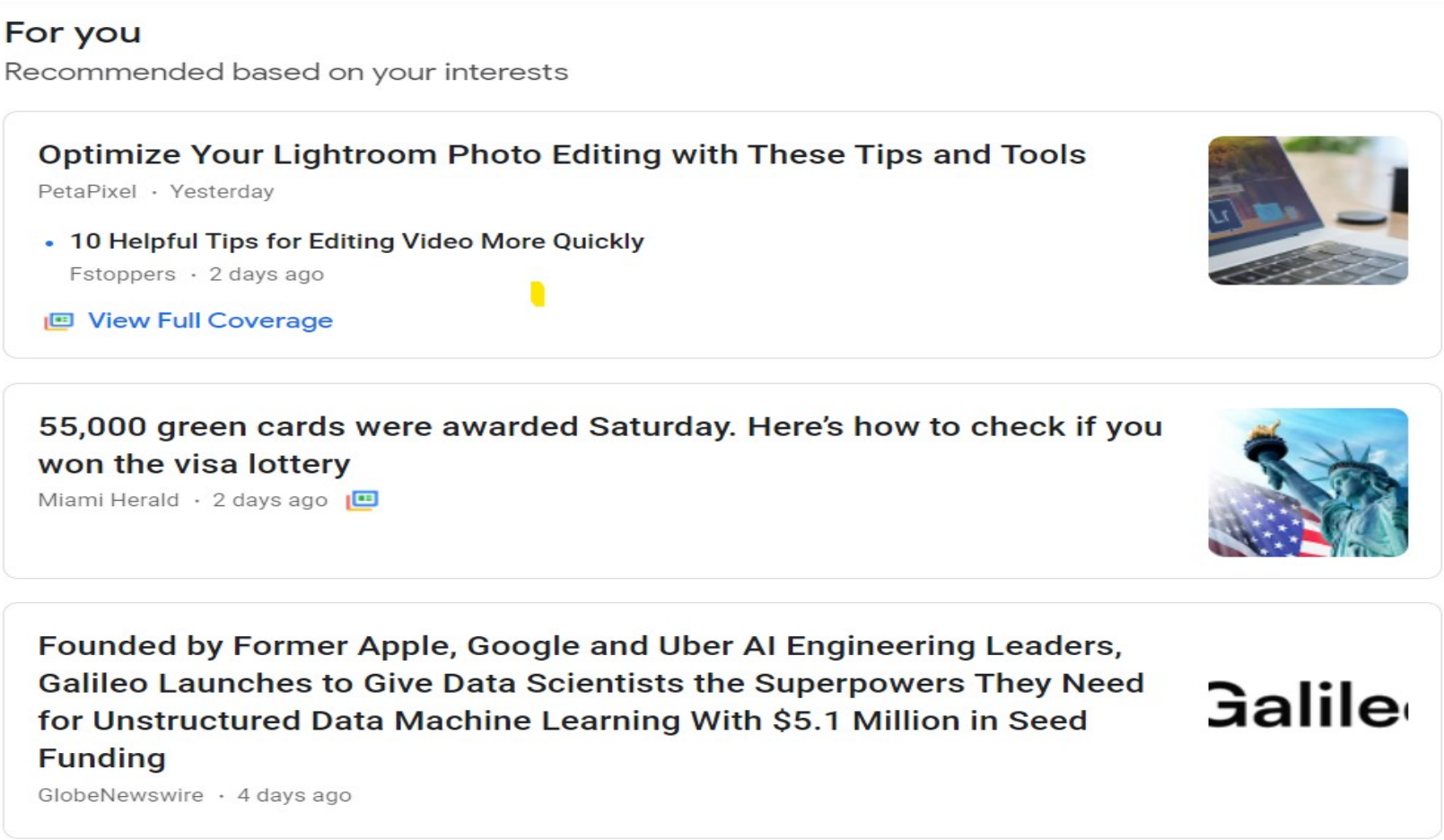


Practical Statistics for Data Scientists: 50+ Essential Concepts Usi...
Peter Bruce
4.5 out of 5 stars 453
Paperback
#1 Best Seller in Mathematical Analysis
$31.49
$49.98 shipping

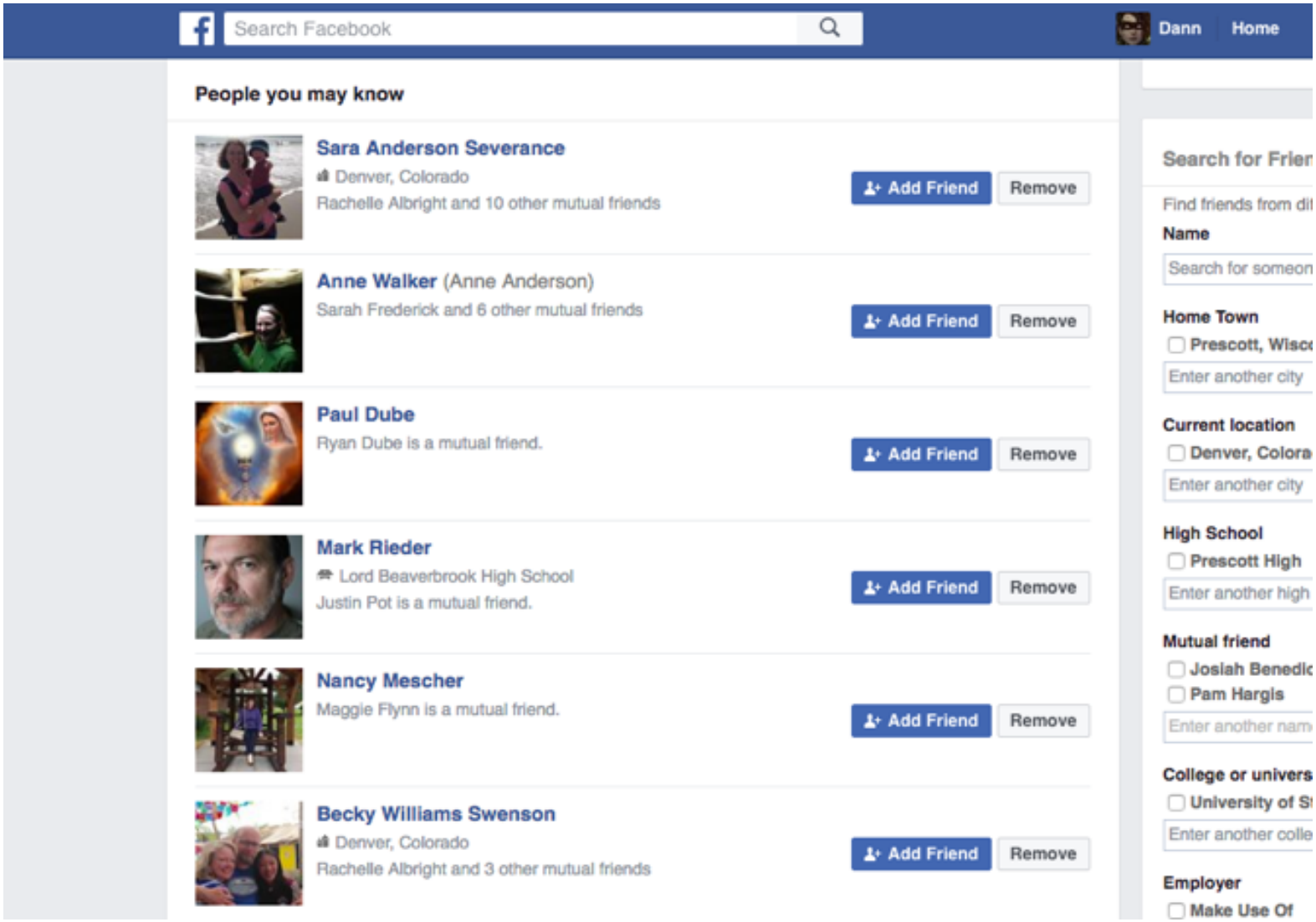


Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning...
›Richard S. Sutton
4.5 out of 5 stars 383
Hardcover
$63.41
$49.98 shipping

Google News Personalization Recommendation System



Facebook Friend Recommendations System



Models of Recommendation Systems

There can be two types of recommendations
Personalized and Non-Personalized Recommendation Systems.

The Non-Personalized Recommendations

The Non-Personalized Recommendations

Do not have a direct impact on the users' buying habits and personal interests, Such as **Popularity Based Recommendation System**.

Popularity Based Recommendation system

Popularity Based RSs

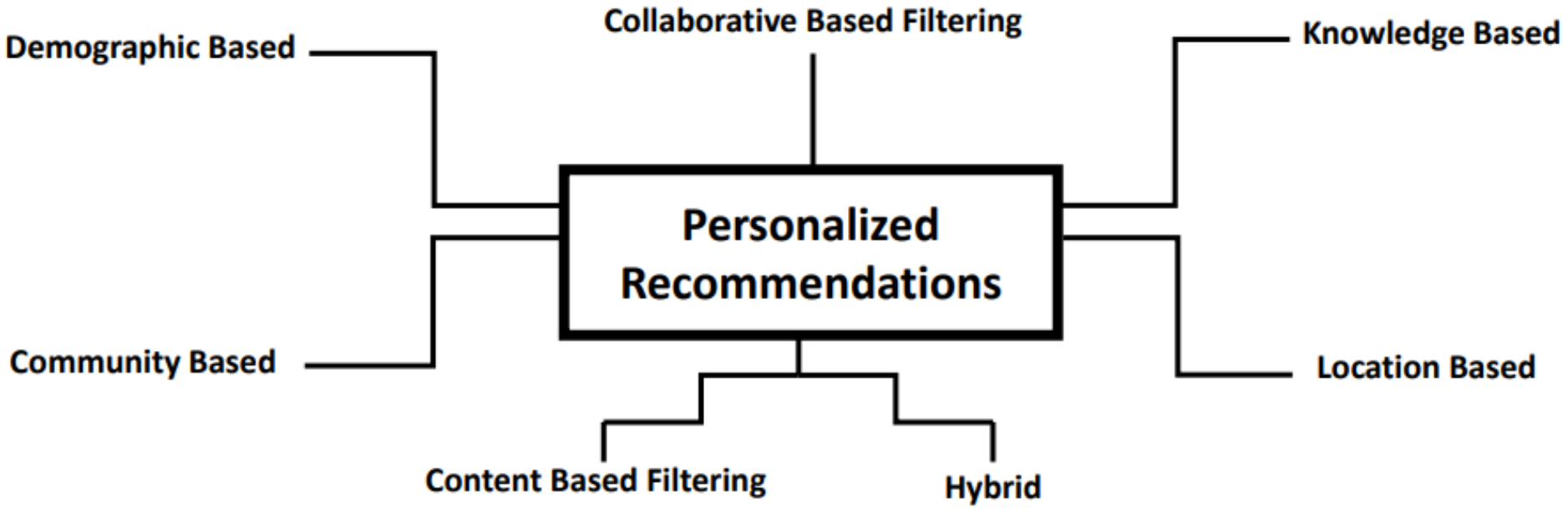
Uses the items which are in trend right now, the bestselling item or most popular item present in the web application,It is a very fundamental type of recommendation system.

The Personalized Recommendations

The Personalized Recommendations

Rely on a user perspective and preferences, and the recommendation are based on the user behavior and interaction.

The Personalized Recommendations



Content Based Filtering

Content Based Filtering

Makes recommendations by using keywords and attributes assigned to items and matching them to a user profile.

Hybrid

Hybrid

Is a special type of recommendation system which can be considered as the combination of the content and collaborative filtering RSs together.
A Hybrid recommendation model was proposed to solve the limitations of those models and to improve the recommendation performance.

Demographic Based

Demographic based

This type of system recommends items based on the demographic profile of the user, such as age,gender,location and education, etc,by classifying users into groups .

Community Based

Community based

Recommend items to the user based on the preference of his friends.

Knowledge Based

Knowledge based

Recommend items based in general on the knowledge about items fit to users' preferences,it used when nothing is known about users' behavior .

Location Based

Location Based

Incorporate the location of users to provide relevant and precise recommendations.

Collaborative Based Filtering Recommendation System

Collaborative filtering is the process of filtering items through the opinions of other users.
There are two types of models that are commonly used in collaborative Filtering :

- **Neighborhood-based (Memory-Based)**
- **Model-Based**

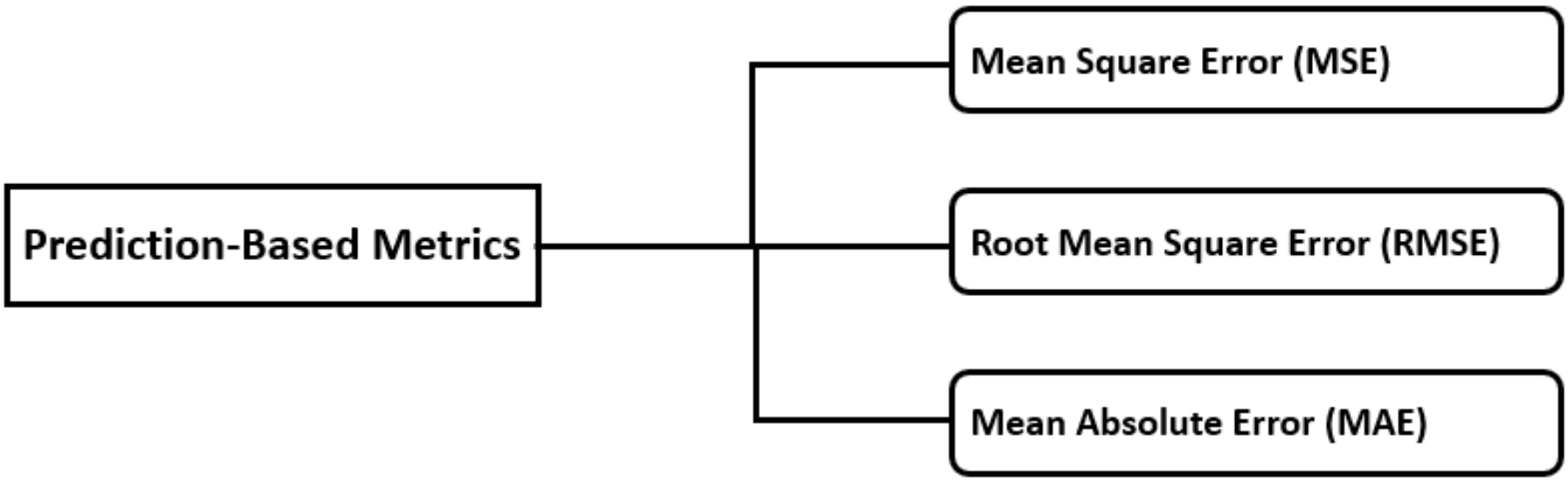
Evaluating Recommendation Systems

The performance of the Recommendation Systems models can be measured in many ways, and different metrics can be applied.

Prediction Based Metrics

Used to measure the difference between predicted rating $\hat{r_{ui}}$ and the rating given by the user r_{ui}.

Evaluating Recommendation Systems



Evaluating Recommendation Systems

- ***Root Mean Square Error (RMSE)***
  it evaluates the difference between the ratings predicted $\hat{r_{ui}}$ by the RSs and the ratings given by the user r_{ui} .

$$\textbf{RMSE} = \sqrt{\frac{\sum_{(u,i,r)\epsilon R}(\hat{r_{ui}} - r_{ui})^2}{|R|}}$$

  Where R denotes the amount of user u rated items,r_{ui} determines the actual rating that user u rates item i and $\hat{r_{ui}}$ denotes the predicted rating of item i for user u .

Evaluating Recommendation Systems

- ***Mean Absolute Error (MAE)***
  it evaluates the difference between the ratings predicted by the RSs and the ratings given by the users. It returns a positive value.

$$\textbf{MAE} = \frac{\sum_{(u,i,r)\epsilon R} |\hat{r_{ui}} - r_{ui}|}{|R|}$$

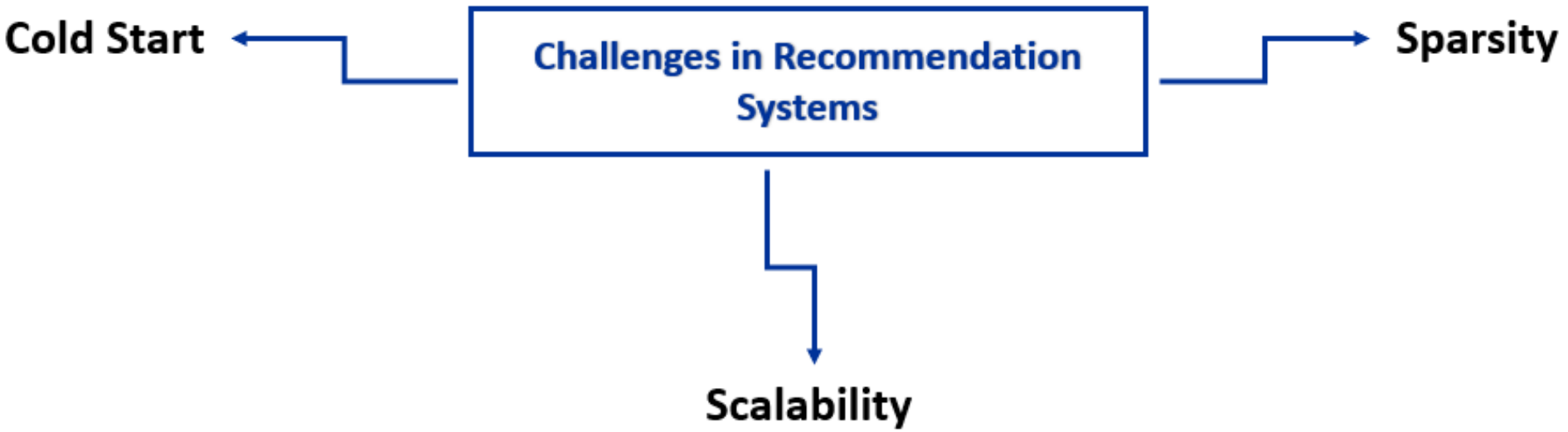
The lower the MAE,the more accuracy the recommendation system predicts rating.

Evaluating Recommendation Systems

- ***Mean squared error (MSE)***
  The MSE is used to evaluate recommendation systems by computing the average value of all absolute value differences between the predicted ratings $\hat{r_{ui}}$ and the user's true ratings r_{ui}.
  The MSE can be computed using the following equation

$$\textbf{MSE} = \frac{\sum_{(u,i,r)\epsilon R}(\hat{r_{ui}} - r_{ui})^2}{|R|}$$

Challenges in Recommendation Systems



Neighborhood Based Recommendation System

Neighborhood Based RSs is the most used model in collaborative filtering recommendation systems one of the main crucial components of these models is the similarity measurement between users or items.

Neighborhood Based Recommendation System

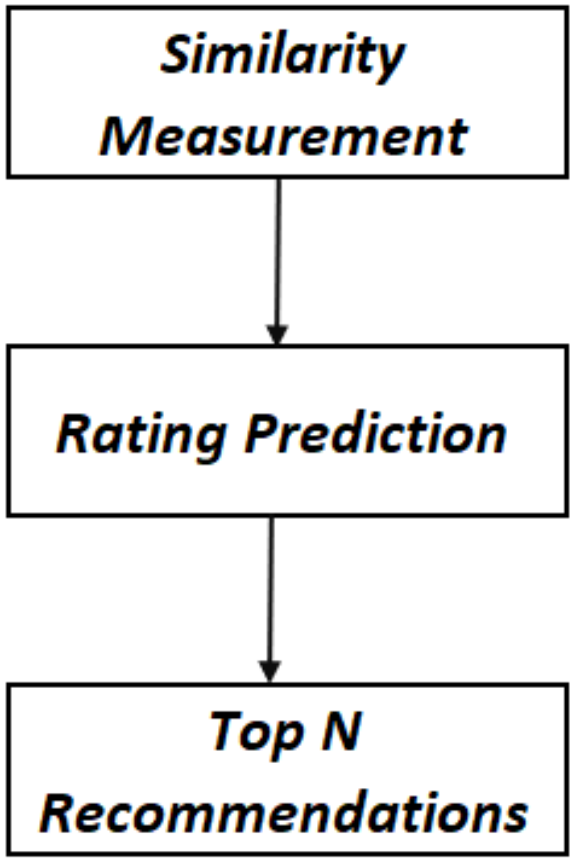


Figure 1: Flow chart Of The Neighborhood Collaborative Filtering

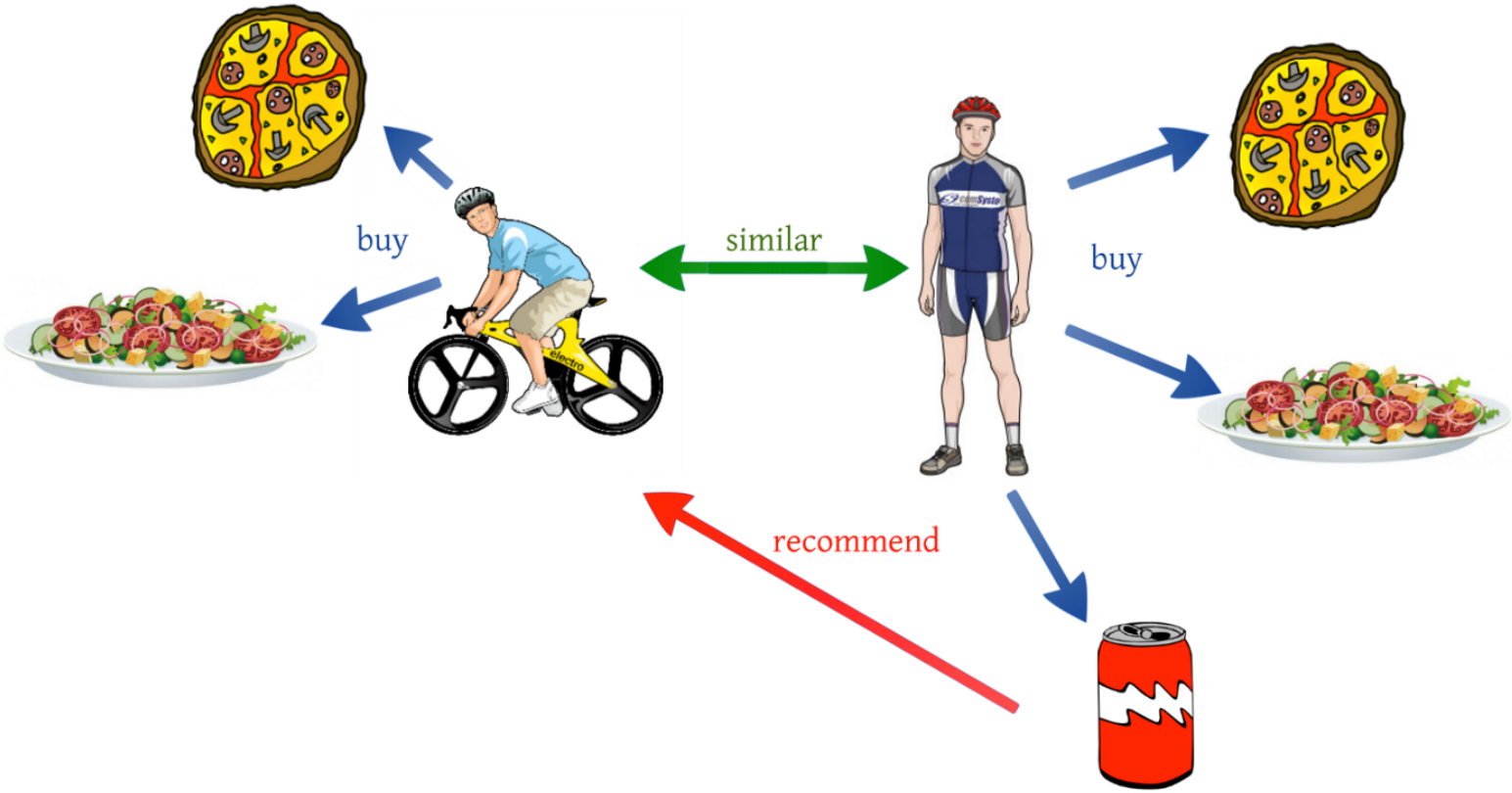
Types Of Neighborhood-Based Collaborative Filtering

- **User-based neighborhood collaborative Filtering**
- **Item-based neighborhood collaborative Filtering**

User-based neighborhood collaborative Filtering

User Neighborhood collaborative Filtering
is based on the main idea that users who have an interest in the same items and similar ratings will thus have similar preferences. In contrast, the recommendations are based on user feedback from similar users (known as neighbors), where they give a similar rating behavior.

User-based neighborhood collaborative Filtering

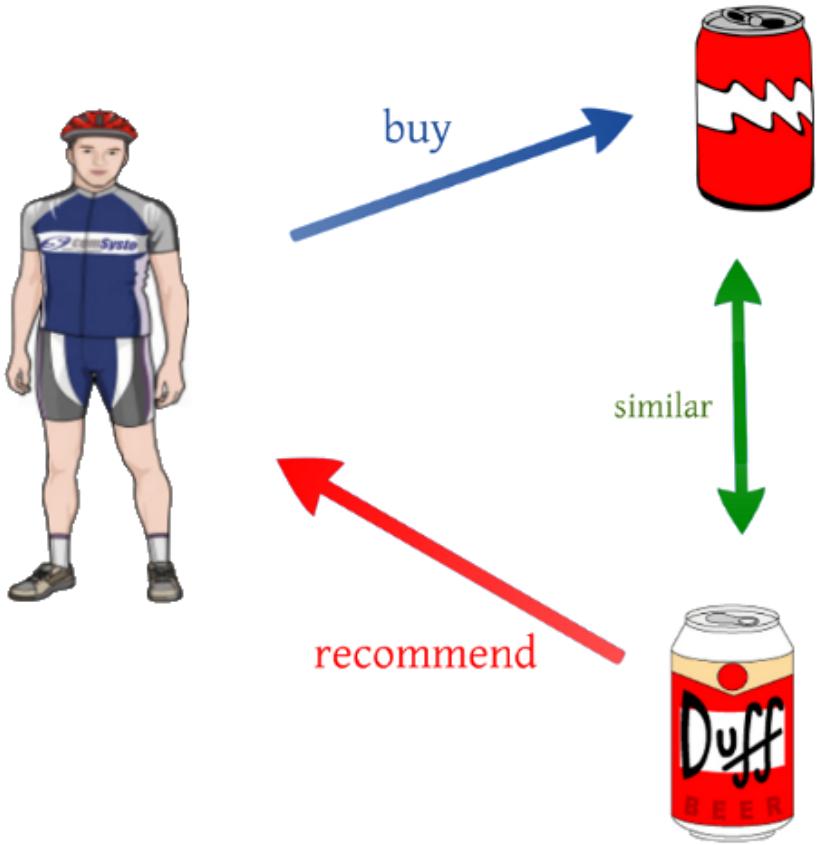


Item-Based Neighborhood Collaborative Filtering

Item-Based Neighborhood Collaborative Filtering

Predicts the user preference based on similarities among the various items and then used them to identify the set of items to be recommended.

Item-Based Neighborhood Collaborative Filtering



Rating Matrix

User/Item	i_1	i_2	...	i_j	...	i_n
u_1	r_{11}	r_{12}	...	r_{1j}	...	r_{1n}
u_2	r_{21}	r_{22}	...	r_{2j}	...	r_{2n}
u_3	r_{31}	r_{32}	...	r_{3j}	...	r_{3n}
.	.	.	...	.	...	.
.	.	.	...	.	...	.
.	.	.	...	.	...	.
u_i	r_{i1}	r_{i2}	...	r_{ij}		r_{in}
.	.	.	...	.	...	.
.	.	.	...	.	...	.
.	.	.	...	.	...	.
u_m	r_{m1}	r_{m2}	....	r_{mj}	...	r_{mn}

Table 1: User-Item Rating Matrix

Similarity Measures in Collaborative Filtering

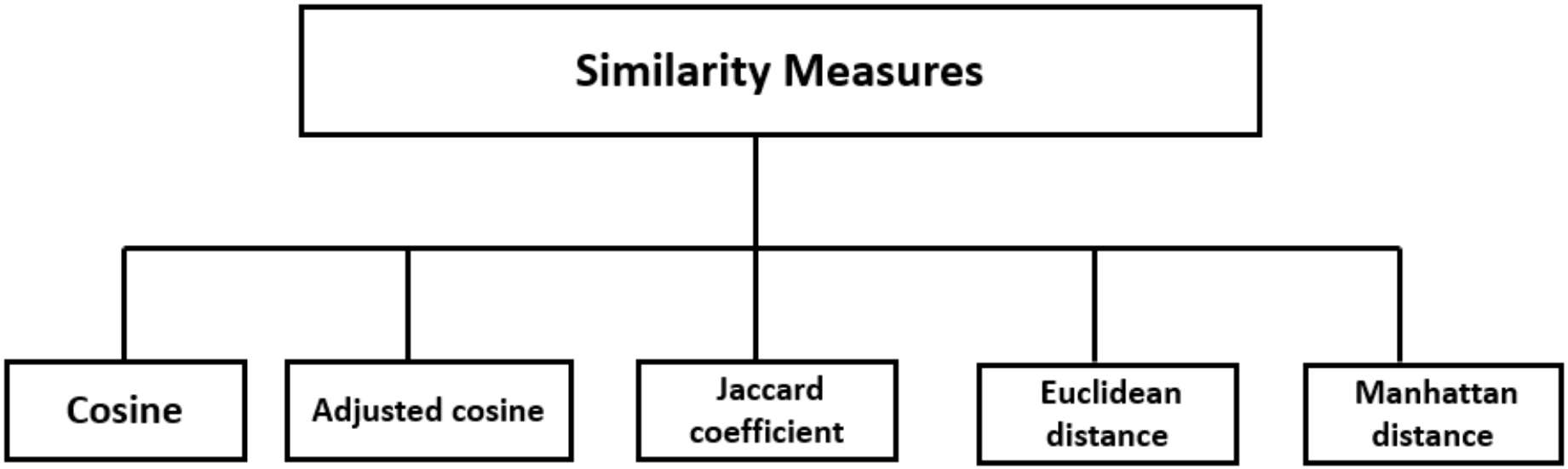


Figure 3: Similarity Measures

Cosine Similarity Measure

Cosine User Similarity

$$cosine(u, v) = \frac{\sum_{i \epsilon I_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{u \epsilon I_u} r_{ui}^2} \sqrt{\sum_{u \epsilon I_v} r_{vi}^2}}$$

Cosine Item Similarity

$$cosine(i, j) = \frac{\sum_{u \epsilon U_{ij}} r_{ui} r_{uj}}{\sqrt{\sum_{u \epsilon U_i} r_{ui}^2} \sqrt{\sum_{u \epsilon U_j} r_{uj}^2}}$$

Adjusted Cosine Similarity Measure

Adjusted Cosine User Similarity

$$ACosine(u,v) = \frac{\sum_{i \epsilon I_{uv}} (r_{ui} - \bar{r}_i)(r_{vi} - \bar{r}_i)}{\sqrt{\sum_{i \epsilon I_{uv}} (r_{ui} - \bar{r}_i)^2} \sqrt{\sum_{i \epsilon I_{uv}} (r_{vi} - \bar{r}_i)^2}}$$

Adjusted Cosine Item Similarity

$$ACosine(i,j) = \frac{\sum_{u \epsilon U_{ij}} (r_{ui} - \bar{r}_u)(r_{uj} - \bar{r}_u)}{\sqrt{\sum_{u \epsilon U_{ij}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{u \epsilon U_{ij}} (r_{uj} - \bar{r}_u)^2}}$$

The Jaccard Coefficient Similarity Measure

The Jaccard Coefficient User Similarity

$$J(u,v) = \frac{|u \cap v|}{|u \cup v|}$$

The Jaccard Coefficient Item Similarity

$$J(i,j) = \frac{|i \cap j|}{|i \cup j|}$$

Euclidean Distance Similarity Measure

Euclidean Distance User Similarity

$$d(u,v) = \sqrt{\sum_{i \epsilon I_{uv}} (r_{vi} - r_{ui})^2}$$

Euclidean Distance Item Similarity

$$d(i,j) = \sqrt{\sum_{u \epsilon U_{IJ}} (r_{uj} - r_{ui})^2}$$

Manhattan Distance Similarity Measure

Manhattan Distance User Similarity

$$d_1(u, v) = \sum_{i \epsilon I_{uv}} (|r_{vi} - r_{ui}|)$$

Manhattan Distance Item Similarity

$$d_1(i, j) = \sum_{u \epsilon U_{ij}} (|r_{uj} - r_{ui}|)$$

Rating Prediction

User-based Rating Prediction

$$\widehat{r_{ui}} = \frac{\sum_{v \epsilon N_i(u)} S_{uv} r_{vi}}{\sum_{v \epsilon N_i(u)} |S_{uv}|}$$

Item-based Rating Prediction

$$\widehat{r_{ui}} = \frac{\sum_{j \epsilon N_u(i)} S_{ij} r_{uj}}{\sum_{j \epsilon N_u(i)} |S_{ij}|}$$

Rating Normalization

- **Mean Centering**
  User-based prediction of a rating Rating r_{ui} is transformation to a mean-centered one $h(r_{ui})$ by $h(r_{ui}) = r_{ui} - \bar{r}_u$

$$\widehat{r_{ui}} = \bar{r}_u + \frac{\sum_{v \epsilon N_i(u)} s_{uv}(r_{vi} - \bar{r}_v)}{\sum_{v \epsilon N_i(u)} |s_{uv}|}$$

  Item-based prediction of a rating Rating r_{ui} is transformation to a mean-centered one $h(r_{ui})$ by $h(r_{ui}) = r_{ui} - \bar{r}_i$

$$\widehat{r_{ui}} = \bar{r}_i + \frac{\sum_{j \epsilon N_u(i)} s_{ij}(r_{uj} - \bar{r}_j)}{\sum_{j \epsilon N_u(i)} |s_{ij}|}$$

Rating Normalization

- **Z-score normalization**
  User-based prediction of a rating the normalization of a rating r_{ui} divides the user-mean-centered rating by the standard deviation σ_u of the ratings given by user u , $h(r_{ui}) = \frac{r_{ui} - \bar{r}_u}{\sigma_u}$.

$$\widehat{r_{ui}} = \bar{r}_u + \sigma_u \frac{\sum_{v \epsilon N_i(u)} s_{uv}(r_{vi} - \bar{r}_v)/\sigma_v}{\sum_{v \epsilon N_i(u)} |s_{uv}|}$$

  Item-based prediction of a rating The same for item $h(r_{ui}) = \frac{r_{ui} - \bar{r}_i}{\sigma_i}$.

$$\widehat{r_{ui}} = \bar{r}_i + \sigma_i \frac{\sum_{j \epsilon N_u(i)} s_{ij}(r_{uj} - \bar{r}_j)/\sigma_j}{\sum_{j \epsilon N_u(i)} |s_{ij}|}$$

Neighborhood Selection

- Top-N Recommendation
  - User-Based Top-N Recommendation
  - Item-Based Top-N Recommendation
- Threshold Filtering
- Negative Filtering

Dimensionality Reduction

The dimension reduction phase transforms the original $R_{n\times m}$ user-item matrix into a lower-dimensional space.

Matrix Factorization Techniques

Matrix Factorization is a subset of dimensionality reduction in which the user-item matrix is reduced to the product of many low-rank matrices.

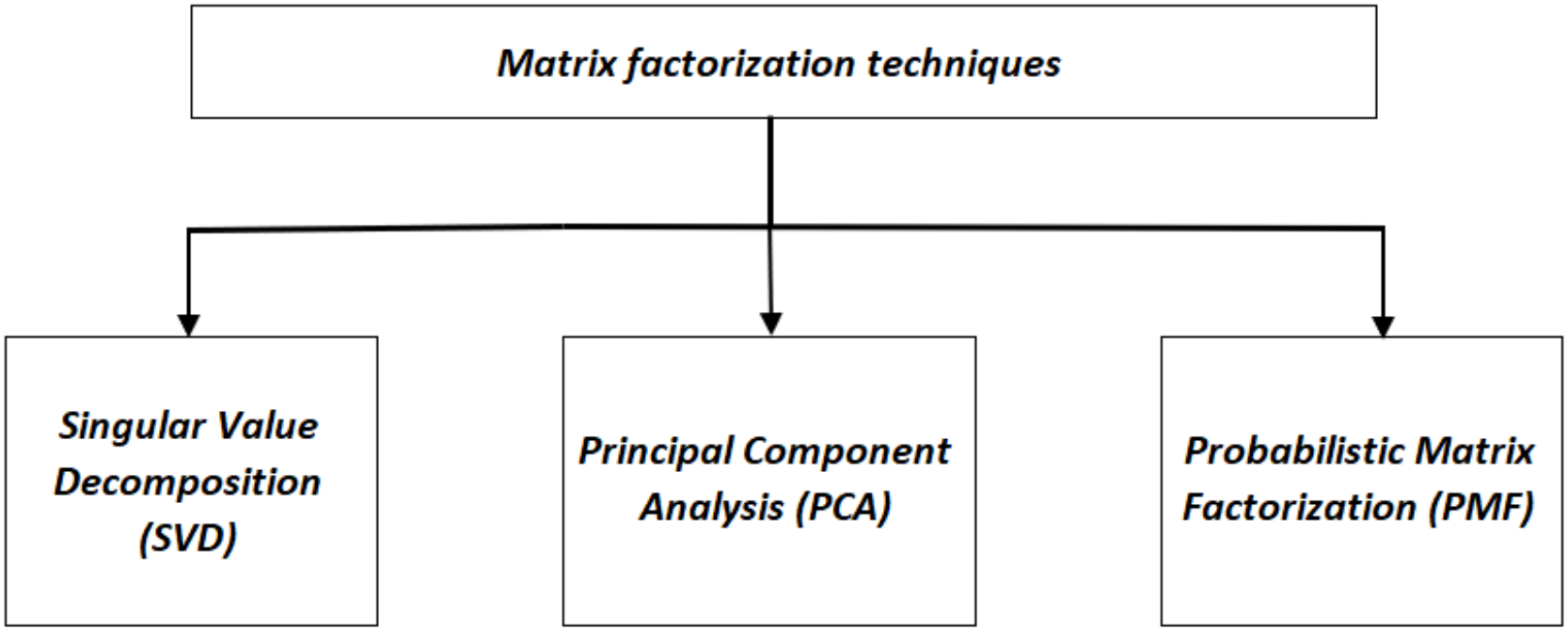


Figure 4: Matrix Factorization Techniques

Distance Covariance

The distance covariance
is a new effective measure of dependence between multivariate random vectors based on the assumption that the data are i.i.d,Aim to detect arbitrary types of non-linear associations between sets of variables, say X and Y, not necessarily equal in dimensions.

Distance Covariance

$$V^2(X,Y) = \int_{R^{p+q}} |\phi_{(X,Y)}(t,s) - \phi_X(t)\phi_Y(s)|^2 \omega(t,s)\, dt\, ds$$

Denoted by :

$$\phi_{X,Y}(t,s) = E[\exp^{i(t'X+s'Y)}], (t,s)\epsilon R^{p+q}$$

$$\phi_X(t) = E[\exp(it'X)]$$

$$\phi_Y(s) = E[\exp(is'Y)]$$

Estimation of distance covariance

The empirical distance covariance measures is function of the double centred distance matrices of (X_i, Y_i) , $i = 1, ..., n$

$$\hat{V}^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}$$

Where :

$$(a_{ij}) = (|X_i - X_j|_p)$$

$$(b_{ij}) = (|Y_i - Y_j|_q)$$

$$A_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$

$$B_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}$$

Estimation of distance covariance

$$\hat{V}^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}b_{ij} + \frac{1}{n^4}\sum_{i,j=1}^{n} a_{ij}\sum_{i,j=1}^{n} b_{ij} - \frac{2}{n^3}\sum_{i,j,k=1}^{n} a_{ij}b_{jk}$$

Distance covariance in User-Based Neighborhood

The empirical distance covariance for User-based neighborhood

$$\hat{V}^2(r_u, r_v) = \frac{1}{m^2}\sum_{i,j=1}^{m} a_{ij}b_{ij} + \frac{1}{m^4}\sum_{i,j=1}^{n} a_{ij}\sum_{i,j=1}^{m} b_{ij} - \frac{2}{m^3}\sum_{i,j,k=1}^{m} a_{ij}b_{ik}$$

(r_{ui}, r_{vi}) , $u = 1, ..., m$ and $v = 1, ..., m$ is a random sample from the joint distribution of the random vectors r_u and r_v.

$$(a_{ij}) = (|r_{ui} - r_{uj}|_p)$$

$$(b_{ij}) = (|r_{vi} - r_{vj}|_q)$$

Distance covariance in Item-Based Neighborhood

The empirical distance covariance for Item-based neighborhood

$$\hat{V}^2(r_i, r_j) = \frac{1}{n^2}\sum_{u,v=1}^{n} a_{uv}b_{uv} + \frac{1}{n^4}\sum_{u,v=1}^{n} a_{uv}\sum_{u,v=1}^{n} b_{uv} - \frac{2}{n^3}\sum_{u,v,k=1}^{n} a_{uv}b_{uk}$$

(r_{ui}, r_{uj}) , $i = 1, ..., n$ and $j = 1, ..., n$ is a random sample from the joint distribution of the random vectors r_i and r_j.

$$(a_{uv}) = (|r_{ui} - r_{vi}|_p)$$

$$(b_{uv}) = (|r_{uj} - r_{vj}|_q)$$

Implementation

YASSIR Express

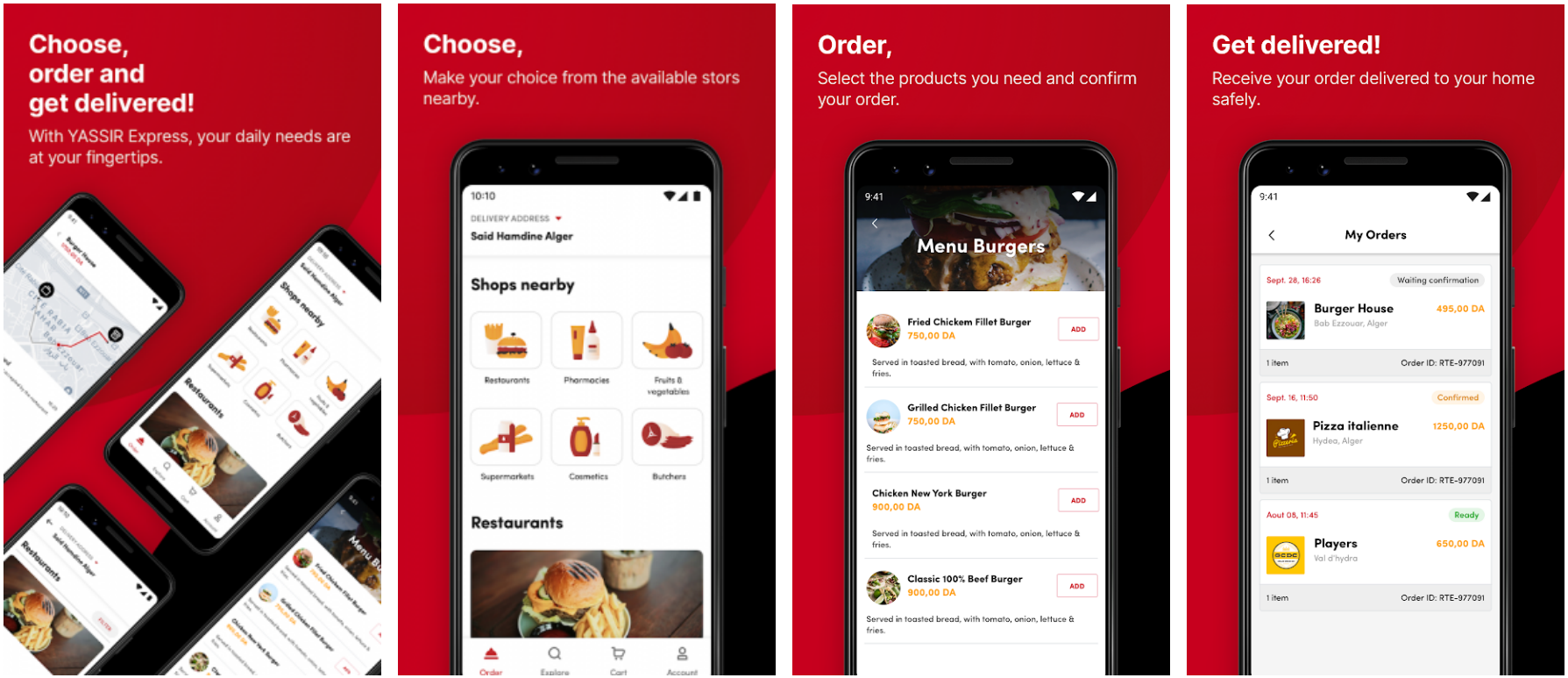


Figure 5: YASSIR EXPRESS Application

Figure 6: YASSIR EXPRESS Restaurants

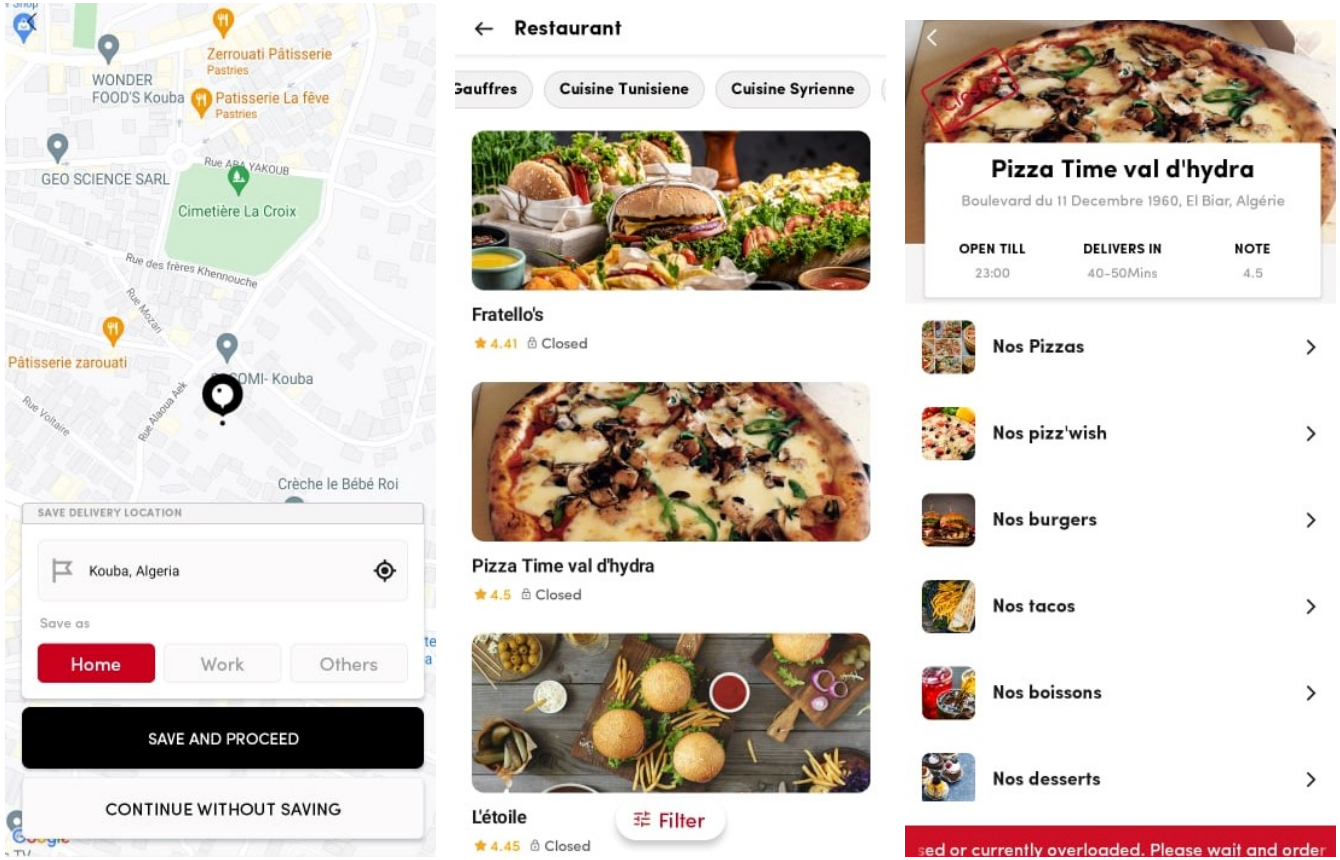


Figure 7: YASSIR EXPRESS User Experience

YASSIR EXPRESS Dataset

User ID

Order ID

Restaurant ID

Request
Delivery
Pick Up date

YASSIR
Express
Dataset

Restaurant Location

Frameworks

- **Google Colab** allows anybody to write and execute arbitrary python code through the browser.
- **Python** is object oriented and functional programming language.

Frameworks

User Rating

User Rating

- **The First implicit Rating Approach**

  User Rating 1 = delivery date - request date

- **The Second implicit Rating Approach**

  User Rating 2 =

$$\frac{\text{Mean Of The Delivery Duration}}{\text{Number of Times The User Has Been Purchased From The Same Restaurant}}$$

Top Recommendation Of User-Based Using Euclidean Distance Similarity Metric

```
user2userRecommendation_euclidean(244)
```

	userid	itemid	predicted_rating
0	244	477	2.725000
234	244	206	1.925000
445	244	358	1.553571
2666	244	625	1.375000
2677	244	309	1.325000

Evaluation Results of The Similarity Measures using User-Based

	Similarity Measure	RMSE	MAE	MSE
User-Based	Euclidean Distance	*0.90187*	*0.64881*	*0.81338*
	Manhattan Distance	*0.89382*	*0.64353*	*0.79891*
	Cosine	*0.93452*	*0.69218*	*0.87333*

Evaluation Results of The Similarity Measures using User-Based

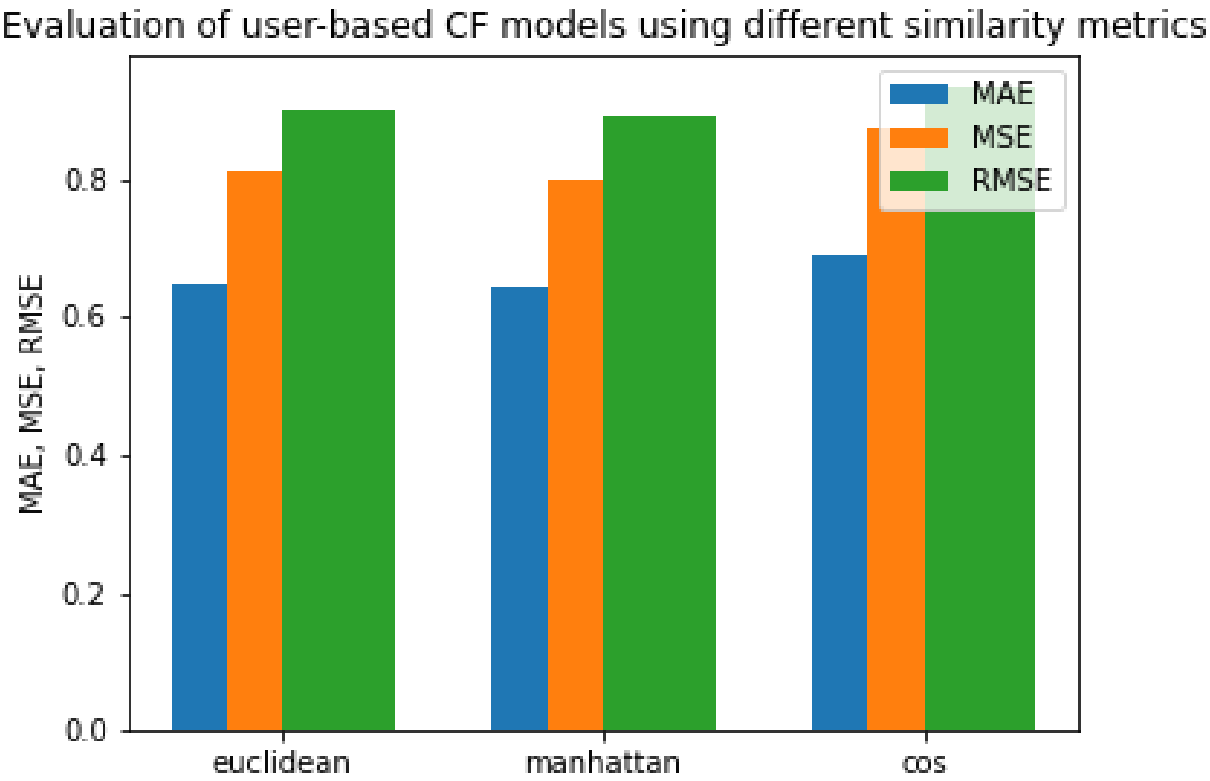


Figure 8: The Evaluation Results Of The Similarity Measures Using User-Based

Top Recommendation Of Item-Based Using Cosine Similarity Metric

	itemid	similarity_with_Iu_cos	prediction
27	516	0.997280	3.000000
10	530	0.999143	3.000000
14	431	0.998490	3.000000
28	244	0.997224	3.000000
30	211	0.997158	3.000000

Evaluation Results of The Similarity Measures using Item-Based

	Similarity Measure	RMSE	MAE	MSE
Item-Based	Euclidean Distance	*0.93329*	*0.676635*	*0.87103*
	Manhattan Distance	*0.93345*	*0.67690*	*0.87133*
	Cosine	*0.94761*	*0.67799*	*0.89796*
	Adjusted Cosine	2.40673	1.67612	5.79237
	Covariance Distance	3.14617	2.25624	9.89840

Evaluation Results of The Similarity Measures using Item-Based

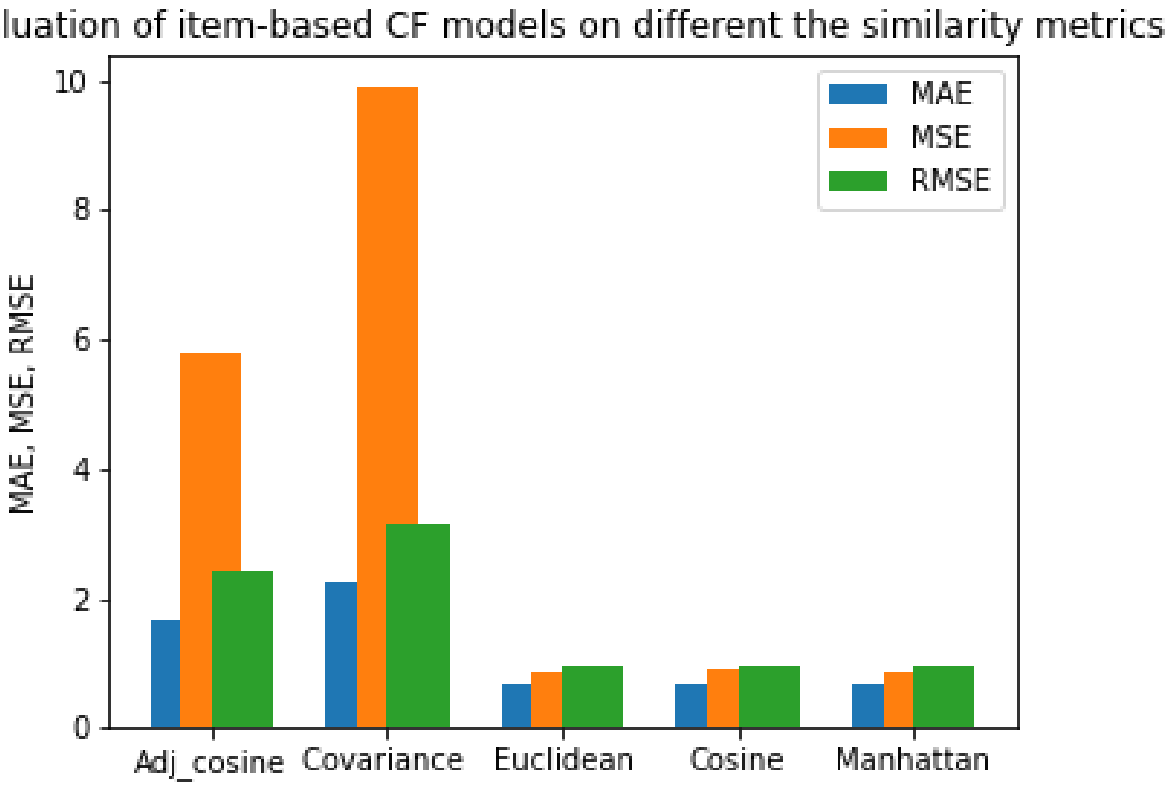


Figure 9: The Evaluation Results Of The Similarity Measures Using Item-Based

Results of the Similarity Measures on User-Based and Item-based using the First Rating Approach

	Similarity Measure	RMSE	MAE	MSE
User-Based	Euclidean Distance	*0.90187*	*0.64881*	*0.81338*
	Manhattan Distance	*0.89382*	*0.64353*	*0.79891*
	Cosine	*0.93452*	*0.69218*	*0.87333*
Item-Based	Euclidean Distance	*0.93329*	*0.676635*	*0.87103*
	Manhattan Distance	*0.93345*	*0.67690*	*0.87133*
	Cosine	*0.94761*	*0.67799*	*0.89796*
	Adjusted Cosine	2.40673	1.67612	5.79237
	Covariance Distance	3.14617	2.25624	9.89840

Results of the Similarity Measures on User-Based and Item-based using the Second Rating Approach

	Similarity Measure	RMSE	MAE	MSE
User-Based	Euclidean Distance	*3.10530*	*2.77726*	*3.22170*
	Manhattan Distance	*3.09143*	*2.76914*	*3.19123*
	Cosine	*3.23437*	*2.92977*	*3.52367*
Item-Based	Euclidean Distance	*3.33196*	*2.96546*	*3.77413*
	Manhattan Distance	*3.33244*	*2.96580*	*3.77541*
	Cosine	*3.2835*	*2.84632*	*1.64744*
	Adjusted Cosine	5.18968	4.72174	12.17406
	Covariance Distance	5.82345	5.25817	16.61884

Conclusion

- The distance covariance new similarity metric has been the worst in terms of performance and effectiveness on the item-based neighborhood using the first and the second rating approach.
- The user-based has the best performance using the manhattan similarity metric,meanwhile the lowest was while using the cosine similarity metric.
- The item-based delivers the best performance results using cosine similarity metric,on the other hand the performance of the item based model was the lowest using the covariance distance similarity metric.
- User-based provides results better than those offered by the item-based for all similarity metrics in terms of their performance.
- The performance of the first rating approach is way better than the second rating approach.
- The performance of the Recommendation Systems is strongly dependent on the used similarity metric.

Questions

Thank you