Synechron

AI – Synechron Serbia Way

What We Do?

Data Science
Providing strategy, planning and leadership to accelerate end-to-end digital transformation.

Computer Vision
Creating user experiences that people love, by exceeding expectations.

Natural Language Processing
Accelerating Cloud & DevOps adoption to increase the speed of delivery.

Data Engineering
Helping enterprises move to modern data solutions and deliver reliable analytics at scale.

Big Data systems
Elevating strategic partnerships and fuelling overall growth while adding value to clients.

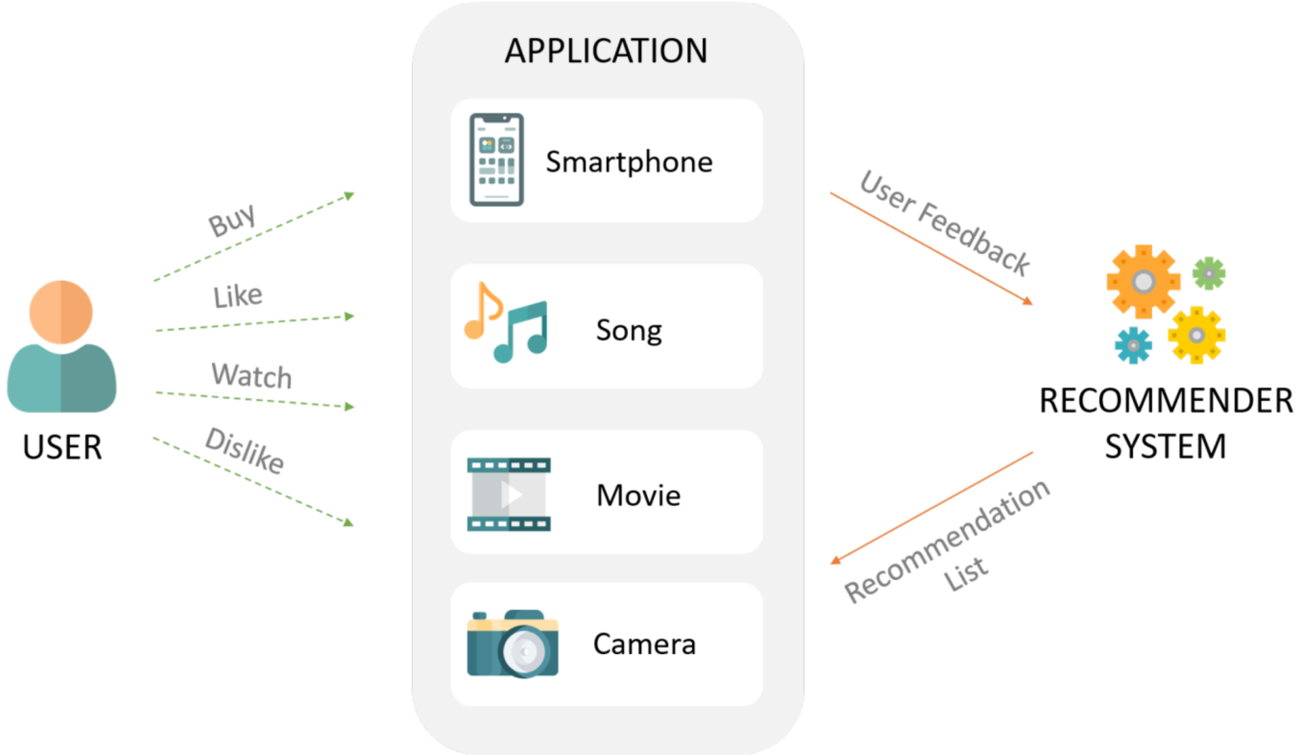
Event-driven systems
Assisting clients to create change that matters – transformation, enabled by technology and sustained through capabilities.

Synechron

Recommender systems

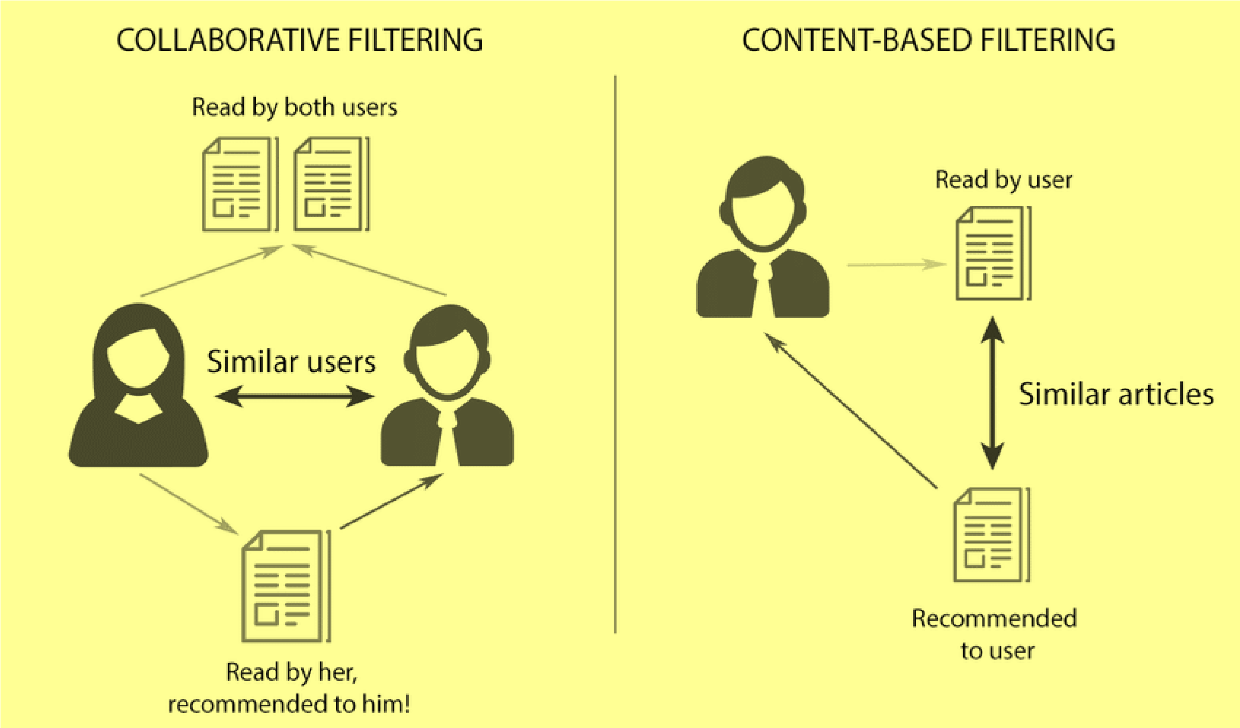
Introduction

Recommender systems are algorithms used to suggest items to users based on their preferences. They are used in various applications such as online shopping, social media, and streaming services. Recommender systems can be divided into two main approaches: **content-based** and **collaborative filtering**. Both approaches have their own strengths and weaknesses and can be used in combination to optimize user experience.



Introduction

Major types are: **content-based filtering, collaborative filtering** and **hybrid filtering** systems. Some other types are popularity-based, demographic-based, knowledge-based systems.



Content-Based Filtering

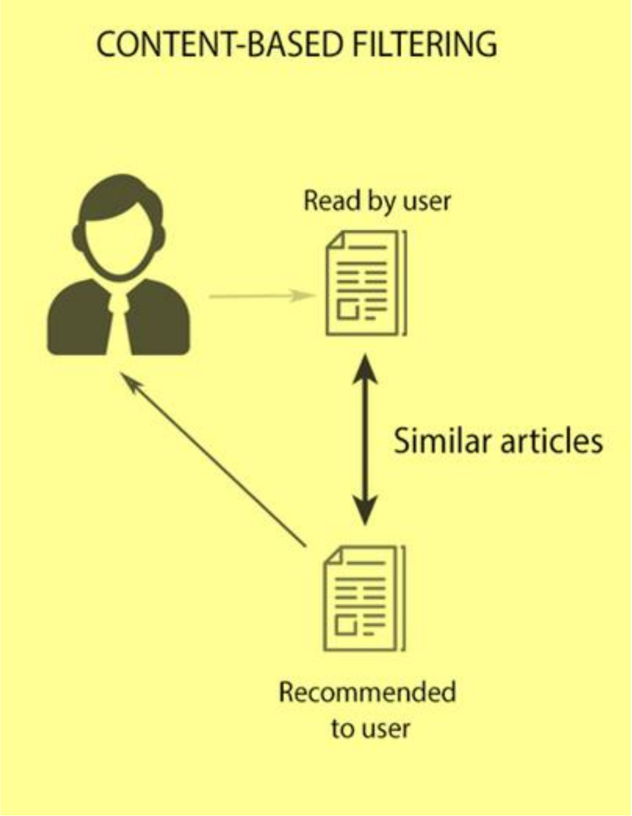
Content-based systems recommend based on the description of an item, and a profile of the user's preferred choices. They ignore contributions from other users as with the case of collaborative techniques.

Pros:

- Easy to scale to a large number of users
- The model can recommend niche items that very few other users are interested in (Solves Grey sheep problem)

Cons:

- Need to have an in-depth knowledge and description of the features of the items in the profile.



Content-Based Filtering

Some of the popular approaches used in content-based filtering are:

1. Similarity approach
2. Classification approach
3. Pattern mining and matching

Similarity metric approach

01

Similarity metric approach for recommendation

- Similarity approaches are commonly used in recommender systems to identify users that are similar to other users, or items that are similar to other items.
- The idea in this approach is that similar users might have similar interests, as well as similar items can be recommended together.
- One advantage of similarity approaches is that they can provide accurate recommendations even when the system has limited information about the user′s preferences.
- Other advantage is their low complexity and high execution speed.

Similarity metric approach for recommendation

```
{
    "name": "Pera Peric",
    "age": 21,
    "country": "Serbia",
    "gender": "M"
    "movies": ["Titanic", "Avatar"]
}
```

Vectorization

User vector

Pera Peric	21	Serbia	M

Find similar vectors/users

Database of all other users

Mika Mikic	25	Serbia	M
Ana Anic	19	Croatia	F
John Doe	32	United States	M

Recommendation phase
Copy movies from TOP N most similar users

```
{
    "name": "Titanic",
    "genre": ["drama", "action"],
    "country": "United States",
    "description": "............"
    "actors": ["Actor 1", "Actor 2"]
}
```

Vectorization

Movie vector

Titanic	drama, action	…	Actor1, Actor2

Find similar vectors/movies

Database of all other users

Avatar	sci/-fi, action	…	Actor5, Actor9
Justice League	sci/-fi, action	…	Actor45, Actor 8
Wonderful Life	drama, document	…	Actor156, Actor 1

Recommendation phase
Recommend TOP N most similar movies

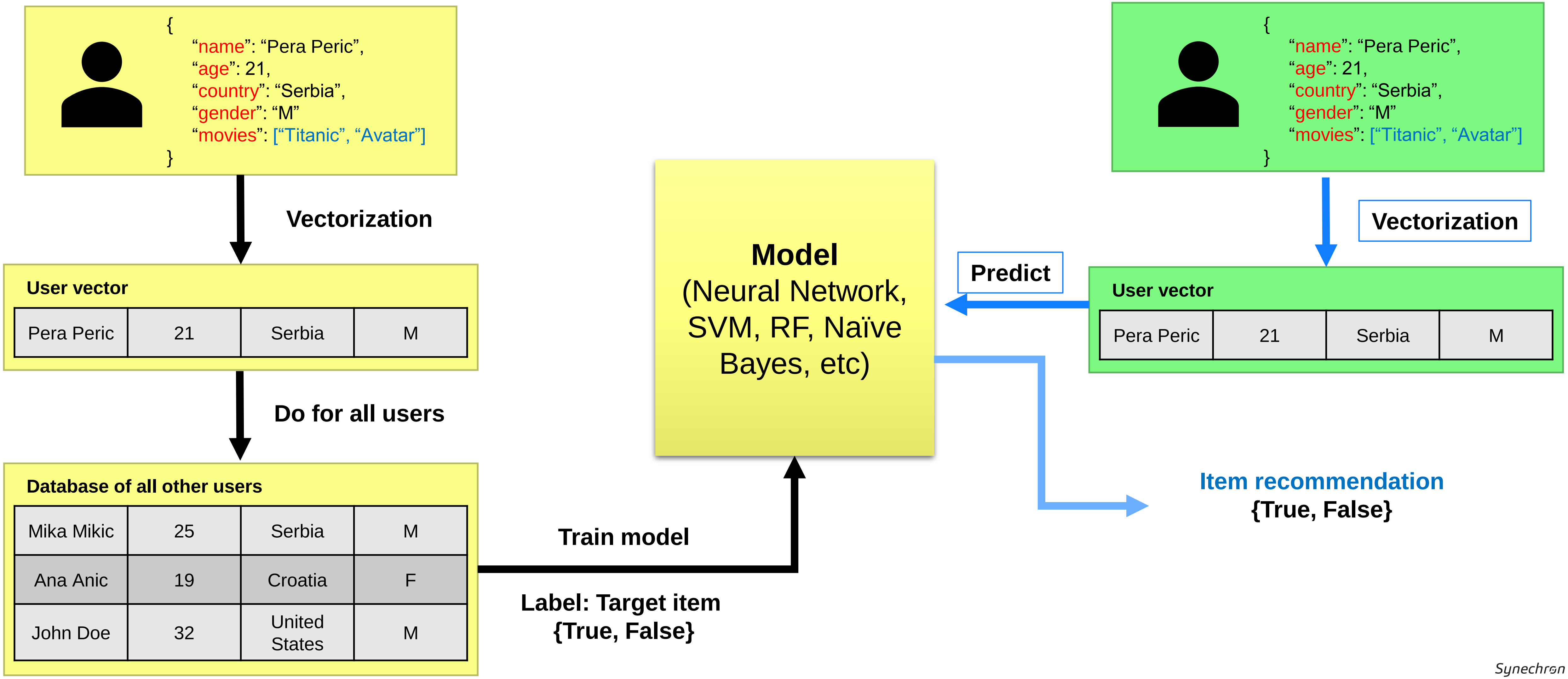
Classification approach

02

Classification models for recommendation

- Classification models can be used in recommender systems to predict whether a user will like or dislike an item based on their past behavior and interactions with the system.
- One common approach is to use a binary classification model, where the target variable is whether the user will like an item or not.
- Once the classification model is trained, it can be used to generate recommendations for users. For example, the model can be used to predict the probability that a user will like a new item, and the items with the highest predicted probabilities can be recommended to the user.
- Classification models can be a useful tool in recommender systems, particularly when the system has a large number of items and users, and when user-item interactions are sparse. However, it is important to ensure that the model is trained on high-quality data and that the features used for the model are relevant to the user's preferences and characteristics.

Classification models for recommendation

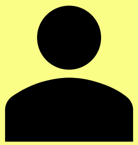


Pattern mining and matching

03

Pattern mining and matching

1. Frequent pattern mining can be used in order to get rules from the set of transactions/dataset rows, and represent them as implications, for example:
    - *{country=Serbia, age=21} => {Avatar}, confidence 95%*
    - *{age=35} => {Wonderful Life}, confidence 73%*
  - Algorithms used for Frequent pattern mining are Apriori and FP-Growth

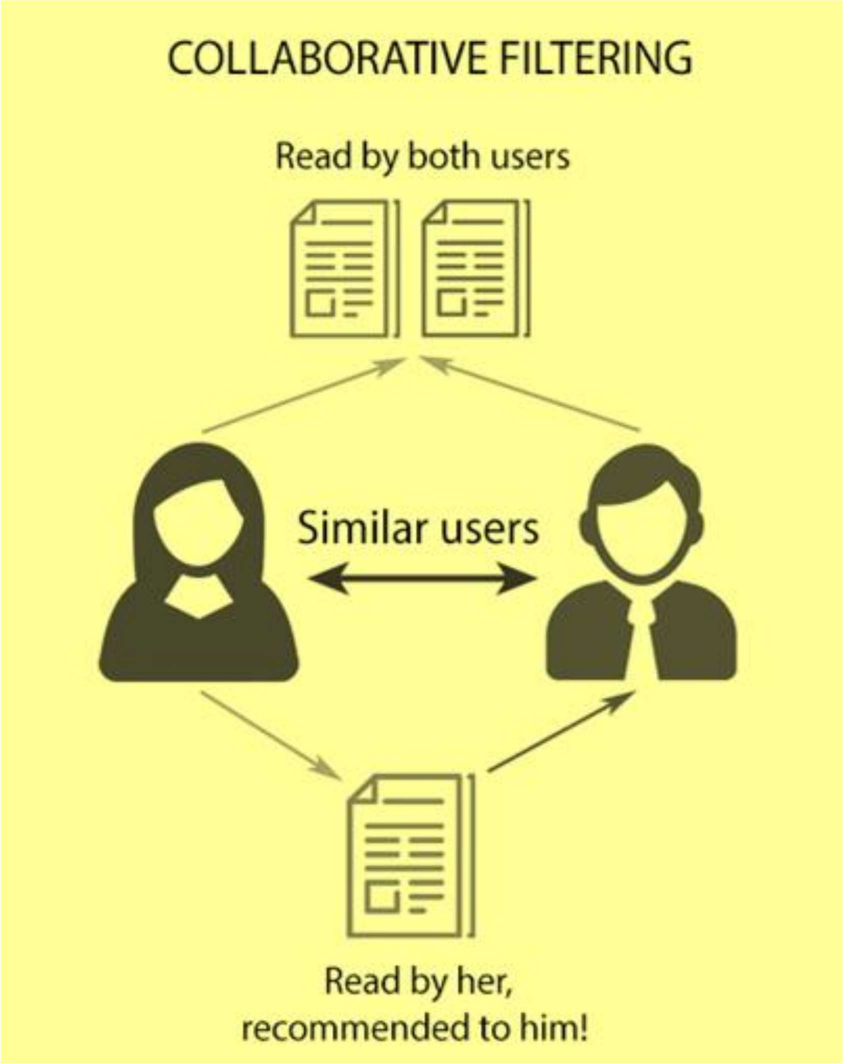



```
{
    “name”: “Pera Peric”,
    “age”: 21,
    “country”: “Serbia”,
    “gender”: “M”
    “movies”: [“Titanic”, “Avatar”]
}
```


2. Correlation analysis
  - Performing Exploratory data analysis and correlation patterns inside data in order to find frequent patterns

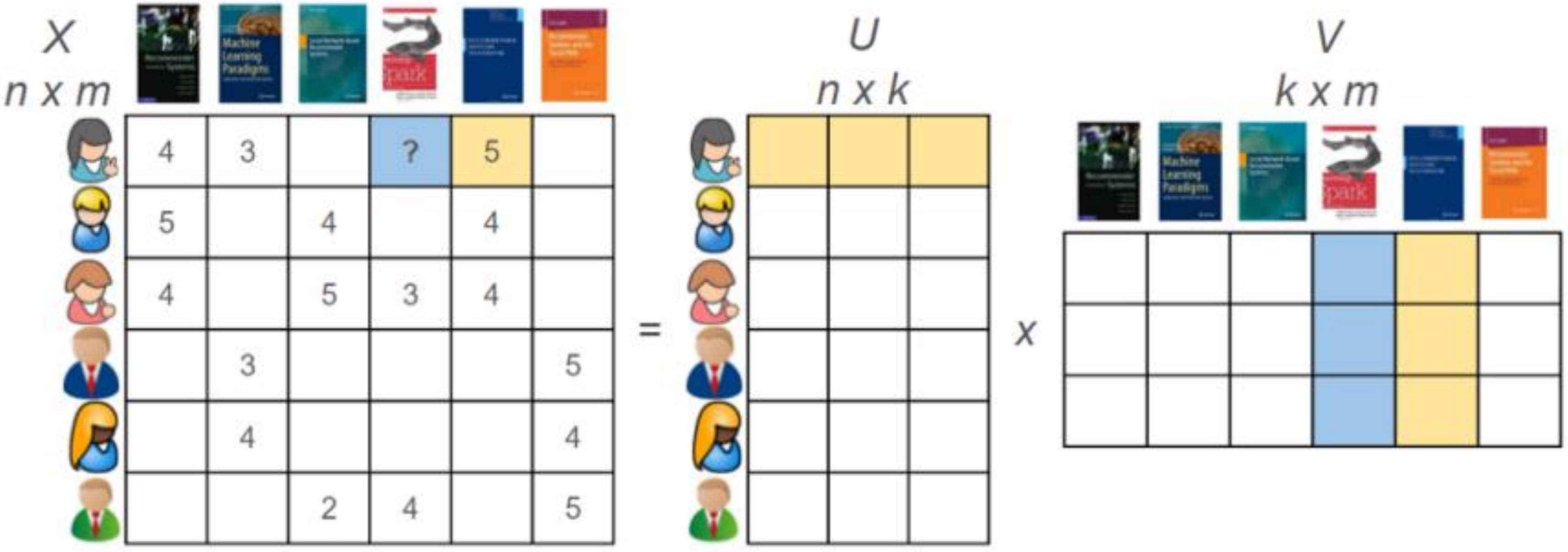
Collaborative Filtering

Collaborative filtering recommends items by identifying other users with similar taste; it uses their opinion to recommend items to the active user. A user gets recommendations to those items that he has not rated before but that were already positively rated by users similar to him. Based on the assumption that users who agreed in the past will agree in the future.



Collaborative Filtering

Most popular algorithms are SVD (Singular Value Decomposition) and NMF(Non-negative Matrix Factorization)



Collaborative Filtering

Potential Problems:

- Cold Start - it concerns the issue that the system cannot draw any inferences for users or items about which it has not yet gathered sufficient information.

Pros:

- No domain knowledge necessary because the embeddings are automatically learned.
- The model can help the user discover new interests. The ML system may not know the user is interested in a given item, but the model might still recommend it because similar users are interested in that item.

Cons:

- Hard to include side features for query/item. Side features are any features beyond the query or item ID

Problem statement

Opis domena problema

Da bi osiguravajuća kuća završila izdavanje polise osiguranja, postoji niz deklaracija i niz unapred definisanih obrazaca ili aneksa (engl. forms) koje je potrebno priložiti uz samu polisu. Ove forme/aneksi moraju pokrivati glavne odredbe politike polise (engl. policy provisions), rasporede, odobrenja, isključenja i druge aspekte politike same polise koju klijent uzima.

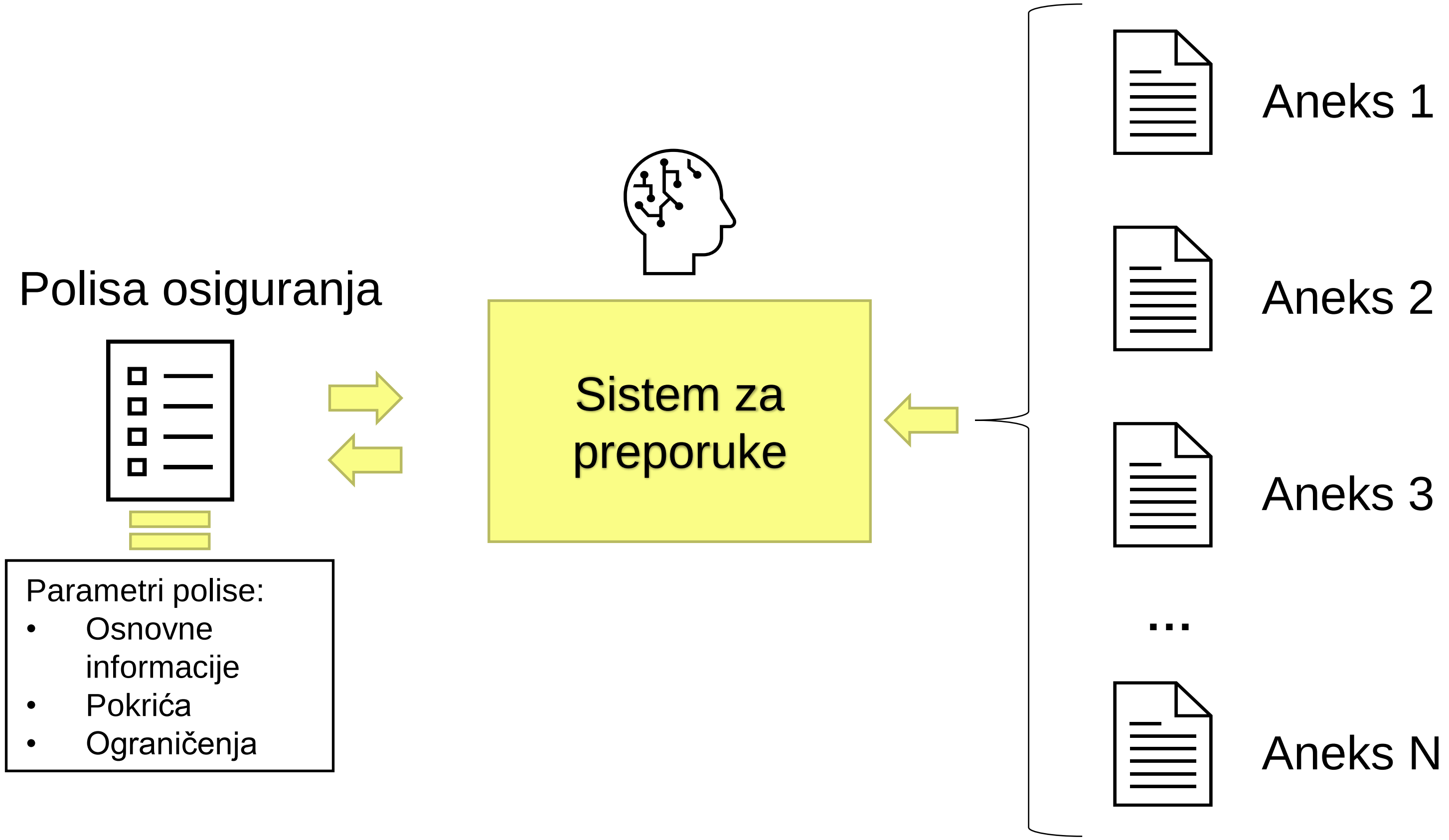
Polisa osiguranja (u realnom svetu je to skup dokumenata)

Glavni dokumenti polise (osnovne informacije o osiguraniku, lista pokrića, lista ograničenja na pokrića, detaljima premije i slično)

Skup aneksa. Oni opisuju uslove pod kojima se primenjuju sve stavke u glavnim dokumentima (pod kojim uslovima važe pokrića, ograničenja, privatnost podataka i slično).

Opis domena problema

Da bi osiguravajuća kuća završila izdavanje polise osiguranja, postoji niz deklaracija i niz unapred definisanih obrazaca ili aneksa (engl. forms) koje je potrebno priložiti uz samu polisu. Ove forme/aneksi moraju pokrivati glavne odredbe politike polise (engl. policy provisions), rasporede, odobrenja, isključenja i druge aspekte politike same polise koju klijent uzima.



Opis skupa podataka

1 Žuta polja

Žute kolone sadrže osnovne podatke iz polise koje se nalaze u svakoj polisi. Tu se nalaze osnovne informacije o kompaniji koja uzima osiguranje, o njenoj šifri delatnosti, osnovnim geografskim informacijama i slično.

2 Plava polja

Plave kolone se odnose na informacije koje su specifične za tu pojedinačnu polisu osiguranja. U ovim kolonama se nalaze informacije o tome koja pokrića je klijent uzeo, pod kojim uslovima, da li postoje limiti (ograničenja) i slično. Ovi atributi utiču na neke anekse i definišu opšte informacije i uslove pokrića, ograničenja i slično koji se nalaze u tim aneksima.

3 Siva polja

Aneksi. Kada je aneks označen sa 1, to znači da je on bio prikačen uz tu pojedinačnu polisu osiguranja. Ako je označen sa 0, znači da je taj aneks bio izostavljen iz date polise. Ovo su naše target varijable (labele).

Zadatak

Ideja je da sistem čoveku preporuči najverovatnije anekse koje bi trebalo prikačiti uz datu polisu, da čovek ne bi morao prolaziti kroz sve anekse i ručno odlučivati koji prikačiti a koji ne. U prevodu, nije važno ako sistem preporuči više aneksa nego što bi trebalo, jer konačnu odluku ipak donosi čovek, ali ne bi smeo izostaviti nešto što bi trebalo prikačiti.

U realnom svetu, osiguravajuće kuće često definišu skup pravila (engl. rules) pod kojim se aneksi priključuju uz polisu osiguranja (pravila tipa implikacije, recimo “država=TX ∧ šifra_detalnosti=12345 => Aneks1”) i ova pravila zavise od zakona u toj državi i direktno ih prate. Međutim problem koji mi rešavamo sadrži više od 100 mogućih aneksa pa je ručno pisanje ovakvih pravila nezgodno i komplikovano, a nije ni održivo jer se zakoni često menjaju.

Napraviti sistem za preporuke koji će biti sposoban da modeluje preporučivanje aneksa za zadatu polisu osiguranja na ulazu, korišćenjem tehnika mašinskog učenja, rovarenja po podacima (engl. data mining) i/ili statističkih mera.

Kriterijum ocenjivanja

1. Kreativnost u samoj izradi rešenja (u modelima, pripremi podataka i slično).

2. Inovativnost i istraživački pristup u rešavanju problema.

3. Recall i precision kao mere performansi modela. Pošto skup podataka nije veliki i teško je obučiti model koji će raditi jako tačno, ovaj kriterijum možete posmatrati samo kao dodatni indikator da li ste izabrali dobar model na osnovu hipoteze koju ste postavili.

4. Objašnjivost modela (engl. model explainability)