Content-based Filtering Recommendation

Jiho Kang
2022.03.26

1. Content-based Filtering

- Content-based filtering uses **item features** to recommend other items **similar** to what the user likes, based on their previous actions or explicit feedback.
- **Item features**
  - TF-IDF, Word2Vec(CBOW, Skip-gram)
- **Similarity**
  - Cosine, Euclidean, Manhattan, Jasscard
- **Pros**
  - Doesn’t need other users’ data, only one user and features of item
  - Cold start problem solved - new item
  - Using item features makes explainable recommendation
- **Cons**
  - Require domain knowledge to extract item features properly
  - Cannot recommend items of new genre
  - Cold start problem - new user

2. Item Features

- **TF-IDF**

  features = single word
  item = document

  - How often a specific word appears in a particular document.
  - TF: Frequency of a feature within a particular item.
  - DF: Frequency of items in which a particular feature appears.
  - IDF: Inverse of DF => to give a penalty

- **Word2Vec => Item2Vec**

  {I like this movie. I love this movie. It was the best movie I've ever seen.}
  {I don't like this movie. This is the worst movie I've ever seen.}

  - **CBOW**
    The model predicts the current word from a window of surrounding context words.

  - **Skip-gram**
    The model uses the current word to predict the surrounding window of context words.

3. Similarity

- **Cosine**
  - Angle
  - Regardless of the scale

$$\text{similarity}\,(x,y) = \cos\,(\theta) = \frac{x \cdot y}{|x||y|}$$

- **Euclidean**
  - Linear distance (L2 norm)
  - Consider the scale

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

- **Manhattan**
  - Coordinates distance (L1 norm)
  - Not widely used in the field

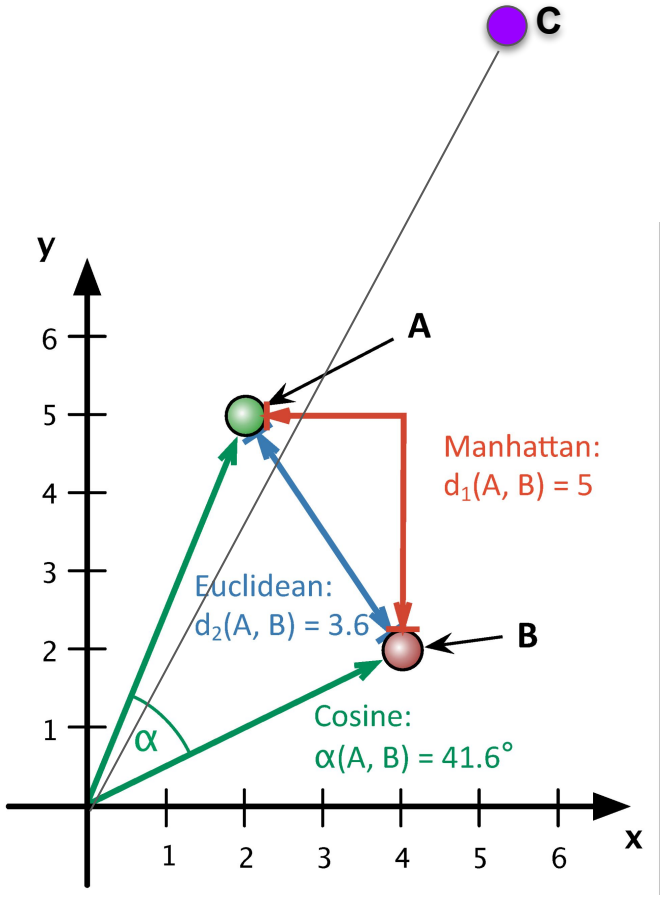
$$|x_1 - x_2| + |y_1 - y_2|$$

- **Jasscard**
  - Ration of intersection over union.
  - Binary feedback

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

3. Similarity

- **Cosine**
  - Angle
  - Regardless of the scale
- **Euclidean**
  - Linear distance (L2 norm)
  - Consider the scale
- **Manhattan**
  - Coordinates distance (L1 norm)
  - Not widely used in the field
- **Jasscard**
  - Intersection over union.
  - Binary feedback, sparse



Content-based Filtering

- Content-based filtering uses **item features** to recommend other items **similar** to what the user likes, based on their previous actions or explicit feedback
- **Item features**
  - TF-IDF, Word2Vec(CBOW, Skip-gram)
- **Similarity**
  - Cosine, Euclidean, Manhattan, Jasscard
- **Pros**
  - Doesn’t need other users’ data, only one user and features of item
  - Cold start problem solved - new item
  - Using item features makes explainable recommendation
- **Cons**
  - Require domain knowledge to extract item features properly
  - Cannot recommend items of new genre
  - Cold start problem - new user

References

- https://data-science-hi.tistory.com/150
- https://medium.com/@gshriya195/top-5-distance-similarity-measures-implementation-in-machine-learning-1f68b9ecb0a3
- https://eda-ai-lab.tistory.com/524
- https://bab2min.tistory.com/566