UNIT – V

Recommender System

Recommender Systems Functions - Data and Knowledge Sources - Recommendation Techniques - Basics of Content-based Recommender Systems - High Level Architecture-Advantages and Drawbacks of Content-based Filtering - Collaborative Filtering - Matrix factorization models - Neighborhood models.

5.1 Recommender Systems Functions

- Recommender systems are a way of suggesting like or similar items and ideas to a user's specific way of thinking. Recommender systems are widely used on the Web for recommending products and services to users
- Recommender systems try to automate aspects of a completely different information discovery model where people try to find other people with similar tastes and then ask them to suggest new things.
- The goal of a recommender system is to generate meaningful recommendations to a collection of users for items or products that might interest them. Suggestions for books on Amazon, or movies on Netflix, are real-world examples of the operation of industry strength recommender systems.
- Fig. 5.1.1 shows recommendation systems concept.

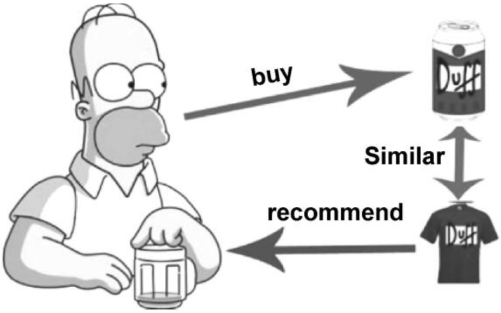
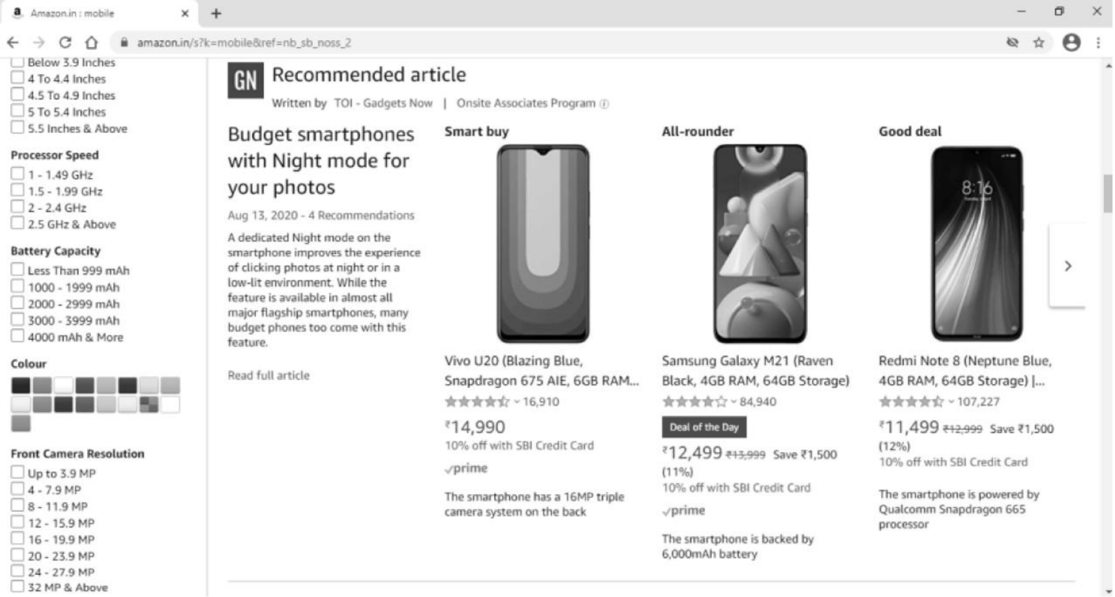


Fig. 5.1.1: Recommendation systems

- Recommendation systems are a key part of almost every modern consumer website. The systems help drive customer interaction and sales by helping customers discover products and services they might not ever find themselves.
- Recommender systems predict the preference of the user for these items, which could be in form of a rating or response. When more data becomes available for a customer profile, the recommendations become more accurate.
- There are a variety of applications for recommendations including movies (e.g. Netflix),

consumer products (e.g., Amazon or similar on-line retailers), music (e.g. Spotify), or news, social media, online dating and advertising.

- Fig. 5.1.2 shows how Amazon uses recommendation concept.



Amazon.in : mobile

amazon.in/s?k=mobile&ref=nb_sb_noss_2

□ Below 3.9 Inches
□ 4 To 4.4 Inches
□ 4.5 To 4.9 Inches
□ 5 To 5.4 Inches
□ 5.5 Inches & Above

Processor Speed
□ 1 - 1.49 GHz
□ 1.5 - 1.99 GHz
□ 2 - 2.4 GHz
□ 2.5 GHz & Above

Battery Capacity
□ Less Than 999 mAh
□ 1000 - 1999 mAh
□ 2000 - 2999 mAh
□ 3000 - 3999 mAh
□ 4000 mAh & More

Colour

Front Camera Resolution
□ Up to 3.9 MP
□ 4 - 7.9 MP
□ 8 - 11.9 MP
□ 12 - 15.9 MP
□ 16 - 19.9 MP
□ 20 - 23.9 MP
□ 24 - 27.9 MP
□ 32 MP & Above

GN Recommended article

Written by TOI - Gadgets Now | Onsite Associates Program

Budget smartphones with Night mode for your photos

Aug 13, 2020 - 4 Recommendations

A dedicated Night mode on the smartphone improves the experience of clicking photos at night or in a low-lit environment. While the feature is available in almost all major flagship smartphones, many budget phones too come with this feature.

Read full article

Smart buy

Vivo U20 (Blazing Blue, Snapdragon 675 AIE, 6GB RAM...

★★★★☆ 16,910

₹14,990

10% off with SBI Credit Card

✓prime

The smartphone has a 16MP triple camera system on the back

All-rounder

Samsung Galaxy M21 (Raven Black, 4GB RAM, 64GB Storage)

★★★★☆ 84,940

Deal of the Day

₹12,499 ~~₹13,999~~ Save ₹1,500 (11%)

10% off with SBI Credit Card

✓prime

The smartphone is backed by 6,000mAh battery

Good deal

Redmi Note 8 (Neptune Blue, 4GB RAM, 64GB Storage) |...

★★★★☆ 107,227

₹11,499 ~~₹12,999~~ Save ₹1,500 (12%)

10% off with SBI Credit Card

The smartphone is powered by Qualcomm Snapdragon 665 processor

Fig. 5.1.2 How Amazon uses recommendation concept

- When you searching mobile on Amazon, it display various mobile and it also display recommended article for mobiles.
- These systems serve two important functions.
  1. They help users deal with the information overload by giving them recommendations of products, etc.
  2. They help businesses make more profits, i.e., selling more products
- Various reasons why service providers increase the use of recommendation systems :
  1. It increases the number of product/items sold.
  2. Sell more diverse products/items
  3. User satisfaction is increases
  4. Increase user fidelity
  5. Better understand what the user wants
- The most common scenario is the following :
  a) A set of users has initially rated some subset of movies (e.g., on the scale of 1 to 5) that they have already seen.
  b) These ratings serve as the input. The recommendation system uses these known ratings to predict the ratings that each user would give to those not rated movies by him/her.
  c) Recommendations of movies are then made to each user based on the predicted ratings.

Recommendation process:

- Every recommendation system follows a specific process in order to produce product recommendations.

- The recommendation approaches can be classified based on the information sources they use. Three possible sources of information can be identified as input for there commendation process. The available sources are the user data (demographics), the item data (keywords, genres) and the user-item ratings.

5.1.1 Challenges

- Following are the challenges for building recommender systems :
  1. Huge amounts of data, tens of millions of customers and millions of distinct catalog items.
  2. Results are required to be returned in real time.
  3. New customers have limited information.
  4. Old customers can have a glut of information.
  5. Customer data is volatile.

5.2 Data and Knowledge Sources

- Recommender systems are information processing systems which actively collect/gather various types of data for designing recommendations system. Data is primarily about the items to suggest and the users who will receive these recommendations.
- Items are the objects that are recommended. Items may be characterized by their complexity and their value or utility.
- Transactions are log-like data that store important information generated during the human-computer interaction and which are useful for the recommendation generation algorithm that the system is using.

5.3 Recommendation Techniques

In general, there are three types of recommender system:

1. Collaborative recommender system is a system that produces its result based on past ratings of users with similar preferences
2. Content based recommender system is a system that produces its result based on the similarity of the content of the documents or items.
3. Knowledge based recommender system is a system that produces its result based on additional and means end knowledge.
4. Demographic based recommender system: This type of recommendation system categorizes users based on a set of demographic classes. This algorithm requires market research data to fully implement. The main benefit is that it doesn't need history of user ratings
5. Hybrid recommender systems combine various inputs and different recommendation

strategies to take advantage of the synergy among them.

5.4 Basics of Content-based Recommender Systems

- Content-based recommenders refer to such approaches that provide recommendations by comparing representations of content describing an item to representations of content that interests the user. These approaches are sometimes also referred to as content-based filtering.
- Content-based recommendation systems try to recommend items similar to those a given user has liked in the past.
- In a movie recommendation application, a movie may be represented by such features as specific actors, director, genre, subject matter, etc.
- The user's interest or preference is also represented by the same set of features, called the user profile.
- Recommendations are made by comparing the user profile with candidate items expressed in the same set of features. The top-k best matched or most similar items are recommended to the user.
- The simplest approach to content-based recommendation is to compute the similarity of the user profile with each item.

5.4.1 High Level Architecture Content-based Recommender Systems

- Fig. 5.4.1 shows High Level Architecture Content-based Recommender Systems.

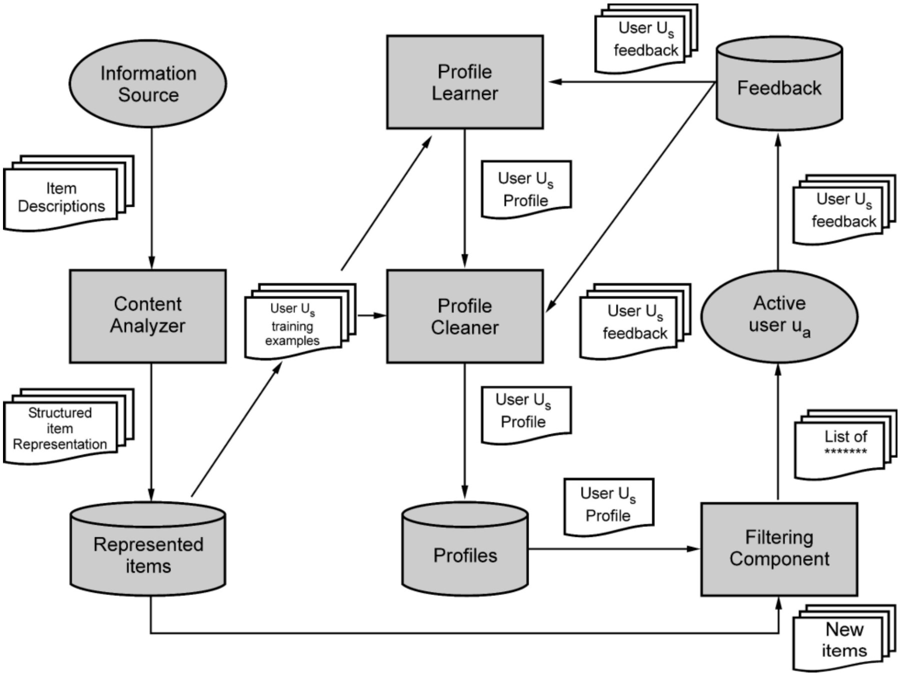


Fig. 5.4.1: High Level Architecture Content-based Recommender Systems

1. Content Analyzer

- Extracts the features (keywords, n-grams) from the source.
- Conversion from unstructured to structured item.
- Data stored in the repository Represented Items

2. Profile Learner

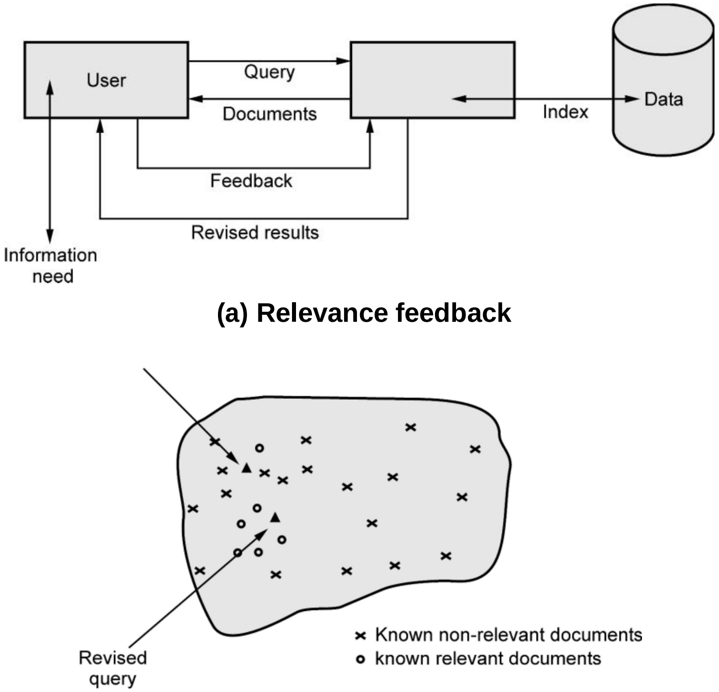
- To build user profile
- Updates the profile using the data in Feedback repository

3. Filtering Component

- Matching the user profile with the actual item to be recommended
- Uses different strategies
- Users have no detailed knowledge of collection makeup and the retrieval environment. Most users often need to reformulate their queries to obtain the results of their interest.

5.4.2 Relevance Feedback

- Thus, the first query formulation should be treated as an initial attempt to retrieve relevant information. Documents initially retrieved could be analyzed for relevance and used to improve the initial query.
- Fig. 5.4.2 shows relevance feedback on initial query.



(a) Relevance feedback

(b) Relevance feedback on initial query

Fig. 5.4.2

- The process of query modification is commonly referred as Relevance feedback, when the user provides information on relevant documents to a query.
- The process of query modification is commonly referred as Query expansion, when information related to the query is used to expand it.
- The user issues a short and simple query. The search engine returns a set of documents. User marks some docs as relevant, some as non-relevant.

Characteristics of relevance feedback:

1. It shields the user from the details of the query reformulation process.
2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones).

Issues with relevance feedback:

1. The user must have sufficient knowledge to form the initial query.
2. This does not work too well in cases like: Misspellings, CLIR and mismatch in user's and document's vocabulary.
3. Relevant documents have to be similar to each other while similarity between relevant and non-relevant document should be small.
4. Long queries generated may cause long response time.
5. Users are often reluctant to participate in explicit feedback.

Advantages of relevance feedback

1. Relevance feedback usually improves average precision by increasing the number of good terms in the query.
2. It breaks down the search operation into a sequence of small search steps, designed to approach the wanted subject area gradually.
3. It provides a controlled query alteration process designed to emphasize some terms and to deemphasize others, as required in particular search environments.

Disadvantages of relevance feedback

1. More computational work
2. Easy to decrease precision

Two basic approaches of feedback methods:

1. **Explicit feedback:** The information for query reformulation is provided directly by the users. However, collecting feedback information is expensive and time consuming.

   The accuracy of recommendation depends on the quantity of ratings provided by the user.

2. **Implicit feedback:** The information for query reformulation is implicitly derived by the system. Implicit feedback reduces the burden on users by inferring their user's preferences from their behavior with the system.

5.4.3 Advantages and Drawbacks of Content-based Filtering

Advantages:

1. User Independence : Recommends only the items that interest the user
2. Transparency : Recommendation is based on the item features, explicitly list the contents features
3. New Item : Helps in recommending new items that are not yet rated by other users

Drawbacks:

1. The user will never be recommended for different items.
2. Business cannot be expanded as the user does not try a different type of product.
3. Overspecialization: Recommends those items that score high with the user profile
4. Cold Start Problem: For a new user, systems don't have historical information to recommend items

5.5 Collaborative Filtering

AU : May-17, Dec.-17

- Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a single user by collecting preferences or taste information from many users (collaborating).
- Collaborative filtering (CF) uses given rating data by many users for many items as the basis for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user. Formally, we have a set of users U = $\{u_1, u_2, \ldots , u_m\}$ and a set of items I = $\{ i_1, i_2, \ldots , i_n\}$. Ratings are stored in a m × n user-item rating matrix.
- The problem of collaborative filtering is to predict how well a user will like an item that he has not rated given a set of historical preference judgments for a community of users.

5.5.1 Type of CF

- There are two types of collaborative filtering algorithms : user based and item based.

1. User based

- User-based collaborative filtering algorithms work off the premise that if a user (A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

- The assumption is that users with similar preferences will rate items similarly. Thus missing ratings for a user can be predicted by first finding a neighborhood of similar users and then aggregate the ratings of these users to form a prediction.
- The neighborhood is defined in terms of similarity between users, either by taking a given number of most similar users (k nearest neighbors) or all users within a given similarity threshold. Popular similarity measures for CF are the Pearson correlation coefficient and the Cosine similarity.
- For example, a collaborative filtering recommendation system for television tastes could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes).
- Note that these predictions are specific to the user, but use information gleaned from many users. This differs from the simpler approach of giving an average score for each item of interest, for example based on its number of votes.
- User-based CF is a memory-based algorithm which tries to mimics word-of-mouth by analyzing rating data from many individuals.
- The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

2. Item-based collaborative filtering

- Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.
- The model-building step consists of calculating a similarity matrix containing all item-to-item similarities using a given similarity measure. Popular are again Pearson correlation and Cosine similarity. All pair-wise similarities are stored in $n \times n$ similarity matrix S.
- Item-based collaborative filtering has become popularized due to its use by YouTube and Amazon to provide recommendations to users. This algorithm works by building an item-to-item matrix which defines the relationship between pairs of items.
- When a user indicates a preference for a certain type of item, the matrix is used to identify other items with similar characteristics that can also be recommended.

- Item-based CF is more efficient than user-based CF since the model is relatively small (N × k) and can be fully pre-computed. Item-based CF is known to only produce slightly inferior results compared to user-based CF and higher order models which take the joint distribution of sets of items into account are possible. Furthermore, item-based CF is successfully applied in large scale recommender systems (e.g., by Amazon.com).

5.5.2 Collaborative Filtering Algorithms

1. Memory-based algorithms:

- Operate over the entire user-item database to make predictions.
- Statistical techniques are employed to find the neighbors of the active user and then combine their preferences to produce a prediction.
- Memory-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as neighbors that have a history of agreeing with the target user.
- Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or top-N recommendation for the active user. The techniques, also known as nearest-neighbor or user-based collaborative filtering are more popular and widely used in practice.
- Dynamic structure. More popular and widely used in practice.

Advantages

1. The quality of predictions is rather good.
2. This is a relatively simple algorithm to implement for any situation.
3. It is very easy to update the database, since it uses the entire database every time it makes a prediction.

Disadvantages

1. It uses the entire database every time it makes a prediction, so it needs to be in memory it is very, very slow.
2. Even when in memory, it uses the entire database every time it makes a prediction, so it is very slow.
3. It can sometimes not make a prediction for certain active users/items. This can occur if the active user has no items in common with all people who have rated the target item.
4. Overfits the data. It takes all random variability in people's ratings as causation, which can be a real problem. In other words, memory-based algorithms do not generalize the data at all.

2. Model-based algorithms:

- Input the user database to estimate or learn a model of user ratings, then run new data through the model to get a predicted output.
- A prediction is computed through the expected value of a user rating, given his/her ratings on other items.
- Static structure. In dynamic domains the model could soon become inaccurate.
- Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user prediction, given his/her ratings on other items.
- The model building process is performed by different machine learning algorithms such as Bayesian network, clustering and rule-based approaches. The Bayesian network model formulates a probabilistic model for collaborative filtering problem.
- The clustering model treats collaborative filtering as a classification problem and works by clustering similar users in same class and estimating the probability that a particular user is in a particular class C and from there computes the conditional probability of ratings.
- The rule-based approach applies association rule discovery algorithms to find association between co-purchased items and then generates item recommendation based on the strength of the association between items

Advantages

1. Scalability: Most models resulting from model-based algorithms are much smaller than the actual dataset, so that even for very large datasets, the model ends up being small enough to be used efficiently. This imparts scalability to the overall system.
2. Prediction speed: Model-based systems are also likely to be faster, at least in comparison to memory-based systems because, the time required to query the model is usually much smaller than that required to query the whole dataset.
3. Avoidance of over fitting: If the dataset over which we build our model is representative enough of real-world data, it is easier to try to avoid over-fitting with model-based systems.

Disadvantages

1. Inflexibility: Because building a model is often a time- and resource-consuming process, it is usually more difficult to add data to model-based systems, making them inflexible.

2. Quality of predictions: The fact that we are not using all the information (the whole dataset) available to us, it is possible that with model-based systems, we don't get predictions as accurate as with model-based systems. It should be noted, however, that the quality of predictions depends on the way the model is built. In fact, as can be seen from the results page, a model-based system performed the best among all the algorithms we tried.

5.5.3 Advantages and Disadvantages

Advantages

1. Collaborative filtering application is to recommend interesting or popular information as judged by the community.
2. Collaborative filtering system can make more personalized recommendation by analyzing information from your past activity or the history of other users of similar taste.

Disadvantages

1. Many commercial recommender systems are based on large datasets. As a result, the user-item matrix used for collaborative filtering could be extremely large and sparse, which brings about the challenges in the performances of the recommendation.
2. As the numbers of users and items grow, traditional CF algorithms will suffer serious scalability problems.
3. Gray sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering.
4. A collaborative filtering system doesn't automatically match content to one's preferences

5.5.4 Difference between Collaborative Filtering and Content based Filtering

Collaborative Filtering	Content Filtering
Collaborative-Filtering systems focus on the relationship between users and items.	Content-Based systems focus on properties of items.
Example : Netflix movie recommendations	Example : Pandora.com music recommendations
Pro : Does not assume access to side information about items	Con : Assumes access to side information about items
Cannot recommend new items	It can recommend new items

Collaborative Filtering	Content Filtering
Item features are inferred from ratings.	Match the item features with user preferences.
Con: Does not work on new items that haveno ratings	Pro: Got a new item to add? No problem, just be sure to include the side information.

University Questions

1. Explain collaborative filtering and content based recommendation system with an example. **AU : May-17, Marks 16**
2. Explain in detail, the collaborative filtering using clustering technique. **AU : Dec.-17, Marks 10**
3. Explain in detail, the collaborative filtering using clustering technique. **AU : Dec-17, Marks 10**
4. Explain collaborative filtering recommendation system with an example. **AU : May-17, Marks 8**

5.6 Matrix Factorization Models

- Matrix factorization (MF) models are based on the latent factor model. MF approach is most accurate approach to reduce the problem from high levels of sparsity in RS database.
- Matrix factorization is a simple embedding model. Given the feedback matrix $A \in R^{m \times n}$, where m is the number of users (or queries) and n is the number of items, the model learns :
  1. A user embedding matrix $U \in R^{m \times d}$, where row i is the embedding for user i.
  2. An item embedding matrix $V \in R^{n \times d}$, where row j is the embedding for item j.

5.6.1 Singular Value Decomposition (SVD)

- SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions from N to K where K < N.
- The matrix factorization is done on the user-item ratings matrix.

$$\overset{X}{\underset{m \times n}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} = \overset{U}{\underset{m \times r}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \overset{S}{\underset{r \times r}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \overset{V^{\top}}{\underset{r \times n}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$

- The matrix S is a diagonal matrix containing the singular values of the matrix X. There are exactly r singular values, where r is the rank of X.
- The rank of a matrix is the number of linearly independent rows or columns in the matrix. Recall that two vectors are linearly independent if they cannot be written as the sum or scalar multiple of any other vectors in the space.

Incremental SVD Algorithm (SVD++)

- The idea is borrowed from the Latent Semantic Indexing (LSI) world to handle dynamic databases.
- LSI is a conceptual indexing technique which uses the SVD to estimate the underlying latent semantic structure of the word to document association.
- Projection of additional users provides good approximation to the complete model
- SVD based recommender systems has following limitations
  a. It cannot be applied on sparse data
  b. Does not have regularization

5.7 Neighbourhood Models

The most common approach to CF is the neighborhood-based approach. Its original form, which was shared by virtually all earlier CF systems, is the user-oriented approach. Such user-Oriented methods estimate unknown ratings based on recorded ratings of likeminded users.

5.7.1 Similarity Measures

- In order to cluster the items in a data set, some means of quantifying the degree of association between them is required. This may be a distance measure, or a measure of similarity or dissimilarity.
- The relationship between the document is described by
  a. Similarity: These values indicate how much two documents or objects are near to each other.
  b. Association: It is same as similarity but difference is objects which are considered for comparison are object characterized by discrete state attributes.
  c. Dissimilarity: Dissimilarity value shows that how much far the objects are.
- The measure of similarity is designed to quantify the likeness between objects so that if one assumes it is possible to group objects in such a way that an object in a group is more like the other members of the group than it is like any object outside the group, then a cluster method enables such a group structure to be discovered.

- A measure of association increases as the number or proportion of shared attribute states increases. In information retrieval system, two documents will be similar to each other if they have more number of common index terms.
- If two documents are having less number of common index terms then obviously they will be semantically far from each other. So we will not include such documents in the same group.
- There are five commonly used measures of association in information retrieval. Since in information retrieval documents and requests are most commonly represented by term or keyword lists. A query is also represented by a list of keywords. Each list of keyword is considered as one set.

5.8 Part A : Short Answered Questions

1. What is collaborative filtering?

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a single user by collecting preferences or taste information from many users (collaborating). It uses given rating data by many users for many items as the basis for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user.

2. What do you mean by item-based collaborative filtering? AU : May-17

Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.

3. What are problem of user based CF?

The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

4. Define user based collaborative filtering. AU : Dec.-16

User-based collaborative filtering algorithms work off the premise that if a user (A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

5. What is cosine similarity?

This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities

6. What are the characteristics of relevance feedback?

Characteristics of relevance feedback:

1. It shields the user from the details of the query reformulation process.
2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones)

7. Write goal of recommender system.

The goal of a recommender system is to generate meaningful recommendations to a collection of users for items or products that might interest them.

8. Define recommender systems.

Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read.

8. What is demographic based recommender system?

This type of recommendation system categorizes users based on a set of demographic classes. This algorithm requires market research data to fully implement. The main benefit is that it doesn't need history of user ratings.

9. What is Singular Value Decomposition (SVD)?

SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions from N to K where K < N.

10. What is Content-based recommender?

Content-based recommenders refer to such approaches that provide recommendations by comparing representations of content describing an item to representations of content that interests the user. These approaches are sometimes also referred to as content-based filtering.

11. What is matrix factorization model?

Matrix factorization is a class of collaborative filtering algorithms used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices