PICT, PUNE

Assignment 1.

Problem statement.

Implementation of conflation algorithm to generate document representation of text file.

Aim: To study

(1) The various concepts & components of IR.

(2) Conflation algorithm.

(3) The role of clustering in IR.

(4) Indexing structure of IR.

Theory:

Information Retrieval

It is the process of locating & selecting data, relevant to a given information. The major functions that constitute an information retrieval system, comprises of acquition, analysis, representation of information, organisation of indexes, matching, retrieving, readjustments & feedback.

Information retrieval system consists of six basic subsystems

(1) Document selection

(2) Indexing

(3) Vocabulary

(4) Searching

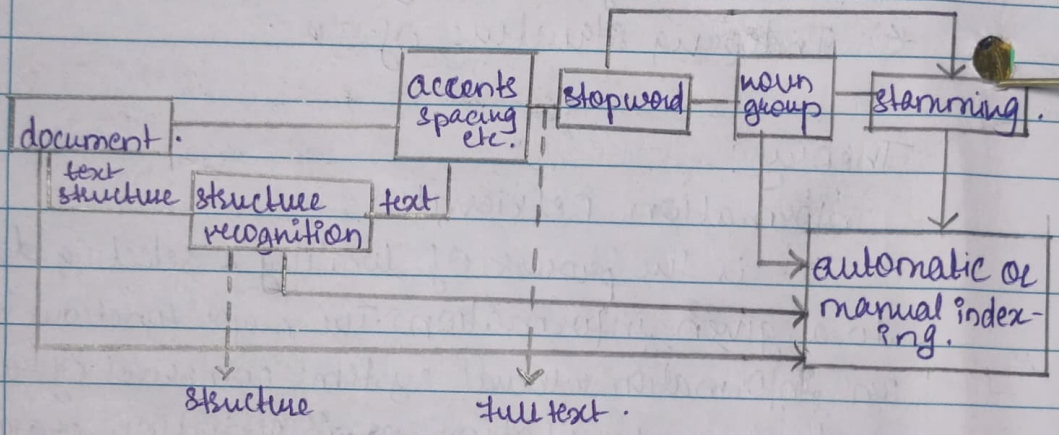
(5) User system

(6) Matching.

PICT, PUNE

Document representative

Document in a collection are frequently represented through a set of index terms or keywords with every large collection, however, even modern, computers might have to reduce the set of representative keywords. These operations are called text operations.



Conflation Algorithm

Ultimately these algorithm are used in information retrieval (IR) systems for matching the morphological variants of terms for efficient indexing & faster retrieval operations.

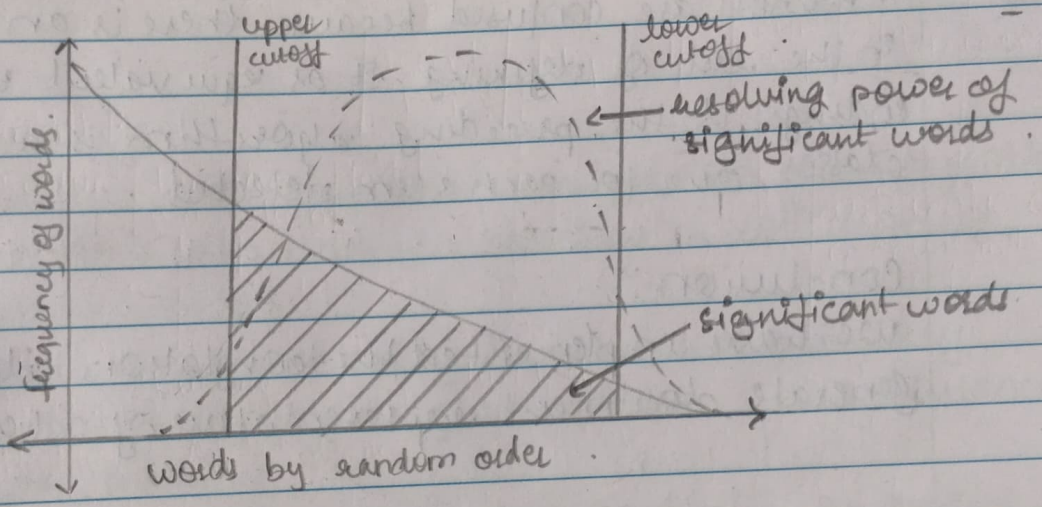
It consists of three parts.

(1) Removal of high frequency words
(2) Buffer stripping
(3) Detecting equivalent stems.

PICT, PUNE

Removal of high frequency words

It is one way of implementing luhn's upper cutoff. This is normally done by comparing the input text with a short list of words which are to be removed. The advantage of the protocol are not only that non-significant words are removed but also that the size of the total document file can be reduced by between 30 to 50 percent.



Suffix stripping

The second stage suffix stripping is more complicated A standard approach is to have a complete list of suffix and to remove the longest possible one. To avoid erroneously removing suffixes, context rules are derived so that a suffix will be removed only if the context is right.

(1) The length of remaining item exceeds a given no.
(2) The stem-ending satisfies a certain condition.

PICT, PUNE

Detecting Equivalent Stems

Many words, which are equivalent in the above sent, map to one morphological form by removing their suffixes. The simplest method of dealing with it is to construct a list of equivalent stem endings. For two stems to be equivalent they must match except for their endings, which themselves must appear in the list as equivalent. Ex: Stems such as ABSORB & ABSORPT are confused because there is an entry in the list of defining it as equivalent stem-endings if the preceding algorithm is a set of classes, one for each stem detected.

Conclusion:

We have implemented the conflation algorithm to generate document representative of a text file.

Assignment 02

Problem statement: Implement single pass algorithm for clustering of files (consider 4-5 files).

Objectives: To study -
(1) what is clustering (2) Single pass algorithm for clustering (3) Measure of association (4) Graphical representation of clustering.

Theory:

Clustering
- It is the most important unsupervised learning problem which deals with finding a structure in collection of unlabelled data.
- Two or more object belong to some cluster if they are close according to given distance. This is distance based clustering
- Goal of clustering - determine instrinsic grouping.

Clustering Requirements :-
The main requirements that a clustering algorithm should satisfy are :-
(1) Scalability (2) Dealing with different types of attribute.
(3) Discovering clusters with arbitary shape.
(4) Minimal knowledge of domain required to determine input parameter.
(5) Ability to deal with noise & outliers.
(6) Insensitivity to order of input records.
(7) High dimensionality.
(8) Interpretability and usability.

PICT, PUNE

Single Pass clustering

- Given a collection of clusters & a threshold value h, if a new document n has the highest similarity more than h to some other cluster, document n is appended to the cluster, & if there exists no cluster, a new cluster is generated which contains only the document n.

Algorithm

(1) Let h be threshold value.
(2) Let S be an empty set & d be the first document. We generate a new cluster a consisting of d1.
(3) When new document di (i>1) comes, calculate the similarity values to all cluster C.
(4) Let simmax be the highest value & cdi the most similar cluster. If simmax > h, add di to cdi & adjust the center of cdi. Otherwise generate a new cluster. cdi that contains only di.
(5) Repeat process above until no data comes.

Measure of association

There are 5 commonly used measures of association:

(1) $|X \cap Y|$ - Simple matching coefficient.
(2) $|X \cap Y| / |X| + |Y|$ - Dice coefficient
(3) $|X \cap Y| / |X \cup Y|$ - Jaccard's coefficient
(4) $|X \cap Y| / |X|^{1/2} * |Y|^{1/2}$ - Cosine coefficient.
(5) $|X \cap Y| / \min(|X|, |Y|)$ - Overlap coefficient.

cluster hypothesis

- The hypothesis is stated as closely associated document tend to be relevant to the same request. The basic assumption in IR system is that

documents relevant to a request are separated from those which have not relevant.

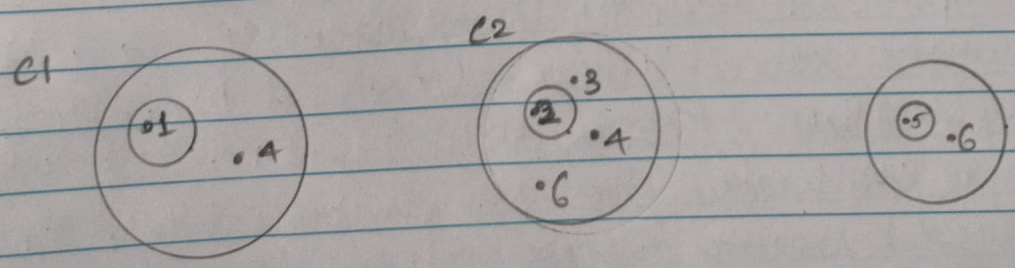
Graphical representation of clustering.

Similarly matrix is used in order to draw a graph, the documents having measure of association greater than the threshold can be represented as edge in graph.

ex objects {1,2,3,4,5,6}. Threshold: 0.59.

1					
2	0.3				
3	0.5	0.6			
4	0.8	0.9	0.7		
5	0.4	0.3	0.85	0.6	
6	0.2	0.8	0.4	0.5	0.6
	1	2	3	4	5	6

clusters are:



PICT, PUNE

conclusion:

Thus we have implemented the single pass algorithm for clustering.

Ass no. 3.

Title: Retrieval of document using inverted files.

Aim: To implement a program for retrieval of documents using inverted files.

Objective:
- To study the concepts of indexing.
- To implement document retrieval using inverted file indexing.
- To practice searching efficiently.

Theory.

Indexing in information Retrieval.

→ When searching, the most basic method is sequential search. It works well for small or frequently changing dataset, but inefficient for large, semistatic collections.

Indexes are built over the text for faster ~~text~~ retrival.

Inverted files.

There are two main types of inverted indexes:
- Record level inverted index: stores, for each word, the list of documents containing it.
- word level (full) inverted index: store for each word, the documents and the positions within those documents where the word appears. This allows for advanced queries like phrase and proximity searches.

PICT, PUNE

Process involves:
① Preprocessing: remove stop words, apply lemmatization.
② Tokenization: split the text into individual terms or words.
③ Index construction: for each term, record the documents.

Searching with inverted files:
① Vocabulary search
② Retrieval of occurences
③ Manipulation of occurency.

Advantages:
- Fast full text search, especially in large document collection.
- Supports complex queries (phrase, proximity) with word level indexes.

Algorithm:
1. Input the conflated file.
2. Build the index file for the input file.
3. Input the query.
4. Print the index file and the result of the query.

Conclusion:
The implementation of document retrieval using inverted file demonstrates the efficiency and effectiveness of indexing.

Assignment 4.

Problem statement: Implement a program to calculate precision & recall for sample input.

Objectives:

1) To understand precision & recall in IR.
2) To study indexing structures for information retrieval

Theory:

Precision & recall in IR:

When a user decides to search for information, the total database & results can be divided into following categories

1) Relevant & Retrieved.
2) Relevant & not retrieved.
3) Non-relevant & retrieved.
4) Non-relevant & not retrieved.

Precision is defined as the ratio of the number of relevant & retrieved documents to the number of total retrieved documents from query.

$$\text{Precision (P)} = \frac{\text{(relevant items retrieved)}}{\text{(retrived items)}} = \left(\frac{\text{relevant}}{\text{retrieved}}\right)P$$

Recall is defined as the ratio of number of retrived documents to the number of possible relevant documents.

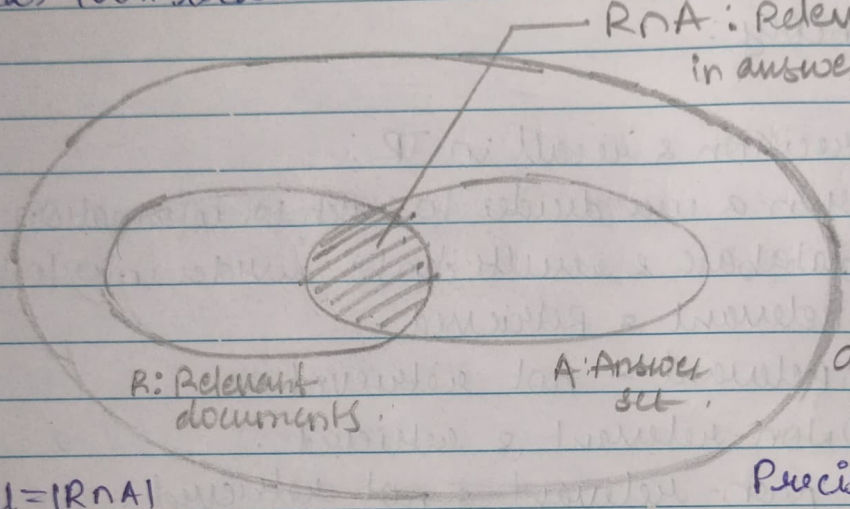
$$\text{Recall (R)} = \frac{\text{(relevant items retrieved)}}{\text{(relevant items)}} = \left(\frac{\text{retrieved}}{\text{relevant}}\right)R$$

PICT, PUNE

$P = \frac{TP}{TP + FP}$ $\quad R = \frac{TP}{TP + FN}$.

Precision recall Trade off.

- Recall is a non-decreasing function of the number of docs retrieved. A system that returns all docs has 100% recall.



$Recall = \frac{|R \cap A|}{|R|}$ $\quad Precision = \frac{|R \cap A|}{|A|}$.

- The goal is to achieve high precision and high recall.
- Precision & recall are extensively used to evaluate the retrieval performance of IR systems or algorithms.
- However it also some problems:

(1) Proper estimation of max recall for a query requires detailed knowledge of all documents in the collection.

(2) In many situations the use of single measure could be more appropriate.

(3) Recall & precision measure effectiveness over a set of queries processed in batch mode.

(4) For system which require a weak ordering, recall & precision might be inadequate.
- The user is not usually presented with all the docs in the answer set A at once. User sees a ranked set of documents & examines them starting from the top. Thus, precision & recall vary as the user ~~proceed~~ proceeds with their examination of set A.

Conclusion:
Implementation is concluded by executing a program to calculate precision and recall for sample input with relevant documents R & for query q.

PICT, PUNE

Assignment 5.

Title: Evaluation of information retrieval system using F-measures and E-measures.

Aim: To write a program that calculates the harmonic mean (F-measure) and E-measure for a given information retrieval (IR) system's performance, thereby evaluating its effectiveness.

Objective:
1. To evaluate the retrieval performance of IR systems using performance metrics.
2. To understand the importance and application of the harmonic mean (F-measure) and E-measure in information retrieval.
3. To study and comprehend different indexing structures used for efficient information retrieval.

Theory:
Precision and Recall:
Information retrieval systems are typically evaluated using two fundamental metrics: precision and recall. Precision measures the fraction of relevant documents that are successfully retrieved.

- Precision (P) = $\frac{\text{Number of relevant document.}}{\text{Total number of document retrieved.}}$

- Recall (R) = $\frac{\text{Number of relevant document retrieved}}{\text{Total number of relevant document.}}$

- **F-measure (F-score):**
The F-measure is the harmonic mean of precision and recall. It's a single metric that provides a balanced evaluation of an IR system, especially when there's an uneven trade-off between precision and recall. The F-score is high only when both precision and recall are high.

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **E-measure:**
The E-measure, is another single metric that combines precision and recall. It introduces a user specified parameter, β which allows user to prioritize either precision or recall.

$$E = 1 - \frac{1 + \beta^2}{\frac{\beta^2}{\text{recall}} + \frac{1}{\text{Precision}}}$$

- $\beta < 1$, user more intrested in recall.
- $\beta > 1$, user more intrested in precision
- $\beta = 1$, E measure is the complement of the F measure i.e ($E = 1 - F$).

PICT, PUNE

Algorithm.

1. Input the number of true positive (relevant document retrieved), number of false positive (irrelevant document retrieved) and number of total relevant document (true positive + false negatives).
2. Calculate Precision and Recall.
3. Calculate the F-measure using harmonic mean formula.
4. Input the user specified parameter β for the E-measure.
5. Calculate the E-measure
6. Display the calculated values.

Conclusion

The implementation successfully calculates the F-measure and E-measure. This demonstrates an understanding of these key metrics and their role in evaluating the effectiveness and retrieval performance of information retrieval systems.

Assignment 6.

Title: Feature Extraction in 2D color images.

Aim: To develop and implement a program for extracting features such as color and texture from a 2D color image and to visualize these features using histogram.

Objective: To study and understand the process of feature extraction from 2D color images, specially focusing on color and texture features, and to create a program that can extract these features and plot their corresponding histograms.

Theory:

Feature extraction is a technique used to transform raw data into a smaller set of meaningful features, making the data easier to process and analyze. In the context of images, it involves converting a complex image into a simpler representation that highlights its most significant characteristics.

- **Colour features:** Colour is a fundamental property for identifying and differentiating objects in an image. Color features are often represented by color histograms, which show the distribution of pixel intensities for each colour channel (eg Red, Green, Blue in RGB color space). By analyzing these histograms, we can understand the dominant color

and overall color composition of an image.

- Texture features: Texture describes the spatial arrangement of pixels in an image. It provides information about the surface characteristics of objects, such as smoothness, roughness or regularity. A common method for texture analysis is the Gray level Coccurrence Matrix (GLCM). The GLCM is a statistical method that calculates how often pairs of pixels with specific intensity values occur in a given direction and distance. From the GLCM, various statistical measures like constract, energy, homogeneity and correlation can be derived to quantify the texture.

Architecture diagram

Input Image	Input of 2D color image.
Color feature Extraction	Extract color information, generation of color histogram for RGB
Grayscale conversion	Convert to grayscale to simplify the data
Texture feature Extraction	Compute the GLCM, from which various texture features are derived.
Visualization	Extracted features are displayed and color histogram are plotted.

Algorithm:

Image import: load a 2D color image into the program using a suitable library.

Preprocessing:

- Convert to grayscale: feature extraction using the GLCM as it simplifies the image data.
- Color channel separation: Split the original color image into its individual color channels (eg RGB or CMYK).

Feature extraction:

- Color features: Calculate the color histogram for each color channel of original image.
- Texture features: Compute the GLCM for the grayscale version of image.

Visualization:

- Histogram: Plot the color histogram for each color channel using a visualization library.

Conclusion:

Feature extraction from 2D color images, by using techniques like color histograms and the grayscale histogram for Gray Level Co-occurence Matrix (GLCM), we were able to quantify and represent important characteristics of an image.

Assignment 7.

Title: Building a web crawler with python.

Aim: To develop a web crawler using python that extracts product information and links from an ecommerce website.

Objective: To understand the fundamental principles and working mechanism of a web crawler and implement a functional version to pull specific data from a live website.

Theory:
A web crawler (or spider) is a program that browses the world wide web systematically. It starts with a list of URL's, known as a seed, and then iteratively downloads the content of those pages. As it processes each page, it identifies and extract new links, which are then added to a queue of pages to visit. This process continues, allowing the crawler to transverse a website or even a significant portion of web.

Key components of a web crawler include:
- URL Frontier/Queue: A data structure that stores the URLs to be crawled.
- Downloader: A model that sends HTTP request to fetch the HTML content of a URL.
- Parser: A component that processes the downloaded HTML, extracting links and relevant data.

PICT, PUNE

An important consideration for web crawler is the Robots Exclusion Protocol (REP), which is defined in a robots.txt file. This file provides directives that inform a crawler which parts of a website it is allowed or disallowed to access. Respecting this protocol is crucial for ethical and responsible crawling.

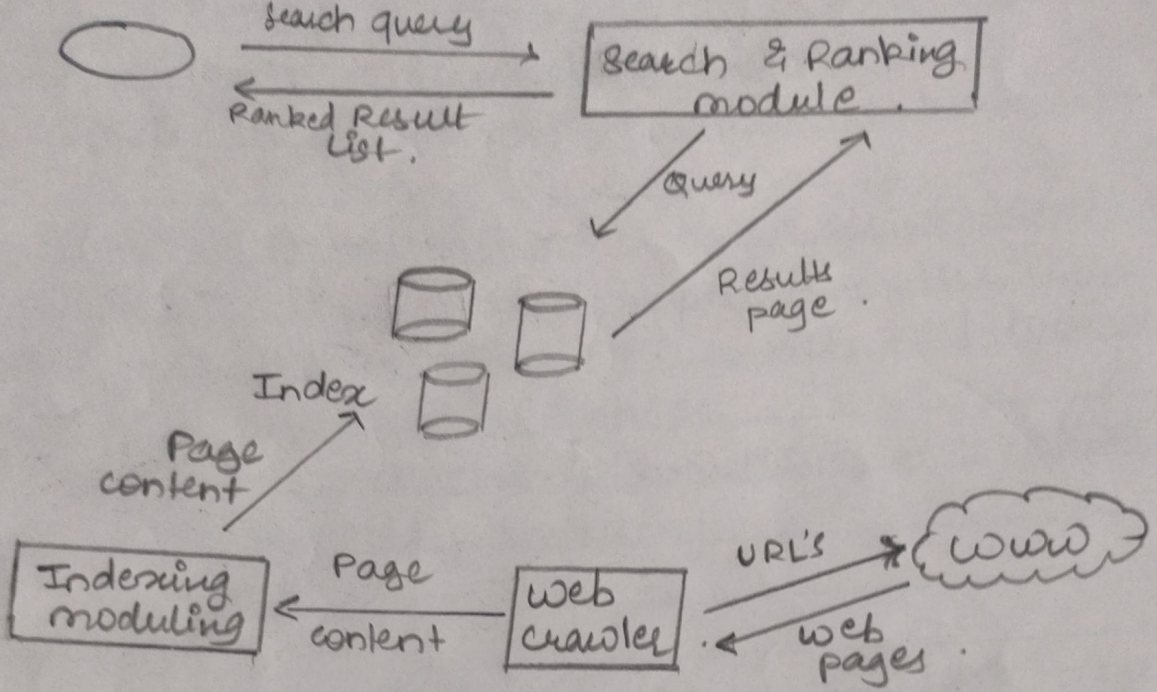
Algorithm

1. Prompt the user to enter the URL of the e-commerce website to be crawled.
2. Send an HTTP GET request to the URL to download the full HTML source code of the webpage.
3. Use Python library (eg Beautiful Soup) to parse the HTML content.
4. Target specific HTML element to pull product names, prices and other relevant information.
5. Store the extracted product information and links in a suitable data structure (eg a list of dictionaries, a CSV file).
6. Present the extracted data and links to user.

Conclusion:

Successfully demonstrates the fundamental working principles of a web crawler.

Architecture :



Web crawler :

→ programs that exploit the graph structure of the web to move from page to page.

→ Designed to retrieve webpages and add them to local repository

→ It downloads a webpage, processes it & following links in that page to other web pages perhaps on other server.

→ System has three peices :
- query processing module, • inverted full text index,
- metadata store.

Assignment 8.

Aim: Develop a Python application that retrieves and presents live weather information for a user specified city by interacting with an external web service

Objective: A functional Python script that can:
1. Accept a city name as input from the user.
2. Access a free weather API to fetch real time data.
3. Extract key weather metrics.
4. Display the extracted information in a clear and user friendly format.

Theory:
Python is an ideal language for this task due to its robust library for handling HTTP requests and processing data formats like JSON. This program relies on the concept of Application Programming Interface (API) which is set of rules and protocol that allows different software applications to communicate with each other. In this scenario, our python program acts as a client, making a request to the OpenWeatherMap API, which acts as a server.

The OpenweatherMap API is a powerful and free web services that provides extensive weather data. To access this data, our program sends an HTTP 'GET' request to specific URL. This request must include an APIkey for authentication and city name as parameter. The API's response is delivered in JSON (Javascript Object Notation) format.

PICT, PUNE

Algorithm.

1. Input: Prompt the user to enter name of city.
2. Define a variable for API key and construct the base URL for API request.
3. Append the users city name and the API key to the base URL, along with any other required parameter (eg. unit of mesurement).
4. Response handling, parse JSON data.
5. Extract the values from a JSON:
   - Temperature, • windspeed, • weather description.
   - Main weather conditions.
6. Print the extracted weather information.

Conclusion

A python program was successfully implemented to retrieve and display live weather information for any given city.

Assignment 9.

Title: A case study on recommender system for E-commerce Product Recommendation

Aim: To understand and implement a recommender system for an ecommerce platform to suggest relevant products to users.

Objective: To study and compare collaborative filtering and content based filtering recommender systems, and to develop a detailed report on a chosen approach, including its design, implementation and evaluation.

Theory:
A recommender system is a subclass of information filtering systems that seeks to predict the 'rating' or 'preference' a user would given an item.
The two primary types are:

- Content based filtering: Recommends item similar to those a user has liked in past. This approach relies on analyzing the attributes of the items and the user's profile. For example, if a user enjoys watching sci-fi movies, the system will recommend other movies with the sci-fi garre.

- Collaborative filtering: Recommends item similar based on the past behavior of other users. This can be broken down into two types:

PICT, PUNE

① User-based: Finds users with similar preferances and recommends item that those similar users have liked.

② Item-based: Recommends items that are similar to the ones the users has previously interacted with, based on other users ratings. for example, if many users who bought item A also bought item B, the system might recommend B to a new user who just bought A.

- A third approach, Hybrid Recommender systems, combines both content based and collaborative filtering to leverage their strengths and mitigate their weaknesses.

Pipeline Architecture Diagram:

Data Ingestion layer — collect raw data (ratings, clicks, purchases).

↓

Data Processing layer — clean, transforms and prepare the data.

↓

Model Training — Train the recommender model must be offline.

↓

Recommendation serving layer — A real time service that takes a user query and returns a list of recommendations.

↓

User interface — presented for user for interaction.

↑

(user) (user)

PICT, PUNE

Algorithm

Item based collaborative filtering algorithm.

1. Data collection: Gather user item interaction data from e-commerce platform. The data can be represented as a user-item matrix.
2. Similarity calculation: use of cosine similarity.

$$\text{Similarity}(A,B) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i}^{n} A_i B_i}{\sqrt{\sum_{i}^{n} A_i^2} \sqrt{\sum_{i}^{n} B_i^2}}.$$

3. Recommendation Generation: Identify the items they have already interacted with. Find the most similar to those items
4. Ranking: Sort the similar items by their similarity score in descending order.

Conclusion:

Outlined the methodology for a recommender system, specified on an item-based collaborative filtering appsoach. This type of system is crucial for enhancing user experience on e-commerce website leading to increased user engagement and sales.