Certainly, here are the answers to your questions in a systematic way:

1. Difference between data retrieval and information retrieval:
- **Data Retrieval:** It involves finding and presenting specific data or facts, often structured and numeric, from a database or dataset.
- **Information Retrieval:** It focuses on searching and retrieving relevant information or knowledge from unstructured or semi-structured sources, such as documents, web pages, or multimedia content. IR is more concerned with context and meaning.

2. Conflation Algorithm:
- The Conflation Algorithm is a technique used for merging and combining similar entities in a dataset to eliminate redundancy and improve data quality.
- Steps of the Conflation Algorithm:
  1. Data Preprocessing: Normalize and clean the data, remove duplicates, and standardize formats.
  2. Feature Extraction: Identify key features to measure similarity.
  3. Pairwise Comparison: Compare records to determine their similarity, e.g., using edit distance or cosine similarity.
  4. Clustering: Group similar records into clusters based on predefined similarity thresholds.
  5. Merge and Resolve: Merge records within clusters to create a single representative record for each entity.

3. Luhn's Idea:
- Luhn's Idea is a concept in information retrieval where a document is divided into sections, and these sections are ranked based on their importance for indexing and retrieval.
- Sections in Luhn's Idea:
  1. High-frequency words: Identify and prioritize the most frequently occurring words in the document.
  2. Significant words: Recognize words that are important for representing the document's content.
  3. Indicative words: Isolate words that uniquely identify the document or express its main theme.
  4. Construct an index: Use these sections to create an index for the document, facilitating efficient retrieval.

4. Stopwords:

- Stopwords are common words (e.g., "the," "is," "in") that are often filtered out during information retrieval or text analysis because they occur frequently in many documents and do not contribute significantly to the understanding of the document's content.

5. Document representative:
- A document representative is a single document or a representation of a group of similar documents. It serves as a point of reference for retrieval and indexing purposes, allowing systems to efficiently retrieve relevant documents based on the representative's characteristics.

6. Indexing, Exhaustivity, and Specificity:
- **Indexing:** The process of creating an index, allowing for faster and more efficient retrieval by extracting keywords or features from documents and organizing them.
- **Exhaustivity:** The extent to which an information retrieval system can find all relevant documents for a given query, with more exhaustive systems retrieving a larger portion of relevant documents.
- **Specificity:** The ability of an information retrieval system to filter out irrelevant documents, providing results closely matching the user's query, with more specific systems delivering highly relevant documents.

7. Five commonly used measures of association in information retrieval:
- Precision
- Recall
- F-measure
- Mean Average Precision (MAP)
- Discounted Cumulative Gain (DCG)

8. Why normalized versions of the simple matching coefficient are used for measures of Association:
- Normalized versions are used to provide standardized measures that allow for fair comparisons across different datasets or systems. They consider differences in the collection size and the number of relevant documents, making the evaluation more meaningful and unbiased. Normalization helps in comparing the performance of information retrieval systems on varying scales.

Certainly, let's address these questions systematically:

1. What is Clustering?
- Clustering is a technique in information retrieval and data analysis used to group similar data points or documents together based on their shared characteristics, attributes, or features.
- It is often employed for data organization and document retrieval to facilitate categorization and exploration of large datasets.

2. Types of Clustering:
- There are several types of clustering, including:
  - **Hierarchical Clustering:** Forms a tree-like structure of clusters, allowing for both top-down and bottom-up exploration.
  - **K-Means Clustering:** Divides data into a pre-defined number of clusters, optimizing the cluster centroids.
  - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Identifies clusters based on data density, suitable for irregularly shaped clusters.
  - **Agglomerative Clustering:** A hierarchical clustering approach that starts with individual data points and merges them into larger clusters.
  - **Spectral Clustering:** Uses the eigenvectors of a similarity matrix to create clusters.

3. Explain the Single Pass Clustering Algorithm:
- The Single Pass Clustering Algorithm is a simple and efficient method for clustering large datasets in a single pass. It works as follows:
  1. Initialize an empty cluster.
  2. Process each data point sequentially.
  3. For each data point, calculate its similarity to existing clusters.
  4. If the similarity is above a threshold, assign the data point to the most similar cluster.
  5. If the similarity is below the threshold, create a new cluster.
  6. Continue this process until all data points are processed.

4. Explain clustering using Similarity Measures:
- Clustering using similarity measures involves quantifying the similarity or dissimilarity between data points or documents. Common similarity measures include Euclidean distance, cosine similarity, and Jaccard similarity. Documents or data points that are more similar are grouped into the same cluster.

5. IR models:

- Information Retrieval (IR) models are frameworks or mathematical models used to rank and retrieve documents in response to user queries. Common IR models include:
  - **Boolean Model:** Based on set theory, matching documents are retrieved.
  - **Vector Space Model:** Documents and queries are represented as vectors, and cosine similarity is used for ranking.
  - **Probabilistic Model (e.g., Okapi BM25):** Ranks documents based on probabilistic relevance.
  - **Language Models (e.g., BM25F):** Treats documents and queries as language models.
  - **Latent Semantic Indexing (LSI):** Applies dimensionality reduction to discover latent semantic structures in documents.

6. Boolean search:
- Boolean search is a type of information retrieval where users create queries using Boolean operators (AND, OR, NOT) to combine keywords. It retrieves documents that match the Boolean query criteria, making it useful for precise searching.

7. What is the multi-pass clustering technique?
- The multi-pass clustering technique involves iteratively refining clusters in multiple passes. It allows for more fine-grained and accurate clustering by considering different criteria or similarity measures in each pass.

8. Explain clustering using a dis-similarity matrix. Also, explain the effect of threshold on clustering:
- Clustering using a dissimilarity matrix quantifies the dissimilarity between data points or documents, usually in the form of a distance matrix. The threshold determines which pairs of data points are considered similar enough to be clustered together. Lower thresholds lead to more clusters, while higher thresholds result in larger clusters.

9. Explain K-list:
- The K-list is a technique in information retrieval where, for each document, a list of the top K most important terms is maintained. These terms represent the document's content and are used for indexing and retrieval. K-lists help in reducing the dimensionality of the term-document matrix.

10. Explain Cluster-Based Retrieval:
- Cluster-Based Retrieval is a retrieval strategy where documents are organized into clusters during indexing. When a user query is issued, the system first identifies relevant clusters and then retrieves documents from those clusters, reducing the search space and improving retrieval efficiency.

11. Explain the working of Rochio's Algorithm:
- Rochio's Algorithm is a relevance feedback method used in information retrieval. It works as follows:
1. The user submits a query, and an initial set of relevant and non-relevant documents is identified.
2. A query vector is constructed based on the user's query.
3. The query vector is adjusted using weighted term vectors of relevant and non-relevant documents.
4. The adjusted query vector is used to retrieve more relevant documents.

These answers should provide a good foundation for your viva. If you need more details on any of these topics, feel free to ask.

Certainly, let's answer these questions systematically:

1. What are Inverted Files?
- Inverted files, also known as inverted indexes, are data structures used in information retrieval to map terms or keywords to the documents or records in which those terms appear. They are fundamental for efficient document retrieval.

2. What is Indexing?
- Indexing is the process of creating an index, a data structure that organizes and stores information (such as keywords or terms) to facilitate faster retrieval. In information retrieval, indexing helps in finding relevant documents quickly when responding to user queries.

3. What is Vocabulary and Occurrences?
- **Vocabulary:** In the context of an inverted index, vocabulary refers to the set of unique terms or keywords present in a collection of documents.
- **Occurrences:** Occurrences refer to the instances of terms within the documents. The index keeps track of where and how many times each term occurs in the documents.

4. How search is carried out on an inverted index?
- Search on an inverted index involves the following steps:
1. The user query is processed, and relevant terms are identified.
2. The index is consulted to retrieve a list of documents that contain these terms.
3. A ranking algorithm is applied to score and rank the retrieved documents.

4. The top-ranked documents are presented as search results to the user.

5. How to index multimedia objects?
- Indexing multimedia objects involves extracting features from multimedia content, such as images, audio, or video. Feature extraction methods differ based on the media type:
  - For images: Features can include color histograms, texture descriptors, and shape information.
  - For audio: Features may comprise spectrograms, pitch, and amplitude modulation.
  - For video: Features could include motion vectors, keyframes, and object recognition.

6. Limitations of Inverted Index:
- Large Storage Requirements: Inverted indexes can become large for extensive document collections.
- Limited Semantics: Inverted indexes may not capture the semantics or context of terms within documents.
- Complex Queries: Handling complex queries beyond simple keyword searches can be challenging.

7. What is Suffix-Array and Suffix-Tree?
- **Suffix-Array:** A data structure used for pattern matching in strings, particularly in information retrieval and bioinformatics.
- **Suffix-Tree:** A tree-like structure representing the suffixes of a string. It's used for efficient pattern matching and substring search in text data.

8. What is the concept of Signature Files?
- Signature files are data structures used for identifying potential candidate documents quickly during information retrieval. They are particularly useful when dealing with large document collections.

9. Working of Inverted Files:
- Inverted files create an index of terms and their locations in documents. This allows for fast document retrieval based on term queries.
- Each term points to a list of document IDs or positions where the term appears.

10. What are the applications of the Inverted Index?
- **Search Engines:** Inverted indexes are essential for web search engines, allowing quick retrieval of web pages.

- **Document Retrieval Systems:** Used in document management and archiving systems.
- **Information Retrieval in Databases:** Supports efficient querying in database management systems.

11. Working of Signature Files:
- Signature files employ hashing techniques to create a compact representation of documents.
- For a query, the system calculates a query signature, and candidate documents are identified by matching their signatures to the query signature.

These answers should help you prepare for your viva on information storage and retrieval. If you have more specific questions or need further details on any topic, feel free to ask.

Certainly, let's address these questions systematically:

1. What is Precision and Recall in IR System?
- **Precision:** Precision is a measure of how many of the retrieved documents are relevant. It assesses the accuracy of the retrieved results.
- **Recall:** Recall is a measure of how many of the relevant documents were retrieved. It evaluates the comprehensiveness of the retrieval system.

2. What is the relevance of the document?
- The relevance of a document in the context of information retrieval refers to the extent to which the document is suitable or pertinent to a user's query. It signifies the value of a document in fulfilling a user's information needs.

3. What are the metrics to measure information systems?
- There are several metrics used to evaluate information retrieval systems, including Precision, Recall, F-measure, Mean Average Precision (MAP), and Discounted Cumulative Gain (DCG).

4. How are Precision and Recall calculated for information systems? (Formulae)?
- **Precision Formula:** Precision = (Number of Relevant Documents Retrieved) / (Total Number of Documents Retrieved)
- **Recall Formula:** Recall = (Number of Relevant Documents Retrieved) / (Total Number of Relevant Documents in the Collection)

5. What is the problem with these two measures?
- Precision and Recall can be in tension with each other. Increasing precision often leads to decreased recall, and vice versa. This trade-off can make it challenging to optimize both measures simultaneously.

6. What is the Precision-Recall Trade-off?
- The Precision-Recall trade-off refers to the inherent tension between precision and recall. Increasing precision typically requires setting more stringent retrieval criteria, which may result in missing some relevant documents (lower recall). Conversely, maximizing recall may lead to lower precision, as more documents are retrieved, including non-relevant ones.

7. What is harmonic mean (F-measure) and E-measure in IR systems?
- **F-measure (harmonic mean):** The F-measure is a single metric that combines both precision and recall into a single value. It is particularly useful when there is a need to balance precision and recall.
- **E-measure:** The E-measure is another metric that considers precision, recall, and a parameter (beta) that allows adjusting the balance between the two. It's useful for situations where one measure is more important than the other.

8. How are (F-measure) and E-measure calculated? (Formulae)?
- **F-measure Formula:** F-measure = 2 * ((Precision * Recall) / (Precision + Recall))
- **E-measure Formula:** E-measure = (1 + beta^2) * (Precision * Recall) / (beta^2 * Precision + Recall), where beta is a parameter that controls the importance of precision and recall.

9. What is the difference between (F-measure) and E-measure?
- The key difference is that the F-measure is a single value that balances precision and recall, while the E-measure allows adjusting this balance using the beta parameter. The E-measure offers more flexibility in specifying the relative importance of precision and recall.

10. What are the metrics to measure information systems?
- Metrics commonly used to measure information systems include precision, recall, F-measure, Mean Average Precision (MAP), and Discounted Cumulative Gain (DCG).

11. What is the advantage of (F-measure) and E-measure?
- F-measure provides a single metric that balances precision and recall, making it easy to interpret and compare different systems. E-measure, on the other hand, allows

for more fine-grained control over the trade-off between precision and recall, making it adaptable to specific requirements or preferences in different applications.

These answers should help you prepare for your viva on information storage and retrieval. If you have more specific questions or need further details on any topic, feel free to ask.

Certainly, let's address these questions systematically:

1. What is Extraction (or Feature Extraction)?
- Feature extraction is a process in information retrieval that involves capturing essential information or characteristics (features) from data, such as text, images, or other types of content. These features are used to represent and describe the data, making it more suitable for analysis or retrieval.

2. How are images indexed?
- Images are indexed by extracting relevant features from them, such as color, texture, shape, and visual content. These features are then stored in an index, allowing for efficient retrieval based on image characteristics.

3. Explain how color is extracted from an image:
- Color extraction from an image involves quantifying the distribution of colors in the image. This can be done using methods like color histograms, color moments, or color coherence vectors. These techniques capture information about the frequency and distribution of colors in the image.

4. What is Multimedia IR? Discuss steps on which data retrieval relies.
- Multimedia Information Retrieval (MIR) is the process of retrieving relevant multimedia data, such as images, audio, or video, based on user queries. Steps in MIR include:
  1. Content Analysis: Analyzing multimedia content to extract features like color, shape, or audio characteristics.
  2. Indexing: Storing extracted features in an index for efficient retrieval.
  3. Query Processing: Interpreting user queries and matching them to multimedia data in the index.
  4. Ranking: Scoring and ranking retrieved multimedia based on relevance.
  5. Presentation: Displaying relevant multimedia content to users.

5. What is the use of image features?
- Image features are used to describe the visual content of images, enabling efficient image retrieval and analysis. These features help in tasks like image search, object recognition, and content-based image retrieval.

6. Enlist some of the features of the image and its applications:
- Image features include color histograms, texture descriptors, edge information, and shape characteristics. Applications of image features include:
  - Image search in search engines.
  - Object recognition in computer vision.
  - Medical image analysis for diagnosis.
  - Content-based image retrieval in digital libraries.

7. How to compare two images and calculate the relevancy?
- Images can be compared based on their features, and relevancy can be calculated using similarity measures like:
  - Euclidean distance for color histograms.
  - Cosine similarity for feature vectors.
  - Structural similarity (SSIM) for overall image similarity.
  - Intersection over Union (IoU) for object overlap in images.

8. Applications of Feature Extraction:
- Feature extraction is used in various applications, including:
  - Text analysis to extract keywords.
  - Audio analysis to identify voice patterns.
  - Image analysis for object recognition.
  - Handwriting recognition in character recognition systems.

Certainly, here are the answers to these questions in a systematic manner:

1. What are search engines? Name a few of them:
- Search engines are web-based tools or software applications that allow users to search for information on the internet. They index web content and provide relevant results to user queries. Some well-known search engines include Google, Bing, Yahoo, and DuckDuckGo.

2. How Search Engine Works:
- Search engines operate through a process that includes:
  - Crawling: Automated bots (crawlers or spiders) explore the web by visiting websites and collecting information.
  - Indexing: Data from web pages is organized and stored in a searchable index.
  - Ranking: Algorithms analyze content to determine relevance, and ranked results are generated.
  - Retrieval: Users submit queries, and the search engine returns relevant results from its index.

3. What is Web Crawling:
- Web crawling is the process by which search engine crawlers systematically navigate the internet, visiting websites, and collecting data from web pages. This data is then indexed for search.

4. What is the Robot Exclusion Protocol (robot.txt):
- The Robot Exclusion Protocol, often referred to as "robots.txt," is a standard used by websites to communicate with web crawlers. It specifies which parts of a website are off-limits to crawlers and which areas can be explored.

5. What is the significance of robot.txt:
- Robot.txt is significant for webmasters because it provides control over what parts of their website are crawled by search engines and which are not. It helps in managing the visibility of sensitive or irrelevant content.

6. What are the strategies used by Crawler:
- Crawlers use various strategies, including breadth-first, depth-first, and focused crawling:
  - **Breadth-First:** Visit pages on the same level before going deeper into the site hierarchy.
  - **Depth-First:** Explore pages deeply, visiting child pages before sibling pages.
  - **Focused Crawling:** Prioritize pages related to a specific topic or keyword.

7. What is Page Rank:
- PageRank is an algorithm used by Google to rank web pages in search results. It measures the importance of web pages based on the number and quality of links pointing to them.

8. What is the significance of the Dampening Factor:

- The damping factor is a value used in the PageRank algorithm to control the probability that a user follows a link on a web page. It helps to prevent pages with excessive links from receiving disproportionately high PageRank scores.

9. What are the Crawler Architectures:
- There are several crawler architectures, including single-threaded, multi-threaded, distributed, and focused crawling:
  - **Single-Threaded:** A single crawler sequentially retrieves web pages.
  - **Multi-Threaded:** Multiple threads or processes are used to crawl web pages in parallel.
  - **Distributed:** Crawlers are distributed across multiple machines to handle a large-scale crawling task.
  - **Focused Crawling:** Emphasizes specific areas or topics on the web.

10. Explain Harvest Architecture:
- The Harvest Project was an early effort to develop a scalable web crawling architecture that used a distributed network of crawler servers. It utilized the Gatherer, Broker, and Cheshire components for harvesting, indexing, and searching.

11. Explain the working of GOOGLE Crawler:
- Google's crawler, called Googlebot, operates by visiting web pages, collecting data, and following links to discover and index new content. It uses sophisticated algorithms to determine the importance and relevance of pages, contributing to Google's search rankings.

12. Explain Challenges involved in searching the web:
- Web search faces challenges such as handling the vast size of the web, dealing with constantly changing content, managing spam and low-quality content, and providing accurate and timely results to users.

These responses should help you prepare for your viva on information storage and retrieval. If you have more specific questions or need further details on any topic, feel free to ask.

Certainly, let's address these questions systematically:

1. What are APIs and their Use:
- **APIs (Application Programming Interfaces):** APIs are a set of rules and protocols that allow different software applications to communicate with each other. They define how requests and data should be structured and exchanged.
- **Use of APIs:** APIs are used to enable various functionalities in software applications, such as accessing data, services, or features from other systems, applications, or platforms. They provide a standardized way to integrate and interact with external resources.

2. How to use API:
- To use an API, you typically need to follow these steps:
  1. Obtain API Access: Sign up for an API key or authentication credentials from the API provider.
  2. Read API Documentation: Familiarize yourself with the API's documentation, which explains how to make requests and what data or services are available.
  3. Make API Requests: Use HTTP requests (e.g., GET, POST) to interact with the API, including specifying the endpoint, parameters, and headers.
  4. Process API Responses: Handle the data returned by the API, which may be in JSON, XML, or another format, and integrate it into your application.

3. Which API have you used in your Assignment 8? (We have Used OpenWeatherMap API):
- In your assignment, you used the OpenWeatherMap API, which provides weather data and forecasts for locations around the world.

4. Explain the API you have used in Assignment 8:
- The OpenWeatherMap API is a service that provides weather information for various locations. You likely used it to retrieve data such as current weather conditions, forecasts, and other related weather information. The API requires an API key for authentication and offers a range of endpoints to access specific weather data.