Information Retrieval (IR) and Automatic Text Analysis:

Basic Concepts of IR, Data Retrieval & Information Retrieval

1. Explain the main differences between data retrieval and information retrieval. How does IR handle unstructured data?

2. What are the key components of an Information Retrieval System, and how do they interact with each other?

3. Define recall and precision in the context of IR. How are they measured, and why are they important?

4. How does relevance feedback improve the performance of an IR system? Give an example.

Text Mining and IR Relation

5. How does text mining complement information retrieval? What are some common techniques used in text mining?

6. What is Natural Language Processing (NLP), and how is it applied in text mining and IR?

IR System Block Diagram

7. Can you describe the typical block diagram of an IR system and explain the role of each block?

8. In an IR system, how is the query processing module designed to handle synonymy and polysemy?

Automatic Text Analysis

9. What was the significance of Luhn's ideas in automatic text analysis, and how are they relevant today?

10. Describe the purpose of a conflation algorithm. How does it differ from stemming and lemmatization in text analysis?

Indexing and Index Term Weighting

11. What is the purpose of indexing in IR, and how does it enhance search efficiency?

12. Explain the concept of term weighting. Why is TF-IDF commonly used, and how does it affect document ranking?

Probabilistic Indexing

13. What is probabilistic indexing, and how does it differ from traditional Boolean indexing methods?

14. How does a probabilistic model determine the relevance of documents to a query? What factors are taken into consideration?

Automatic Classification

15. What is automatic classification in the context of IR, and what are some commonly used algorithms?

16. How does Naive Bayes work for document classification, and what are its limitations?

Measures of Association and Different Matching Coefficients

17. What are measures of association in IR, and why are they important for information retrieval tasks?

18. Define cosine similarity and Jaccard similarity coefficient. In what scenarios would each be preferable in IR?

Cluster Hypothesis and Clustering Techniques

19. Explain the cluster hypothesis. How does it influence the design of clustering techniques in IR?

20. Describe Rocchio's Algorithm and compare it to the Single Pass and Single Link algorithms for clustering. What are the main differences and use cases for each?

Indexing, Searching Techniques, and IR Models:

Indexing: Inverted File, Suffix Trees & Suffix Arrays, Signature Files, Scatter Storage or Hash Addressing

1. What is an inverted file, and why is it essential for efficient document retrieval in IR?

2. How do suffix trees and suffix arrays facilitate substring searching in IR systems? What are their advantages and limitations?

3. Explain the role of signature files in indexing. How do they compare to inverted files in terms of efficiency and storage?

4. Describe scatter storage or hash addressing in indexing. How does this technique enhance retrieval speed, and what are potential drawbacks?

Searching Techniques: Boolean Search, Sequential Search, Serial Search, Cluster-Based Retrieval, Query Languages, Types of Queries, Pattern Matching, Structural Queries

5. How does Boolean search work in IR, and what are its main limitations?

6. What is the difference between sequential search and serial search? In what contexts are these techniques useful in IR?

7. Explain the concept of cluster-based retrieval. How does clustering improve search relevance in large datasets?

8. What is a query language in IR, and how do different types of queries (e.g., keyword, phrase, and wildcard queries) impact search results?

9. Describe pattern matching and its applications in information retrieval. How does it differ from structural queries?

10. What are structural queries, and how are they used in document-based IR systems? Provide an example.

IR Models: Basic Concepts, Boolean Model, Vector Model, Probabilistic Model

11. What is the main difference between the Boolean model and the Vector model in IR?

12. How does the Vector Space Model handle document similarity? Explain the role of cosine similarity in this model.

13. Describe the Probabilistic Model in IR. How does it determine the relevance of documents to a query?

14. What are the strengths and limitations of the Boolean model? Why might it be less effective for complex queries?

15. In what scenarios would the Vector Model be preferable over the Boolean and Probabilistic Models?

Advanced and Conceptual Questions

16. How can suffix arrays be optimized for large-scale IR systems, and what are their benefits for pattern matching?

17. Compare scatter storage with traditional hashing techniques. How does scatter storage handle collisions in IR?

18. Explain how clustering can assist with query expansion in IR systems.

19. How do different types of queries (such as keyword queries and pattern-based queries) influence the choice of IR model?

20. Discuss the importance of using probabilistic approaches in modern IR systems. How do they improve over purely deterministic models like the Boolean model?

Performance Evaluation and Visualization in Information Systems:

Performance Evaluation: Precision and Recall, MRR, F-Score, NDCG,
User-Oriented Measures

1. Define precision and recall in the context of IR. How do they differ, and why are they both important for evaluating search performance?

2. What is the F-Score, and how does it balance precision and recall? Why might a high F-Score still not fully represent system performance?

3. Explain Mean Reciprocal Rank (MRR). How does it provide insight into the quality of search results?

4. Describe Normalized Discounted Cumulative Gain (NDCG). How does it account for the ranking position of relevant documents in the evaluation?

5. What are user-oriented measures in IR performance evaluation? Give examples and explain their importance in assessing system usability.

6. Why might it be necessary to use multiple evaluation metrics (such as NDCG, precision, and recall) to assess an IR system’s effectiveness?

Visualization in Information Systems: Starting Points, Query Specification, Document Context, User Relevance Judgment, Interface Support for Search Process

7. What is the role of visualization in Information Retrieval systems, and how does it impact user experience?

8. Explain the concept of 'starting points' in search interfaces. Why are they critical in IR systems?

9. How does query specification in a user interface affect search outcomes? What are some examples of effective query specification tools?

10. What is document context in IR, and how does providing it help users better evaluate search results?

11. Define user relevance judgment. How does an IR system support or enhance relevance judgment through visualization?

12. Describe different types of interface support for the search process. How do these enhance user engagement and effectiveness?

Advanced and Conceptual Questions

13. How might NDCG be particularly useful in applications like recommendation systems?

14. Discuss the limitations of precision and recall as performance metrics in complex IR systems.

15. How can visualization aid in query reformulation, and what interface elements are particularly helpful for this?

16. Explain how document clustering and visualization can enhance user relevance judgment in exploratory searches.

17. What are some challenges in designing user-oriented evaluation metrics, and how do they address user satisfaction?

18. How do interactive visualizations help users modify or refine their searches within an IR system?

19. What are the trade-offs between using a simplified search interface and a highly customizable query interface?

20. In what scenarios might user-oriented measures be prioritized over traditional metrics like precision and recall?

Distributed Information Retrieval (Distributed IR) and Multimedia Information Retrieval (Multimedia IR):

Distributed Information Retrieval: Introduction, Collection Partitioning, Source Selection, Query Processing

1. What is Distributed Information Retrieval (DIR), and how does it differ from traditional IR systems?

2. Explain the concept of collection partitioning in DIR. What are some advantages and challenges associated with it?

3. What is source selection in DIR, and how does it affect query processing efficiency?

4. How does query processing differ in a distributed IR environment compared to a centralized IR system?

5. What strategies are used to ensure effective coordination among distributed sources in DIR?

6. Describe the role of a broker in distributed IR systems. How does it help manage multiple sources?

7. Explain the impact of network latency and bandwidth on query performance in distributed IR systems. How can these issues be mitigated?

Multimedia Information Retrieval: Introduction, Data Modeling, Query Language, Background-Spatial Access Method, Generic Multimedia Indexing, Time Series, Color Images, Feature Extraction, Trends, and Research Issues

8. Define Multimedia Information Retrieval (MMIR) and describe its unique challenges compared to text-based IR.

9. What is data modeling in MMIR, and why is it crucial for representing multimedia data accurately?

10. Explain how query languages for MMIR differ from those used in text IR. What are some specific query features for multimedia data?

11. What is the Spatial Access Method, and why is it important for querying multimedia data?

12. Describe a generic multimedia indexing approach and its role in managing large multimedia collections.

13. How are one-dimensional time series data represented and queried in MMIR systems?

14. Explain how two-dimensional color images are indexed and retrieved in MMIR systems. What role does color histograms play in this process?

15. What is automatic feature extraction, and why is it a critical component of MMIR?

16. What are some current trends and research issues in MMIR? Discuss one example.

17. How does MMIR handle data with high dimensionality, such as images or audio files, during indexing and retrieval?

18. Describe the challenges associated with temporal data in MMIR, such as in video or audio retrieval systems.

Advanced and Conceptual Questions

19. What are the main limitations of current spatial access methods for large-scale multimedia data?

20. Explain the role of machine learning in automatic feature extraction for multimedia retrieval. How is it advancing MMIR capabilities?

Web Search and Web Scraping:

Web Search: Introduction, Challenges, Web Characteristics, Search Engines, Ranking, Crawling, Indices, Browsing, Meta-searchers, Hyperlink Search, Trends and Research Issues

1. What are the primary challenges in web search, and how do they differ from traditional information retrieval?

2. Explain the unique characteristics of the web that impact search engine design and performance.

3. Describe the differences between centralized and distributed architectures in search engines. What are the benefits of each?

4. What are the main components of a search engine's ranking system? How does it determine the relevance of search results?

5. Explain the process of crawling in web search engines. How is a web crawler designed to handle dynamic or frequently changing content?

6. What is an index in the context of web search, and how does it improve the efficiency of a search engine?

7. Describe the purpose and function of a meta-search engine. How does it differ from a traditional search engine?

8. How does hyperlink-based searching, such as PageRank, enhance search relevance on the web?

9. What are some emerging trends and research issues in web search technology?

10. How is user interaction data utilized to improve search engine results and interfaces?

Web Scraping: Python, Requests, HTML Parsing, Beautiful Soup

11. What is web scraping, and how is it typically used in data gathering or analysis?

12. Explain the difference between web scraping and web crawling. What are the legal or ethical considerations in web scraping?

13. How does the Requests library in Python facilitate web scraping? Provide an example of a simple GET request.

14. What is HTML parsing, and why is it necessary in web scraping?

15. Describe the Beautiful Soup library. How does it help in locating and extracting specific data from an HTML document?

16. What are some common challenges encountered in web scraping, such as CAPTCHA or dynamic content, and how can they be addressed?

17. Explain how CSS selectors and HTML tags are used to locate elements in Beautiful Soup.

Advanced and Conceptual Questions

18. How can a distributed search engine architecture improve crawling efficiency in large-scale search engines?

19. In what scenarios would using a meta-search engine be more advantageous than a traditional search engine?

20. Discuss the role of web scraping in data-driven research. What are some best practices for using Python libraries like Requests and Beautiful Soup responsibly?

XML Retrieval, Recommendation Systems, and the Semantic Web:

XML Retrieval: Basic Concepts, Challenges, Vector Space Model, Evaluation, Text-Centric vs. Data-Centric XML

1. What is XML, and why is it commonly used in data storage and retrieval?

2. Explain some unique challenges in XML retrieval compared to traditional text retrieval.

3. How does the Vector Space Model apply to XML retrieval? What modifications are needed for structured data?

4. What are some key evaluation metrics for XML retrieval, and how do they differ from those used in traditional IR?

5. Differentiate between text-centric and data-centric XML retrieval. Why might a system focus on one over the other?

6. What is the role of XPath and XQuery in XML retrieval?

7. How can the hierarchical structure of XML data complicate or enhance retrieval strategies?

8. Describe a scenario where XML retrieval is more beneficial than standard IR approaches.

Recommendation Systems: Collaborative Filtering, Content-Based Recommendation, Document/Product Recommendations

9. Define collaborative filtering and explain its use in recommendation systems. What types of data does it typically rely on?

10. What is content-based recommendation, and how does it differ from collaborative filtering?

11. What are the limitations of collaborative filtering in a recommendation system?

12. Explain how a hybrid recommendation system combines collaborative filtering and content-based approaches. What are the benefits of this?

13. What are the challenges of building a recommendation system for a new product or document with no prior user interactions (cold-start problem)?

14. In what ways can user profiling be used to improve recommendation accuracy?

15. How does a recommendation system handle diversity and novelty in recommendations?

Semantic Web: Introduction

16. What is the Semantic Web, and how does it aim to enhance traditional web data?

17. Explain the role of RDF (Resource Description Framework) in the Semantic Web. How does it enable data linking?

18. What are ontologies, and how do they contribute to the functionality of the Semantic Web?

19. Describe how the Semantic Web enables improved data interoperability. Give an example.

20. What are the main challenges currently faced in the adoption of the Semantic Web?