Don Bosco Institute of Technology
Department of Information Technology

Syllabus and Question Bank for IAT-1

Class : BE Sem - 7 Subject – Dept. Level Elective – ITDO7024 IRS

Subject Incharge : Ms. Aruna Khubalkar

Syllabus : Module – I, Module – II and Module-III (till Query Protocols)

Module - I INTRODUCTION

1. Define information retrieval.
   Information Retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

2. What are the applications of IR?

3. Compare & contrast data retrieval and information retrieval.

4. Explain the general objective of an Information Retrieval System.

5. Define relevance.
   Relevance appears to be a subjective quality, unique between the individual and a given document supporting the assumption that relevance can only be judged by the information user. Subjectivity and fluidity make it difficult to use as measuring tool for system performance.

6. List and explain components of IR block diagram.

7. What is conflation?
   Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. The process of stemming if often called conflation.

8. Explain information retrieval process with a generic software architecture diagram.

9. What is search engine?
   A search engine is a document retrieval system design to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.

10. Give the functions of information retrieval system.
    • To identify the information(sources) relevant to the areas of interest of the target users community
    • To analyze the contents of the sources(documents)
    • To represent the contents of the analyzed sources in a way that will be suitable for matching user's queries

• To analyze user's queries and to represent them in a form that will be suitable for matching with the database
• To match the search statement with the stored database
• To retrieve the information that is relevant
• To make necessary adjustments in the system based on feedback form the users.

11. List the issues in information retrieval system.
    • Assisting the user in clarifying and analyzing the problem and determining information needs.
    • Knowing how people use and process information.
    • Assembling a package of information that enables group the user to come closer to a solution of his problem.
    • Knowledge representation.
    • Procedures for processing knowledge/information.
    • The human-computer interface.
    • Designing integrated workbench systems.

- Designing user-enhanced information systems.
- System evaluation.

Module – II

1. What do you mean information retrieval models?
   A retrieval model can be a description of either the computational process or the human process of retrieval: The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

2. Differentiate retrieval task Adhoc and Filtering.

3. State the formal characterization of IR Models.

4. What are the three classic models in information retrieval system?
   1. Boolean model
   2. Vector Space model
   3. Probabilistic model

5. Name the alternative modeling paradigms of classic models in IRS?

6. Classify diagrammatically different types of retrieval models associated with distinct combinations of a document logical view and a user task.

7. Define Boolean retrieval model and explain it with an example.

8. What is the basis for boolean model?
   Simple model based on set theory and Boolean algebra
   - Documents are sets of terms
   - Queries are specified as Boolean expressions on terms

9. What are the advantages of Boolean model?
   - Clean Formalism
   - Easy to implement
   - Intuitive concept

■ Still it is dominant model for document database systems

10. What are the disadvantages of Boolean model?
    - It is not simple to translate an information need into a Boolean expression.
    - Exact matching may lead to retrieval of too few or too many documents.
    - Difficult to rank output, some documents are more important than others.
    - The model does not use term weights (all terms are equally weighted).
    - More like data retrieval than information retrieval
    - No notion for partial matching

11. What is the significance of tf and idf? How can you calculate tf and idf in a vector model?

12. What are the Parameters in calculating a weight for a document term or query term?
    Term Frequency (tf): Term Frequency is the number of times a term i appears in document j (tfij)
    – Document Frequency (df): Number of documents a term i appears in, (dfi).
    – Inverse Document Frequency (idf): A discriminating measure for a term i in collection, i.e., how discriminating term i is. (idf i) = log10(n / dfi), where n is the number of document

13. What are the assumptions of vector space model?
    Assumption of vector space model:
    The degree of matching can be used to rank-order documents;
    This rank-ordering corresponds to how well a document satisfying a users information needs.

14. Explain about vector space model in detail.

15. Define Vector model.

16. What are the advantages of Vector Model?
    - Simple model based on linear algebra
    - Term weights not binary
    - Allows computing a continuous degree of similarity between queries and documents
    - Allows ranking documents according to their possible relevance
    - Allows partial matching
    - Allows efficient implementation for large document collections

17. What are the disadvantages of Vector Model?
    ■ Index terms are assumed to be mutually independent
    ■ Search keywords must precisely match document terms
    ■ Long documents are poorly represented
    ■ The order in which the terms appear in the document is lost in the vector space representation
    ■ Weighting is intuitive, but not very formal

18. What is link analysis?
    The goal of information retrieval is to find all documents relevance for a user query in a collection of documents. With the advent of the web new source of information became available, one of them being the hyperlink between documents and records of user behavior. Collections of documents connected by hyperlinks. Hyperlinks provide a

valuable source of information for web information retrieval. This area of information retrieval is commonly link analysis.

19. Explain in detail about Probabilistic model and briefly describe Simple term weights with an example.

20. What is a term frequency and normalized term frequency? Write down their equations.

21. Define Probabilistic model or Binary Independence Retrieval :-
    The Objectiveof Probabilistic model is to capture the IR problem using a probabilistic framework
    Given a user query, there is an ideal answer set
    ⇒ Querying as specification of the properties of this ideal answer set
    Definition

    ⇒ Weight variables all are binary, i.e. w i,j €{0,1} and w i,q €{0,1}
    ⇒ q- a query is a subset of index terms
    ⇒ R – set of doc's known (initial guess) to be relevant
    ⇒ R – the complement of R, i.e. the set of non-relevant doc's
    ⇒ P(R|d j) – probability of d j relevant to q
    ⇒ P(R|d j) - probability of d j non-relevant to q

    sim(d j ,q) = P(R|d j) / P(R|d j)

22. What are the Fundamental assumptions for probabilistic principle?
    ⇒ q- user query,d j – doc in the collections
    ⇒ Model assumes, relevance depends on the query and the doc representation only
    ⇒ R – ideal answer set, relevant to the query
    ⇒ R - ideal answer set, non-relevant to the query
    ⇒ Similarity to the query ratio is, i.e. probabilistic ranking computed as
    ⇒ Ratio = P(d j relevant-to q) / P(d j non-relevant-to q)
    ⇒ The rank minimizes the probability of the erroneous judgment

23. Write the advantages and disadvantages of probabilistic model:
    Advantages
    ⇒ Doc's are ranked in decreasing order of their probability of relevant
    Disadvantages
    ⇒ Need to guess the initial separation of doc's into relevant and non-relevant sets
    ⇒ All weights are binary
    ⇒ The adoption of the independence assumption for index terms
    ⇒ need to guess initial estimates for P(ki | R)
    ⇒ method does not take into account tf and idf factors

24. Discuss structured text retrieval models in details.

<u>Module – III</u>

1. State different types of queries.
2. Explain the pattern matching query concept with an example.
3. What is Keyword based querying? Discuss contect queries and Boolean queries in detail with example.