Vivekanand Education Society's Institute of Technology, Chembur, Mumbai,

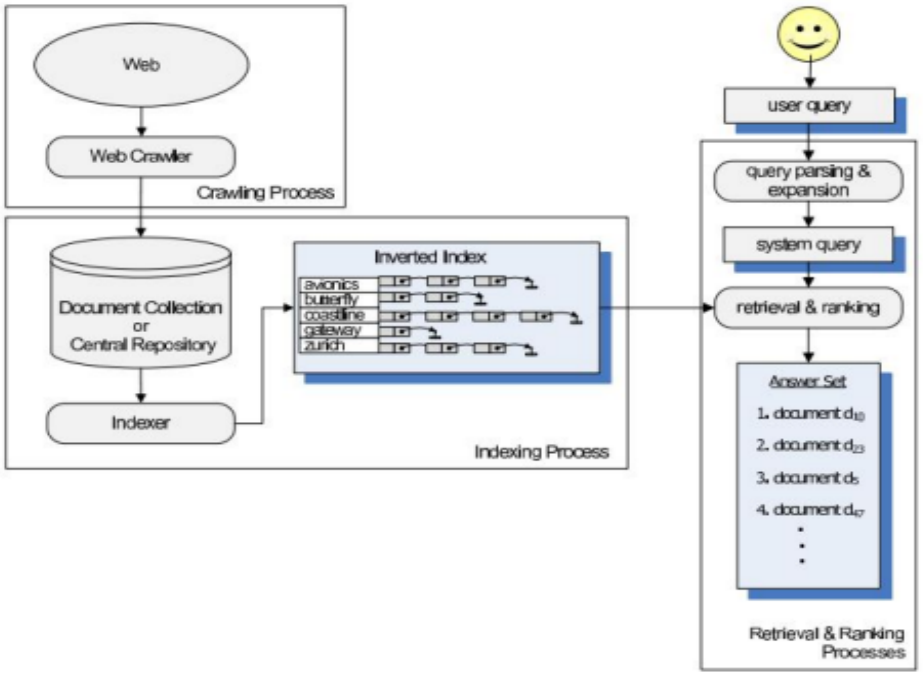
Department Of Computer

Year:2023-24 (ODD Sem)

Information Retrieval MIDTERM TEST(Solutions)

Q.1)		(Attempt any five of the following)	Marks (20)
	a)	**Critique the limitations and challenges of information retrieval systems in handling unstructured or semi-structured data.** **Ans:** Information retrieval systems play a crucial role in helping users access and retrieve relevant information from large datasets, but they face several limitations and challenges when dealing with unstructured or semi-structured data. Here are some key critiques: 1. **Lack of Structure:** Unstructured data, such as text documents or multimedia content, lacks a predefined structure, making it challenging for traditional information retrieval systems designed for structured data. 2. **Ambiguity and Context:** Unstructured data often contains ambiguity and context-dependent information, making it difficult for systems to accurately interpret and retrieve relevant content. 3. **Scalability:** As the volume of data grows, information retrieval systems may struggle to maintain efficiency and performance. 4. **Multimedia Content:** Handling multimedia content, such as images, audio, and video, is a significant challenge for information retrieval systems. 5. **Data Heterogeneity:** Unstructured and semi-structured data can come in various forms and sources, leading to data heterogeneity. 6. **Information Extraction and NLP:** Extracting relevant information from unstructured text data requires natural language processing (NLP) techniques, which can be computationally intensive.	2M

	b)	**Evaluate the various performance measures employed in assessing the effectiveness of Search Engines.** **Ans:** 1. **Relevance:** Relevance is a fundamental measure of a search engine's effectiveness. It assesses how well the search engine's results match the user's query. Relevance can be measured using metrics like Precision (the proportion of retrieved documents that are relevant) and Recall (the proportion of relevant documents retrieved). 2. **Recall:** Recall measures the ability of a search engine to retrieve all relevant documents from the database. It is essential to ensure that the search engine doesn't miss important results. Recall is typically evaluated by comparing the number of relevant documents retrieved to the total number of relevant documents in the collection. 3. **Precision:** Precision measures the accuracy of the search results. It assesses the proportion of retrieved documents that are actually relevant. A high precision indicates that the search engine is returning mostly relevant results.	2M
	c)	**With the help of architecture diagram show the interconnectedness of the components in an Information Retrieval system and how they collaborate to retrieve information effectively.** **Ans:**	2M

			
	d)	**If an IR system returns 6 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. Calculate the precision and recall of the system on this search?** **Ans:** Precision (P) measures the accuracy of the retrieved relevant documents: P = (Number of Relevant Documents Retrieved) / (Total Number of Documents Retrieved) Recall (R) measures the ability to retrieve all relevant documents: R = (Number of Relevant Documents Retrieved) / (Total Number of Relevant Documents in the Collection) Given the information provided: ● The IR system returns 6 relevant documents. ● The IR system returns 10 non-relevant documents. ● There are a total of 20 relevant documents in the collection. Precision (P) = (6) / (6 + 10) = 6 / 16 = 0.375 Recall (R) = (6) / (20) = 6 / 20 = 0.3 So, the precision of the system is 0.375 (or 37.5%), and the recall is 0.3 (or 30%).	2M

	e)	**Identify limitations in the Boolean model's ability to handle complex queries or large datasets?** **Ans:** ● **No Partial Matching:** Boolean queries require an exact match of terms. This can be problematic when dealing with documents containing variations or partial matches of query terms. ● **Lack of Relevance Ranking:** The Boolean model doesn't rank results by relevance. It treats all documents that match the query equally. This can lead to situations where numerous results are retrieved, but they may not be ordered by their importance or relevance to the user's needs. ● **No Concept of Document Relevance:** The Boolean model doesn't consider the concept of document relevance or the context in which terms appear. It treats all terms as equally important without considering their importance within documents. ● **Precision and Recall Trade-off:** Boolean queries can be overly restrictive, leading to low recall (missing relevant documents) or overly permissive, leading to low precision (including many irrelevant documents). Striking the right balance can be challenging. ● **Scalability:** The Boolean model can become impractical for large datasets. As the dataset grows, the number of possible combinations of Boolean queries can become enormous, leading to slow and inefficient searches.	2M
	f)	**How to improve query formulation through query expansion and term reweighting.** **Ans:** Query Expansion: Query expansion involves adding related terms or synonyms to your original query to broaden its scope. Here's how to do it: ● Thesauri and WordNet: Consult thesauri and lexical databases like WordNet to find synonyms and related terms for your query keywords. ● Concept Hierarchies: Explore concept hierarchies to identify broader or narrower terms that may be relevant. ● Semantic Analysis: Utilize semantic analysis tools or techniques to discover terms related to the query concepts. ● Contextual Expansion: Analyze the context of your query to identify terms that frequently co-occur with your keywords. 4. Term Reweighting: Term reweighting involves assigning different weights to terms in your query to emphasize their importance. Here's how to do it:	2M

		• TF-IDF (Term Frequency-Inverse Document Frequency): Calculate the TF-IDF scores for each term in your query. This method gives higher weight to terms that are rare in the document collection but frequent in the query. Relevance Feedback: Incorporate feedback from users or relevance assessments to reweight query terms based on their importance in relevant documents.	
Q.2)	a)	**Compare and contrast the vector space, and probabilistic information retrieval models. Can you describe a real-world situation where a probabilistic retrieval model would be beneficial?** **Ans:** Vector space models (VSM) and probabilistic information retrieval models (PIR) are two different approaches to information retrieval, each with its own characteristics and strengths. Vector Space Model (VSM): • Representation: VSM represents documents and queries as vectors in a high-dimensional space, where each dimension corresponds to a unique term. The vector's components are typically based on term frequency-inverse document frequency (TF-IDF) values. • Scoring: VSM calculates the similarity between a query vector and document vectors using measures like cosine similarity. Documents with higher cosine similarity scores are considered more relevant to the query. • Retrieval Strategy: VSM retrieves documents that match the query terms based on similarity scores. It often uses the "bag of words" assumption, ignoring term order and semantics. • Query Expansion: VSM can be enhanced with query expansion techniques, such as relevance feedback, to improve retrieval effectiveness. Probabilistic Information Retrieval Model (PIR): • Representation: PIR models the probability of relevance between a document and a query. It estimates the likelihood that a document is relevant to a given query. • Scoring: PIR calculates a relevance score for each document-query pair based on the probability of relevance. Common PIR models include the Binary Independence Model (BIM) and the Okapi BM25 model. • Retrieval Strategy: PIR models use probabilistic ranking techniques to rank documents by their estimated probabilities of relevance. These models often take into account factors like term frequency, document length, and document prior probabilities.	5M

		• Query Expansion: While PIR models can incorporate feedback from users to update the probability estimates, they typically rely more on statistical relationships between terms. Comparison: • Representation: VSM represents documents and queries as vectors, while PIR models focus on estimating probabilities of relevance. • Scoring: VSM relies on similarity measures, while PIR models use probabilistic ranking techniques. • Retrieval Strategy: VSM focuses on similarity-based retrieval, whereas PIR models emphasize probabilistic ranking. Real-World Situation for PIR: • A real-world situation where a probabilistic retrieval model would be beneficial is in a large-scale search engine, such as Google or Bing. • PIR models are well-suited for large-scale web search scenarios due to their scalability, adaptability, and ability to handle diverse user intents and statistical relevance patterns in documents.	
		OR	
	b)	**Draw the inverted index that would be built for the following document collection.** **Doc 1 one fish, two fish** **Doc 2 red fish blue fish in the hat** **Doc 3 cat in the red hat** **Doc 4 blue eggs and red ham** **Find out Term Document Incidence Matrix for the expression - fish and red not hat** **Ans:** **Step 1: Create the Inverted Index**	5M

An inverted index is a data structure that stores a mapping of terms to the documents where they appear. Here's the inverted index for the given document collection:

Term	Documents
and	Doc 2, Doc 4
blue	Doc 2, Doc 4
cat	Doc 3
eggs	Doc 4
fish	Doc 1, Doc 2
ham	Doc 4
hat	Doc 2, Doc 3
in	Doc 2, Doc 3
one	Doc 1
red	Doc 2, Doc 3, Doc 4
the	Doc 2, Doc 3
two	Doc 1

Step 2: Create the Term Document Incidence Matrix

Now, let's create a Term Document Incidence Matrix for the expression "fish and red not hat." We will use binary values to indicate term presence (1) or absence (0) in each document.

- Terms: fish, and, red, not, hat
- Documents: Doc 1, Doc 2, Doc 3, Doc 4

The matrix will look like this:

Q.3)	a)	Consider a reference collection and its set of example information requests. If q is the information request and a set $R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{50}, d_{70}, d_{80}, d_{120}\}$ Now, consider that a new retrieval algorithm has been designed and has been evaluated for information request q returns, ranking of the documents in the answer set as. 1. $\underline{d_{120}}$ 6. $\underline{d_9}$ 11. d_{38} 2. d_{84} 7. d_{58} 12. d_{48} 3. $\underline{d_{50}}$ 8. d_{129} 13. d_{230} 4. d_6 9. d_{143} 14. d_{113} 5. d_8 10. $\underline{d_{25}}$ 15. $\underline{d_3}$ The documents that are relevant to the query q are underlined. Calculate precision and recall for the documents that are relevant to the query q.	5M
		OR	
	b)	How would you implement a relevance feedback mechanism in an information retrieval system to enhance search results? Ans:	5M

	Doc 1	Doc 2	Doc 3	Doc 4
fish	1	1	0	0
and	0	1	0	1
red	0	1	1	1
not	0	0	0	1
hat	0	1	1	0

In this Term Document Incidence Matrix:

- "1" indicates the presence of the term in the corresponding document.
- "0" indicates the absence of the term in the corresponding document.

For example, "fish" is present in Doc 1 and Doc 2 but absent in Doc 3 and Doc 4.

"and" is present in Doc 2 and Doc 4 but absent in Doc 1 and Doc 3, and so on.

Implementing a relevance feedback mechanism in an information retrieval system is a valuable approach to enhancing search results by involving users in the feedback process. Relevance feedback allows users to provide feedback on the retrieved documents, and the system uses this feedback to refine subsequent searches. Here's how you can implement a relevance feedback mechanism:

1. User Interaction:
When a user performs a search query, retrieve a set of documents that are potentially relevant to the query.
2. Present Search Results:
Present the retrieved documents to the user in a user-friendly interface. Include features like document titles, snippets, and possibly metadata.
3. Gather User Feedback:
Allow users to interact with the search results by marking documents as relevant or non-relevant to their information needs.
Provide mechanisms like checkboxes, thumbs-up/thumbs-down icons, or star ratings to collect user feedback.
4. Feedback Collection:
Collect and record the user feedback data, associating it with specific queries and documents. You'll need to keep track of which documents were marked as relevant and which were not.
5. Relevance Model:
Develop a relevance model based on the user feedback. Common approaches include:
Probabilistic Models: Incorporate the feedback data into probabilistic models like the Relevance Model (RM) or the Markov Random Field Model (MRF) to estimate the relevance of terms and documents.
6. Query Reformulation:
Modify the user's original query using the relevance model. This query reformulation aims to improve the retrieval of relevant documents based on the feedback.
7. Execute Refined Query:
Execute the refined query on the document collection to retrieve a new set of documents.
8. Present New Results:
Display the newly retrieved documents along with the original results, clearly indicating which documents were retrieved based on the user feedback.
9. Iteration:
Allow users to continue the feedback process iteratively. Users can provide feedback on the refined search results, and the system can further refine the query and results based on this feedback.
10. Evaluation and Monitoring:
Continuously evaluate the effectiveness of the relevance feedback mechanism using metrics like precision, recall, and user satisfaction.

		Monitor user interactions and feedback to detect changes in information needs and adapt the system accordingly. **11. User Control:** Provide users with control over the feedback process. Allow them to opt in or out of relevance feedback and to clear or modify their feedback data. **12. Privacy and Security:** Implement robust privacy and security measures to protect user feedback data and ensure user anonymity when necessary. Implementing a relevance feedback mechanism requires careful design, monitoring, and user engagement. By incorporating user feedback into the search process, you can significantly enhance the relevance and quality of search results over time.	