CSE508: Information Retrieval **End-term Examinations [10 May, 2024]**
Name: **Roll Number:**

Instructions:

1. It is a closed-book examination. Exam duration is 2 hours.
2. Write name/ roll number/ project group# both on the answer sheet and questions paper.
3. All questions are mandatory. Total marks are 100.
4. Calculators are not allowed.
5. There is no negative marking.
6. For MCQs, more than one choice could be correct. There are no partial marks.

Section 1: MCQs/Fill in the Blanks [20 x 2 marks each]

Question 1: Which type of data reflects the multifaceted nature of information encountered in various domains?

A) Structured data only
B) Text documents and images
C) Audio recordings and videos
D) All of the above

Answer: D) All of the above

Question 2: Between HITS and PageRank which of the followings are True:

A) HITS emphasizes mutual reinforcement between authority and hub webpages.
B) PageRank attempts to capture the distinction between hubs and authorities.
C) PageRank is a Link analysis algorithm based on a random surfer model.
D) PageRank and HITS only operate on a small subgraph from the web graph.

Answer: A), C)

Question 3: What is a limitation associated with the scalability of information retrieval systems?

A) They are only suitable for small datasets
B) As data volume increases, retrieval performance may degrade
C) Large datasets are easier to manage than small ones
D) Scaling up an information retrieval system has no impact on performance

Answer: B) As data volume increases, retrieval performance may degrade

Question 4: Which of the following statements accurately describes the BSBI technique?

A) BSBI builds the entire index in memory before writing to disk
B) BSBI processes documents individually without sorting
C) BSBI sorts and merges blocks of postings during index construction
D) BSBI is primarily used for retrieval of unstructured data

Answer: C) BSBI sorts and merges blocks of postings during index construction

Question 5: In single-pass in-memory indexing, when is the index constructed?

A) After multiple passes through the data
B) During a single pass through the data

C) Before the data is loaded into memory
D) After sorting the entire dataset

Answer: B) During a single pass through the data

Question 6: How can versioning aid in managing updates and changes in information retrieval systems?
A) By preventing any updates to the system
B) By automatically reverting to the previous version of content
C) By maintaining historical versions of documents to track changes over time
D) By delaying updates indefinitely

Answer: C) By maintaining historical versions of documents to track changes over time

Question 7: Which of the following are TRUE?
A) Web search is an example of a precision-critical task.
B) Legal and patent search is an example of a recall-critical task.
C) Legal and patent search is an example of a precision-critical task.
D) Web search is an example of a recall-critical task.

Answer: A, B

Question 8: What accurately describes the distinction between search and retrieval?
A) Search involves querying databases, while retrieval involves data insertion.
B) Retrieval is synonymous with data extraction, while search encompasses data organization.
C) Search is a user-driven activity to locate information, while retrieval involves data processing and delivery.
D) Retrieval refers to web browsing, while search is specific to local data systems.

Answers: C) Search is a user-driven activity to locate information, while retrieval involves data processing and delivery.

Question 9: Which technique is used to improve retrieval speed by pre-computing similarities between terms and documents?
A) Caching
B) Parallel processing
C) Index pruning
D) Term-document matrix

Answer: D) Term-document matrix

Question 10: Statement: An inverted index maps documents to the terms they contain.
A) True
B) False
Answer: B) False
Explanation: Inverted index maps terms to the documents that contain them, not the other way around.

Question 11: Which statements correctly differentiate between search and retrieval in the context of information systems?

A) Search involves finding relevant information based on user queries.
B) Retrieval refers to the process of storing and organizing data.
C) Retrieval focuses on accessing and delivering stored information.
D) Search is primarily concerned with indexing and categorizing data.

Correct Answers:
A) Search involves finding relevant information based on user queries.
C) Retrieval focuses on accessing and delivering stored information.

Question 12: Statement: Inverted indexes facilitate efficient boolean operations (AND, OR, NOT) on terms during query processing.
A) True
B) False

Answer: True
Explanation: Inverted indexes allow quick execution of boolean operations on terms, making them suitable for complex query processing.

Question 13: How does incremental indexing contribute to efficient Information Retrieval?
A) By compressing index data for faster access
B) By periodically rebuilding the entire index from scratch
C) By updating the index incrementally with changes to the document collection
D) By restricting access to static documents only

Answer: C) By updating the index incrementally with changes to the document collection

Question 14: What characterizes semi-structured data?
A) Data that conforms to a rigid schema
B) Data that lacks any organization or format
C) Data that has some organizational structure but does not fit neatly into relational databases
D) Data that is exclusively textual

Answer: C) Data that has some organizational structure but does not fit neatly into relational databases

Question 15: What is the purpose of feature extraction in multimedia data indexing?
A) To compress multimedia files for efficient storage
B) To convert multimedia data into text documents
C) To identify and extract meaningful characteristics from multimedia content
D) To remove irrelevant data from multimedia files

Answer: C) To identify and extract meaningful characteristics from multimedia content

Question 16: Document Frequency in dictionary help with boolean query _____.
Answer: Optimization.

Question 17: There is an _____ gap between the information need by users and the corresponding query used by users in information retrieval.
Answer: Intent.

Question 18: Term frequency-Inverse document frequency (TF-IDF) assigns higher weights to terms that are _____ in a document but less common across the entire document collection.

Answer: frequent or significant

Question 19: Often Probability Ranking Principle (PRP) and Binary Independence Model (BIM) used in _____ to solve IR problems.

Answer: conjunction

Question 20: A crawler must be _____ and _____.

Answer: robust and polite.

Section 2: Short Answers [10 x 4 marks each]

Question 21: What are Requirements for real-time search, say X or Instagram?

Answer:

- Low latency, high throughput query evaluation
- High ingestion rate and immediate data availability
- Concurrent reads and writes of the index
- Dominance of temporal signal

Question 22: What are the two key ideas in SPIMI: Single-pass in-memory indexing?

Answer:

- Key idea 1: Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.
- Key idea 2: Don't sort. Accumulate postings in postings lists as they occur.
- With these two ideas we can generate a complete inverted index for each block.
- These separate indexes can then be merged into one big index.

Question 23: What are the benefits of positional indexing over traditional indexing?

Answer:

Supports Phrase Queries: One significant benefit of positional indexing is its ability to efficiently handle phrase queries. With positional indexing, each term in the index is associated not only with the document ID but also with the position(s) of the term within the document. This allows the system to retrieve documents where specific terms appear in a specified sequence or proximity, enabling accurate and context-aware retrieval.

Enhances Relevance Ranking: Positional indexing can improve relevance ranking by considering the proximity and order of terms within documents. For example, when a user searches for a multi-word query or a phrase, positional indexing can prioritize documents where the query terms appear closely together or in a specified order, leading to more relevant search results.

Question 24: What is the Permuterm index? How is it helpful? What overhead does Permuterm index add? Explain with an example.

Answer: It is helpful in handling the wildcard queries. Remember the example of pro*cent from the class for a more elaborative answer. One TA can complete this answer.

Question 25: What are two types of spelling errors? Explain with examples.

Answer:

Non-word Errors

graffe → giraffe

Real-word Errors

Typographical errors

- three → there

Cognitive Errors (homophones)

- piece → peace,
- too → two
- your → you're

Question 26: Explain the cases when a user's information need is fully, partially, and not at all fulfilled for the corresponding query used by the user for the information need. Explain with an example?

Answer: One TA can complete the answer. Information need of a user is Fully fulfilled when a query represents 100% information needed by users, partially when it represents just a part of it, and not at all when query is totally different from what the user information need is.

Question 27: What is the goal of Okapi BM25? Explain with an example.

Answer: Goal of BM25 is to be sensitive to term frequency and document length while not adding too many parameters. Add example.

Question 28: What is the difference between safe and non-safe ranking in IR? provide two examples for each.

Answer: The terminology "safe ranking" is used for methods that guarantee that the K docs returned are the K absolute highest scoring documents. TAs can add a few examples for safe and non-safe ranking.

Question 29: What is robots.txt? How does it help in crawling? Explain with examples.

Answer: protocol for giving spiders (“robots”) limited access to a website, originally from 1994
www.robotstxt.org/robotstxt.html

Website announces its request on what can(not) be crawled

For a server, create a file /robots.txt

This file specifies access restrictions

Question 30: In logarithmic merge, where n=4. We have 31 tokens to be processed. Which indexes including auxiliary indexes (Z0, I0, I1, I2, I3, I4) would be in use after all the tokens are used? See the table for representation. (consider Z0 < n).

#token	Z_0	I_0	I_1	I_2	I_3	I_4
3	1	0	0	0	0	0
4	0	1	0	0	0	0

Answer:

1, 1, 1, 1, 0, 0

Explanation:

n=4

First 3 tokens - Z_0 (<n)

4 tokens - I_0 (n)

7 tokens - Z_0, I_0

8 tokens - I_1 (2n)

15 tokens - Z_0, I_0, I_1

16 tokens - I_2 (4n)

31 tokens - Z_0, I_0, I_1, I_2

Section 3: Descriptive/Numerical Answers [4 x 5 marks each]

Question 31: What is the need of probabilistic information retrieval techniques? What is the limitation of boolean information retrieval and how probabilistic information retrieval techniques solve this problem. Explain with an example.

Question 32: Consider 4 documents D1, D2, D3, D4.

For any given query Q, the relevance for the documents is: D1: 3, D2: 1, D3: 2, D4: 1.

An IR System retrieved the documents for query Q in the following order: D4, D1, D3, D2.

A Ranking Function returns documents for query Q in the order D1, D3, D4, D2.

Compute NDCG for the IR system (NDCG_IRSystem) and Ranking function (NDCG_RankingFunction). (Use log base 2 for computations)

Answer:

NDCG_IRSystem = 0.939, NDCG_RankingFunction = 1

Explanation:

DCG_ideal = $3 + 2/\log_2 2 + 1/\log_2 3 + 1/\log_2 4 = 6.1309$

DCG_IRSystem = $1 + 3/\log_2 2 + 2/\log_2 3 + 1/\log_2 4 = 5.7618$
NDCG_IRSystem = 5.7618/6.1309 = 0.9397

DCG_RankingFunction = $3 + 2/\log_2 2 + 1/\log_2 3 + 1/\log_2 4$ = DCG_ideal
NDCG_RankingFunction = 1

Question 33: Imagine you are tasked with designing an advanced information retrieval system for a large e-commerce platform. Describe the key components and considerations involved in building this system, highlighting the challenges and solutions for effective search functionality.

Answer Outline:

To design an advanced information retrieval system for a large e-commerce platform, several key components and considerations need to be addressed:

1. Data Collection and Indexing:
    - Implement web crawling techniques to gather product information, including titles, descriptions, categories, attributes, and customer reviews.
    - Use efficient indexing methods (e.g., inverted index) to create a searchable index of the product data, enabling fast retrieval based on user queries.
2. Query Processing and Understanding:
    - Develop robust query processing mechanisms to interpret user search queries, considering synonyms, spelling variations, and user intent.
    - Use natural language processing (NLP) techniques to analyze and understand user queries, improving relevance and accuracy of search results.
3. Relevance Ranking and Personalization:
    - Implement advanced ranking algorithms (e.g., TF-IDF, BM25) to prioritize search results based on relevance to the user's query.
    - Incorporate user preferences, browsing history, and behavioral data to personalize search results and recommendations, enhancing user experience.
4. Faceted Search and Filtering:
    - Enable faceted search and filtering based on product attributes (e.g., price range, brand, size) to refine search results and assist users in finding specific items.

- Implement dynamic filtering options to adapt to user selections and provide real-time feedback on available choices.

5. Performance Optimization and Scalability:
    - Design the system to handle large-scale data processing and real-time updates, ensuring scalability and responsiveness during peak traffic periods.
    - Utilize caching mechanisms and distributed computing frameworks to optimize query performance and minimize response times.
6. User Interface and Experience:
    - Develop intuitive and responsive user interfaces (UI/UX) for search functionality, incorporating autocomplete, spell correction, and visual feedback to guide user interactions.
    - Implement rich snippet integration to display relevant product information directly within search results, enhancing discoverability and engagement.
7. Monitoring and Evaluation:
    - Integrate analytics and monitoring tools to track search metrics (e.g., click-through rate, conversion rate) and user interactions, enabling continuous optimization of the retrieval system.
    - Conduct regular evaluations and A/B testing to assess the effectiveness of search algorithms and user interface enhancements, iterating based on feedback and performance metrics.

Challenges such as data variability, query complexity, and maintaining freshness of product information must be addressed through robust architecture design, algorithmic innovation, and continuous refinement of the information retrieval system. By leveraging advanced technologies and methodologies, the e-commerce platform can deliver a seamless and personalized search experience to users, driving customer satisfaction and conversion rates.

Question 34: Suppose you are evaluating a search engine's performance for a specific query "machine learning" against a collection of documents related to artificial intelligence. You manually assess the relevance of retrieved documents and use the following data to compute precision and recall:

- Total Relevant Documents in Collection (based on manual assessment): 50
- Retrieved Documents for Query "machine learning":
    - Document A: Relevant
    - Document B: Relevant
    - Document C: Non-Relevant
    - Document D: Relevant
    - Document E: Non-Relevant
    - Document F: Relevant
    - Document G: Non-Relevant
    - Document H: Non-Relevant

- Document I: Relevant
- Document J: Relevant

Answer:

Let's calculate precision (P) and recall (R) based on the provided data:

1. Identify Relevant and Non-Relevant Documents in Retrieval Result:
. Relevant Documents (Retrieved and Correctly Classified): A, B, D, F, I, J (Total =6)
. Non-Relevant Documents (Retrieved but Incorrectly Classified): C, E, G, H (Total =4)
2. Compute Precision (P) and Recall (R):

. Precision (P):
Precision measures the proportion of retrieved documents that are relevant.P =

no. of relevant retrieved doc / total no, of relevant document = 6/10 = 0.6

· Recall (R):
Recall measures the proportion of relevant documents that are retrieved.R =
Number of Relevant Retrieved Documents/Total Number of Relevant Documents in Collection = 6/50 = 0.12

Therefore, the precision (P) is 0.6 (or 60%) and the recall (R) is 0.12 (or 12%) for the query "machine
learning" against the collection of documents related to artificial intelligence.

This calculation indicates that out of the retrieved documents, 60% are relevant (precision), and out of all relevant documents in the collection, only 12% were successfully retrieved by the search engine (recall). These metrics provide insights into the search engine's effectiveness in returning relevant results of search query.