\

CSE508: Information Retrieval **Quiz [29 April 2024]**

Name: **Roll Number:**

Instructions:

1. It is a close book examination.
2. Please write name, roll number, and project group number in the answer sheet.
3. Exam duration is 30 minutes.
4. All questions are mandatory.
5. Calculators are not allowed.
6. There is no negative marking.
7. Total marks are 20.
8. There could be more than one correct answer in MCQs. No partial marks will be given.

Section 1: MCQs [1 mark each]

Question 1: Mark all correct statements.

A. F-Measure provides a single score that balances both the concerns of precision and recall in one number.
B. F-Measure provides a single score that gives more importance to precision than recall.
C. F-Measure provides a single score that gives more importance to recall than precision.
D. We can accurately compute F-Measure for any search engine such as Google.

Answer: A

Question 2: Select the FALSE statement(s):

A. Hub score of a page P is calculated as the sum of the authority scores of all pages that point to P.
B. Authority score of a page P is calculated as the sum of the authority scores of all pages that point to P.
C. Hub score of a page P is calculated as the sum of the authority scores of all pages that P points to.
D. Authority score of a page P is calculated as the sum of the hub scores of all pages that P points to.

Answer: A, B, D

Hub score of a page P is calculated as the sum of the authority scores of all pages that P points to. Hence option C is correct.
Explanation:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$

Question 3: Which of the following is True about PageRank?

A. Higher PageRank scores imply the page is less important
B. PageRank scores of pages that have a link to a given page influence its PageRank score
C. The higher the in-degree of links of a page, the higher is its long term visit rate (PageRank score)
D. The higher the out-degree of links of a page, the higher is its long term visit rate (PageRank score)

Answer: B, C
Explanation:
Higher the PR score, the higher the importance of the page
PageRank score of voters influences the scores of voted page
Higher in-degree implies a higher PR score

Question 4: Mark all correct statements.

A. Precision quantifies the number of positive class predictions that actually belong to the positive class.
B. Precision quantifies the number of positive class predictions made out of all positive examples in the dataset.
C. Recall quantifies the number of positive class predictions that actually belong to the positive class.
D. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

Answer: A, D

Question 5: In the pagerank algorithm, the process of jumping to a random node from a dead-end with a certain probability is called:

A. Jump
B. Transfer
C. Teleport
D. Walk

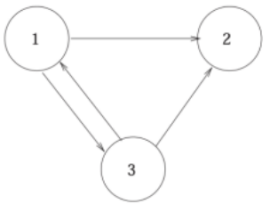
Answer: C

Question 6: Which of the following is TRUE about Recall?

A. Recall is a non-increasing function of the number of relevant docs retrieved.
B. Recall is a non-decreasing function of the number of relevant docs retrieved.
C. A system that returns all relevant docs has 100% recall.
D. A system cannot have 100% recall.

Answer: B, C

Question 7: Assuming no teleportation, which node is expected to have the highest pagerank in the given graph?



A. 1
B. 2
C. 3
D. Insufficient Information

Ans: 2
Node 2 has no outgoing links due to which it will absorb all the pagerank during power iteration (2 is a dead node)

Question 8: 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 precision-critical task.
C. Web search is an example of a recall-critical task.
D. Legal and patent search is an example of a recall-critical task.

Answer: A, D

Question 9: Given the options mark which are not correct about PageRank Algorithm.

A. The idea behind this algorithm is that pages visited more often in this walk are less important.
B. The idea behind this algorithm is that pages visited more often in this walk are more important.
C. In the PageRank algorithm the surfer proceeds in a random walk from node to node.
D. If the current location of the surfer has no out-links teleport operation is used.
E. If the current location of the surfer has no out-links, random walks are used.

Answer: A), E)

Question 10: What's true about Webgraph?

A. It uses Adjacency lists.
B. WebGraph provides a way to store efficiently the URLs of a Web graph.
C. WebGraph is a framework that provides simple methods to manage very large graphs.
D. Algorithms For compressing Web graphs that exploit gap compression.

Answer: A, C, D

Section 2: Descriptive and Numerical Questions [5 marks each]

Question 11: How do distributed word representations enhance the effectiveness of information retrieval systems? Describe with an example.

Answer: The following is one sample answer for one example. Evaluate students' answers around the example they give. Ideally, they should cover most possible benefits such as Semantic Understanding, Contextual Similarity, Dimensionality Reduction, Improved Ranking, Handling Polysemy, Learning from Large Text Corpora, etc.

Distributed word representations enhance the effectiveness of information retrieval systems by capturing semantic relationships between words and enabling more nuanced understanding of text. These representations, typically learned from large text corpora using techniques like Word2Vec or GloVe, encode words into dense vector spaces where similar words have similar vector representations. This improvement is particularly beneficial in information retrieval, as it allows systems to grasp context and meaning beyond simple keyword matching.

For example, consider a scenario where a user queries "computer security measures". In a traditional keyword-based retrieval system, the search might primarily look for documents containing exact matches of these words. However, with distributed word representations, the system can leverage semantic similarities encoded in word embeddings.

Suppose the word embeddings reveal that "computer" is closely related to "technology", "cyber", and "information", while "security" is linked to "protection", "safety", and "defense". By analyzing these semantic associations, the retrieval system can expand the scope of the search to include documents discussing topics related to "technology protection" or "cybersecurity measures", even if those specific terms are not present in the query.

Furthermore, distributed representations enable handling of synonyms and related terms effectively. For instance, if a document discusses "network security" but doesn't use the exact phrase "computer security measures", the system can still recognize its relevance based on the similarity between the word embeddings of "network security" and "computer security".

In essence, distributed word representations facilitate a more sophisticated understanding of natural language semantics, allowing information retrieval systems to retrieve documents that align with the underlying meaning and context of user queries rather than just exact keyword matches. This capability significantly enhances the effectiveness and accuracy of retrieving relevant information from large text collections.

Question 12: How can incorporating personalized search and recommendation systems improve information retrieval in e-learning platforms? Discuss specific strategies and

technologies that can be utilized to enhance the efficiency and effectiveness of content retrieval for individual learners.

Answer: The following is one sample answer for one example. Evaluate students' answers around the example they give.

Personalized search and recommendation systems offer significant potential for improving information retrieval in e-learning platforms by tailoring content delivery to meet individual learner needs and preferences. Here's an exploration of strategies and technologies that can be employed:

1. Personalized Content Recommendations: By leveraging machine learning algorithms such as collaborative filtering or content-based filtering, e-learning platforms can recommend courses, modules, or resources based on a learner's past interactions, preferences, and performance. For example, a platform like Netflix utilizes collaborative filtering to recommend movies based on similar user preferences.
2. Adaptive Learning Paths: Implementing adaptive learning technologies allows platforms to dynamically adjust the sequence and difficulty level of content based on a learner's progress and comprehension. Reinforcement learning algorithms can optimize learning paths to maximize engagement and knowledge retention.
3. Semantic Search and Natural Language Processing (NLP): Integrating NLP models like BERT or word embeddings into search functionalities enables more context-aware and semantic-driven queries. Learners can use natural language to express their information needs, and the system can retrieve relevant content accurately. For instance, Google Search uses BERT to better understand the context and intent behind search queries.
4. User Behavior Analysis: By analyzing user behavior data such as search history, browsing patterns, and time spent on specific topics, platforms can gain insights into individual learning preferences and adapt information retrieval strategies accordingly. Clustering algorithms can group learners with similar behavior profiles to enhance recommendations.
5. Context-Aware Retrieval: Incorporating contextual information such as location, time of day, or device type can enhance the relevance of retrieved content. Context-aware retrieval ensures that learners receive content that is most suitable for their current learning context.
6. Feedback Loop and Continuous Improvement: Establishing a feedback loop where learners provide explicit feedback (e.g., ratings, comments) or implicit feedback (e.g., clicks, dwell time) on recommended content helps refine recommendation algorithms over time. Platforms can continuously improve information retrieval based on real-time user interactions.
7. Multi-modal Learning: Integrating multiple modalities such as text, images, videos, and interactive simulations enables richer content representation and more diverse retrieval options. Deep learning architectures like convolutional neural networks (CNNs) and transformer models can process multi-modal data for improved content understanding and retrieval.
8. Ethical Considerations and User Privacy: While implementing personalized search and recommendation systems, it's crucial to address ethical considerations such as transparency, fairness, and user privacy. Platforms should provide transparency on how user data is utilized and ensure fairness in content recommendations to avoid reinforcing biases.

In conclusion, incorporating personalized search and recommendation systems in e-learning platforms can significantly enhance the efficiency and effectiveness of information retrieval, leading to a more engaging and tailored learning experience for individual learners. The integration of machine learning, NLP, user behavior analysis, and multi-modal learning technologies plays a pivotal role in optimizing content discovery and delivery in e-learning environments.