Mid-term Examinations
CSE508: Information Retrieval

Date: 26 Feb 2023
Time: 10-11 am

Name:

Roll Number:

Instructions

- It is a close book examination
- All questions are mandatory
- Calculator is not allowed
- There is no negative marking
- MCQ questions (2 marks each) may have more than one correct answers
- Short and long descriptive answers are of 3 and 5 marks, respectively
- There are total 8 questions (3 MCQs, 3 short answers, and 2 long answers)
- Maximum score is 25 marks [3 x 2 + 3 x 3 + 2 x 5]

Section A: MCQs [2 marks each]

Q1) Which of the following cannot be input for an information retrieval system.

A. Text
B. Audio
C. Image
D. Gesture
E. None of the above

Answer) E. None of the above

Q2) Which of the following is true for the query, mercy AND Caesar AND NOT Calpurnia

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

A. 101100
B. 100111
C. 110011
D. 100011

Answer) B 100111

Q3) Mark all True sentences.

A. Traditional Inverted Indexes are more powerful than Positional Inverted Indexes.
B. Positional Inverted Indexes are more powerful than Traditional Inverted Indexes.
C. Traditional Inverted Indexes require more computation than Positional Inverted I.
D. Positional Inverted Indexes require more computation than Traditional Inverted I.

Answer) B, D

Section B: Short Answers [3 marks each]

Q4) Describe Information Search, Information Retrieval, and Data Retrieval with an example.

Answer) Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

Difference Between Information Retrieval and **Data Retrieval**.

Information Retrieval	**Data Retrieval**
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

Q5) List the issues/challenges in the information retrieval system.

Answer) A few points are as follows
• 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 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. JEC/DEPT.OF.CSE JEC/CSE/QB/IV YR /IR

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

Q6) What are the benefits of creating an inverted index for indexing documents in an information retrieval system? Explain with an example.

Answer) If we index all tokens for a given document then the index will become too huge and would be unmanageable, similar to the term-incident matrix. Inverted indexes help us to keep track of all documents for a given token which ease the process of finding documents for a given query.

Section C: Detailed Answers [5 marks each]

Q7) What do you mean by web co-pilot? Why do we need the same? List down the benefits and limitations of web co-pilot with a real-world example.

Based on the video that we discussed in class
https://www.youtube.com/watch?v=rOeRWRJ16yY&t=1369s

Example, ChatGPT is a web co-pilot for Bing in assisting users with phrasing queries efficiently and conversing to get the relevance feedback.

Q8) Google claims to have a better information retrieval system than Microsoft. For a given query, returns the following 5 documents: d1, d2, d3, d4, and d5, with their relevance scores 0, 2, 1, 4, 2, respectively (a higher score indicates more relevant to the given query). New Bing + OpenAI based system returns the following results for the same query: d4, d1, d2, d3, d5. Old Bing returns the following documents: d1, d2, d4, d3, and d5. Compare all IR systems based on DCG and NDCG metrics. For your reference, DCG is defined as follows: $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$

Additionally, log 3 and log 5 (all on base 2) are 1.59 and 2.32, respectively.

Answer) Compute DCG with the above-mentioned formula. To compute NDCG, consider the ground truth order of DCG, d4, d2, d5, d3, and d1. Divide the DCG value by the perfect DCG value to compute NDCG and compare.

8 $l_2 = 1$ $l_3 = 1.59$ $l_4 = 2$ $l_5 = 2.32$

Google:	New Bing	Old Bing
0	4	0
2	0	2
1	2	4
4	1	1
2	2	2

DCG

$0 + \frac{2}{l_2} + \frac{1}{l_3} + \frac{4}{l_4} + \frac{2}{l_5}$ $\quad$ $4 + \frac{0}{l_2} + \frac{2}{l_3} + \frac{1}{l_4} + \frac{2}{l_5}$ $\quad$ $0 + \frac{2}{l_2} + \frac{4}{l_3} + \frac{1}{l_4} + \frac{2}{l_5}$

5.49 0 ① $\quad$ 6.61 9 ① $\quad$ 5.87 7 ①

IDCG

$4 + \frac{2}{l_2} + \frac{2}{l_3} + \frac{1}{l_4} + \frac{0}{l_5}$

$= 7.75$ 7 ①

nDCG

0.707 $\quad$ 0.85 3 $\quad$ 0.75 2

(1/2) Bing-oper 11 > G > Bing old

1 BD ②

2 B ②

3 E ②

4: isearch vs iRef vs DRef ③

5: IR challenges ③

6: Benefits ranked index ③ eg

7: web co-pilot? need? benefits limit? ⑤