Beijing-Dublin International College

SEMESTER 2 FINAL EXAMINATION - 2023/2024

School of Computer Science

COMP3009J Information Retrieval

Dr. Robert Ross
Assoc. Prof. Neil Hurley
Dr. David Lillis *

Time Allowed: 120 minutes

Instructions for Candidates

Answer Question 1 and any two other questions. Question 1 has 30 marks available. All other questions have 35 marks available.

Instructions for Invigilators

Candidates are allowed to use non-programmable calculators during this examination.

All numeric answers should be given in decimal format, and be correct to 3 places of decimals.

Question 1:

(a) A *modern Information Retrieval pipeline* may include Boolean searches, simple ranking (using BM25, for example) and reranking based on machine learning. Explain why each of these are useful to make an effective Information Retrieval system.

[6 marks]

(b) The *Cranfield Paradigm* is typically used as the basis of evaluation strategies for Information Retrieval. Explain in detail what is meant by this. 05b

The Cranifield Paradigm is a standard evaluation framework for information Retrieval. It involves:

1. A fixed document corpus used for all queries **[6 marks]**
2. A set of standard queries
3. Relevance judgements made by experts, identifying which documents are relevant to each query.

This setup allows IR systems to be evaluated by comparing their retrieved results against known relevant documents

(c) The *information need* of a user is said to have 4 stages. Describe these stages.

1. Visceral Need: actual but not unexpressed
2. Conscious need: within-brain description of the need **[6 marks]**
3. Formalised need: the formal statement of question
4. Compromised need: the question as presented to IR

(d) When *tokenising* text, the natural language that the documents are written in can influence the strategy being used. Discuss three examples of issues that can arise when tokenising languages other than English. 02a

1. Chinese: No space between word; Maybe ambiguous **[6 marks]**
2. Arabic: Written right to left
3. French: L'ensemble one or two token

(e) Explain what is meant by the phrase *Adversarial Information Retrieval*. In the context of web search, show two examples of where this can occur.

[6 marks]

Adversarial Information Retrieval refers to attempts by web content creators to manipulate search engine ranking by violating engine guidelines. **[Total 30 marks]**

- Hidden Text: Inserting keywords in color same with background to deceive crawlers but user can't see.
- Cloaking: Serving different content to Search engine crawlers than human.

Question 2

(i) Incidence matrix is sparse for bigger collection (More 0). Inverted index only store docID where term accually occur. It also enables faster Boolean query processing through sorted posting lists.

(a) This question relates to using postings lists to index a document corpus.

(i) Explain why an *inverted index* is a more suitable data structure for representing a document corpus, compared to a *term-document incidence matrix*.

(ii) Describe in detail how a set of *postings lists* can be created to represent a document corpus. Your answer should include details of the data structures that are used during this process.

(ii) 1. Tokenize each document in to term, recoding (term, docID) pairs;
2. Sort these pair by term and then docID
3. Merge duplicates and construct a dictionary mapping each term. 4. Used Linked list for posting

[10 marks]

(b) This question relates to preprocessing.

(iii) Stemming is the process reduce words to a common root, offten by suffix stripping. Ad: fast, simple, reduce vocabulary size.

(i) What is meant by *stopword removal*, and how can this help to improve the Information Retrieval process?

Lemmatisation is a NLP technique for converting word into lemma. Ad: more accurate. Return a real words. Dis: Slower, Need more text analysis.

(ii) In what way is *Zipf's Law* related to stopword removal?

(iii) Explain what is meant by *stemming* and *lemmatisation*? Give advantages and disadvantages of each.

(i) stopword removal is the process of remove common words from document, because they are unuseful when differentiating doc. Remove them reduce number of term, speed up processing.

(ii) Only a few term are used often, most are used rare. By ranking words by frequently, highly frequent words are identified as stopword.

[10 marks]

(c) Below is a small document collection, containing three documents. Answer the questions that follows. All calculations should be presented in decimal format, and be correct to three decimal places.

Len(q) = 1.6815
Len(d_1) = 1.7321
Len(d_2) = 1.5811
Len(d_3) = 2.2361

SCORES:
d_1 = 1.685
d_2 = 0.219
d_3 = 0

Stopwords: a, an, is, some, the

Document 1: Her doctor gave her some very, very bad news.

Document 2: A no news day is a good news day.

Document 3: An apple a day keeps the doctor away.

(i) Calculate a vector to represent each document, using the TF-IDF weighting system. You should use the stopword list provided, but do not perform stemming or lemmatisation.

(ii) Calculate the cosine similarity for each vector using the query "What is her news?" and show the final ranked list of documents for this query.

(i) Lowercase all term

	her	doctor	gave	very	bad	news	no	day	good	apple	keeps	away
idf	$\log_2 3$	$\log_2 \frac{3}{2}$	$\log_2 3$	$\log_2 3$	$\log_2 3$	$\log_2 \frac{3}{2}$	$\log_2 3$	$\log_2 \frac{3}{2}$	$\log_2 3$	$\log_2 3$	$\log_2 3$	$\log_2 3$
tf												
d_1	1	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0
d_2	0	0	0	0	0	1	$\frac{1}{2}$	1	$\frac{1}{2}$	0	0	0
d_3	0	1	0	0	0	0	0	1	0	1	1	1
d_1	1.585	0.2925	0.7925	1.585	0.7925	0.2925	0	0	0	0	0	0
d_2	0	0	0	0	0	0.585	0.7925	0.585	0.7925	0	0	0
d_3	0	0.585	0	0	0	0	0	0.585	0	1.585	1.585	1.585
q	1.585	0	0	0	0	0.585	0	0	0	0	0	0

[15 marks]

[Total 35 marks]

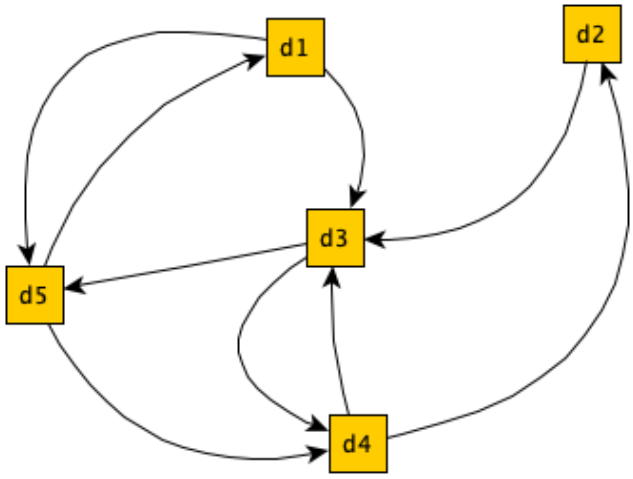
It used to address rank sinks and pages with no backlinks. rank sink which refers to a group of pages that have least one backlink and link to one another, but do not link to anywhere else outside the group. d ensures that not all of a document's PageRank is passed on via its outlinks.

Question 3

(a) This question relates to PageRank.

(i) In the context of PageRank, what is a *damping factor* and why is it important?

(ii) The link structure of some web pages is shown below. There are five web pages shown (d1, d2, d3, d4 and d5), and the arrows show the links between the pages.



Using this structure as an example, describe in detail how a PageRank score is calculated. Use a damping factor of 0.85 and show at least 3 iterations (the first step of assigning the same initial PageRank to each page does not count as an iteration).

Your answer must include a description of the steps you take, in addition to the calculations.

Answers must be given in decimal format and be correct to three places of decimals.

[15 marks]

(b) This is question is related to the evaluation of Information Retrieval systems.

(i) Compare and contrast the P@10, MAP and NDCG evaluation metrics. In your answer, outline any advantages and disadvantages of each. For each metric, suggest a situation where it is more appropriate than the others.

(ii) Explain how the creators of the *bpref* evaluation metric showed that it is suitable for dealing with *incomplete relevance judgments*?

[12 marks]

start initial pageRank with 1

① $R(d_1) = (1-0.85) + 0.85 \times \frac{1}{2} = 0.575$

$R(d_2) = (1-0.85) + 0.85 \times \frac{1}{2} = 0.575$

$R(d_3) = (1-0.85) + 0.85 \times (\frac{1}{2} + 1 + \frac{1}{2}) = 1.85$

$R(d_4) = (1-0.85) + 0.85 \times (\frac{1}{2} + \frac{1}{2}) = 1$

$R(d_5) = (1-0.85) + 0.85 \times (\frac{1}{2} + \frac{1}{2}) = 1$

② $R(d_1) = (1-0.85) + 0.85(\frac{1}{2}) = 0.575$

$R(d_2) = (1-0.85) + 0.85 \times (\frac{1}{2}) = 0.575$

$R(d_3) = (1-0.85) + 0.85(\frac{0.575}{2} + 0.575 + \frac{1}{2}) = 1.308$

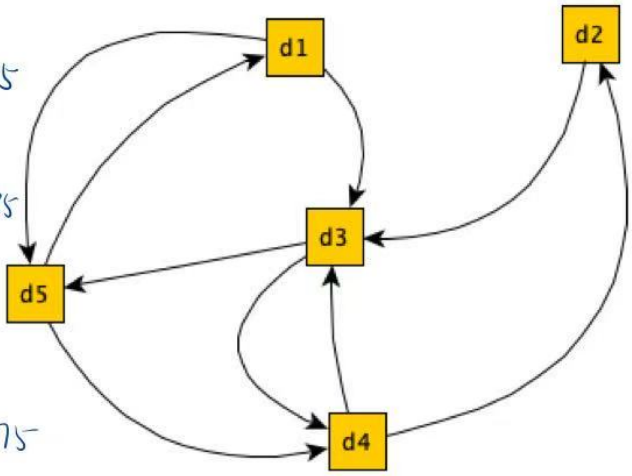
$R(d_4) = (1-0.85) + 0.85(\frac{1.85}{2} + \frac{1}{2}) = 1.361$

$R(d_5) = (1-0.85) + 0.85 \times (\frac{0.575}{2} + \frac{1.85}{2}) = 1.181$

③ $R(d_1) = 0.652$

$R(d_2) = 0.728$

$R(d_3) = 1.461$ $R(d_4) = 1.208$ $R(d_5) = 0.950$



Using this structure as an example, describe in detail how a PageRank score is calculated. Use a damping factor of 0.85 and show at least 3 iterations (the first step of assigning the same initial PageRank to each page does not count as an iteration).

Your answer must include a description of the steps you take, in addition to the calculations.

Answers must be given in decimal format and be correct to three places of decimals.

[15 marks]

(b) This is question is related to the evaluation of Information Retrieval systems.

(c) Below is a set of results that were returned by a search engine in response to a query.

R R N U R U N R U U R R N N U R

Retrieved = d7 d1 d10 d14 d8 d2 d18 d17 d4 d20 d21 d3 d11 d12 d5 d24

Judged Relevant = {d1, d3, d6, d7, d8, d13, d17, d19, d21, d24}

Judged Non-relevant = {d9, d10, d11, d12, d18, d25}

Calculate the following evaluation metrics for this query. Your results should be presented in decimal format, and be correct to three places of decimals:

(i) Precision
(ii) Recall
(iii) Mean Average Precision (MAP)
(iv) bpref

[8 marks]

[Total 35 marks]

(i) Precision $= \frac{|Ret \cap Rel|}{Ret} = \frac{7}{16} = 0.438$

Recall $= \frac{|Ret \cap Rel|}{Rel} = \frac{7}{10} = 0.700$

$$MAP = \frac{P@1 + P@2 + P@5 + P@8 + P@11 + P@12 + P@16}{|rel|}$$

$$= \frac{1 + 1 + 0.6 + 0.5 + 0.4545 + 0.5 + 0.4375}{10}$$

$$= 0.447$$

bpref: $B = \frac{1}{R} \sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{R}$

1+1

$$\frac{0 + 0 + (1 - \frac{1}{10}) + (1 - \frac{2}{10}) + (1 - \frac{2}{10}) + (1 - \frac{2}{10}) + (1 - \frac{4}{10})}{10}$$

$$= 0.405 \quad 0.59$$

Question 4

(a) Describe in detail how you would design a *web crawler* to find information that could be included in the index of an Information Retrieval system. In your answer, include details about what standards the crawler should follow, and situations where accessing information is difficult.

[10 marks]

(b) The table below shows results from four search engines in response to the same query. Each set of results consists of a ranked list of unique document identifiers (DocID), along with the ranking score. Complete the following tasks, showing your workings for each. Answers should be correct to three places of decimals.

(i) Calculate the score that document d16 would have using *CombMNZ*.
(ii) Calculate the score that document d18 would have using *Borda Fuse*.
(iii) Calculate the score that document d10 would have using *CombSum*.

Engine A	
DocID	**Score**
d19	0.999
d23	0.873
d2	0.784
d25	0.733
d10	0.733
d18	0.608
d9	0.507
d16	0.435
d15	0.055

Engine B	
DocID	**Score**
d22	3.710
d24	3.655
d18	3.421
d12	3.411
d10	3.339
d6	2.976
d11	2.780
d20	2.248
d21	2.149
d16	1.717
d9	1.054

Engine C	
DocID	**Score**
d10	98.762
d6	97.898
d2	89.051
d17	66.673
d8	53.929
d3	22.150
d7	10.985

Engine D	
DocID	**Score**
d15	0.974
d13	0.957
d12	0.911
d24	0.871
d22	0.728
d20	0.690
d23	0.626
d10	0.385
d16	0.163
d18	0.146
d7	0.122
d14	0.081
d8	0.020

[10 marks]

(c) Three *effects* can be exploited when designing a fusion algorithm.
   (i) Describe the three effects.
   (ii) Based on the fusion algorithms you have studied, give *four* examples where one of these effects is exploited, and explain how they exploit this effect.

[9 marks]

(d) Briefly describe *three* sources of synonyms for use in *query expansion*.

[6 marks]

[Total 35 marks]