Semester Two of Academic Year (2015---2016) of BDIC

《 Information Retrieval 》

Module Code: COMP3009J

Exam Paper A

Exam Instructions: Answer Part 1 and any two other parts

Honesty Pledge:

I have read and clearly understand the Examination Rules of Beijing University of Technology and University College Dublin and am aware of the Punishment for Violating the Rules of Beijing University of Technology and University College Dublin. I hereby promise to abide by the relevant rules and regulations by not giving or receiving any help during the exam. If caught violating the rules, I would accept the punishment thereof.

Pledger: ___________ **Class No:** ___________

BJUT Student ID: ___________ **UCD Student ID**___________

。。

Notes:

The exam paper has 4 parts on 6 pages, with a full score of 100 points. You are required to use the given Examination Book only.

Instructions for Candidates

Candidates should answer Part 1 and any two other parts. Part 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.

Obtained score

Part 1:

a) Explain what is meant by *stopword removal*. Are there situations where this may harm retrieval performance?

[6 marks]

b) Describe what is meant by *information need* and how it is different from a query that is supplied to an Information Retrieval system.

[6 marks]

c) Briefly explain why *database overlap* is important in the area of fusion.

[6 marks]

d) Explain what is meant by the phrase *Adversarial Information Retrieval*. In the context of web search, show two examples of this.

[6 marks]

e) There are three key phases that are typically present in most Information Retrieval systems. Briefly describe each of these three phases.

[6 marks]

[Total 30 marks]

Obtained score

Part 2:

a) The *Boolean Model* makes use of the query operators *AND*, *OR* and *NOT*. Explain how these work and how they affect the number of documents returned by an Information Retrieval system. Also show how each of these can be implemented by using operations from Set Theory.

[6 marks]

b) TF-IDF is one method of calculating *term weights* for Information Retrieval systems.

(i) Why are term weights important in Information Retrieval systems?

(ii) When using TF-IDF, why would a term have a high weight?

[5 marks]

c) Below is a small document collection, containing three documents. Answer the questions that follow, showing your workings for each.

Stopwords: a, but, he, is, of, the

Document 1: A crowd of lions is called a pride

Document 2: The crowd of injured people called the hospital

Document 3: He suffered nothing but injured pride

(i) Calculate a vector for each document, using the TF-IDF weighting system. You should use the stopword list provided, but you are not required to perform stemming.

(ii) Calculate a vector for the query “injured people hurt”.

(iii) Calculate the cosine similarity between the query vector and each of the document vectors, and show the final ranked list of documents for this query.

[15 marks]

d) The Probabilistic Model of Information Retrieval makes use of two probabilities relating to query terms. These are the *probability that a relevant document will contain the term* and the *probability that a non-relevant document will contain the term*. However, these probabilities cannot be calculated directly and must be estimated.

(i) Briefly describe how initial values for these probabilities may be generated.

[3 marks]

(ii) Explain how these initial estimates can be improved with user feedback.

[6 marks]

[Total 35 marks]

Obtained score

Part 3:

a) Using an example with at least 4 documents and at least 6 links, show a worked example of how *PageRank* scores are calculated. Use a damping factor of *d=0.8* and perform at least 3 iterations.

[12 marks]

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

Retrieved = d_0, d_7, d_{19}, d_1, d_{12}, d_{18}, d_6, $d_{16,}$ d_{10}, d_9, d_8, d_{13}

Below is the set of relevance judgments for the same query:

Relevant = {d_0, d_1, d_{12}, d_{16}, d_{17}}
Non-relevant = {d_6, d_8, d_9, d_{10}, d_{18}, d_{19}}

For the above query, calculate the *MAP* and *bpref* score.

[11 marks]

c) Information Retrieval evaluation is traditionally based on the *Cranfield paradigm*.

(i) Describe in detail how evaluation can be done using this paradigm, and how it can result in the calculation of *precision* and *recall*.

[6 marks]

(ii) Precision and recall are not usually used to evaluate modern Information Retrieval systems. What features do modern systems have that make these metrics unsuitable? How have other evaluation metrics solved these problems?

[6 marks]
[Total 35 marks]

Obtained score

Part 4:

a) There are three *effects* that may be exploited by a Fusion algorithm.

(i) Briefly describe each of these effects.

[6 marks]

(ii) Describe how the *ProbFuse* algorithm exploits these effects.

[3 marks]

b) Explain, with the aid of an example, how the *interleaving* fusion system works. Identify a change to the algorithm that may improve its performance.

[9 marks]

c) The table below shows results from three 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 score that was used for ranking. Complete the following tasks, showing your workings for each.

(i) Normalise the scores for each ranked list.

[6 marks]

(ii) Fuse the results using the CombSUM algorithm.

[4 marks]

(iii) Fuse the results using the CombMNZ algorithm.

[4 marks]

(iv) Explain why it is necessary to perform normalization for score-based fusion.

[3 marks]

Engine A

DocID	Score
D9	0.4
D15	0.39
D3	0.38
D6	0.12
D10	0.10
D11	0.08
D0	0.07
D18	0.05

Engine B

DocID	Score
D1	12
D7	11
D15	9
D17	6
D14	5
D18	4
D8	3
D13	1

Engine C

DocID	Score
D18	2451
D2	2245
D15	2108
D7	1744
D1	1430
D0	1427
D5	1264
D19	1002

[Total 35 marks]