Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2022/2023

MODULE: CA4009 – Search Technologies

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
DS	BSc in Data Science
EC	BSc in Enterprise Computing
ECSAO	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 3,4,O

EXAMINER(S):

Gareth Jones	(Internal)	(Ext:5559)
Prof. Arend Rensink	(External)	External
Dr. Donnacha Daly	(External)	External

TIME ALLOWED: 3 hours

INSTRUCTIONS: **Candidates should answer Question 1 in Section A and any 3 questions from the 5 questions in Section B.**

All questions are worth a maximum of 25 marks.

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

There are no additional requirements for this paper.

Section A

Question 1 is COMPULSORY.

QUESTION 1 ***[TOTAL MARKS: 25]***

[25 Marks]

Question Overview This question requires you to analyse a scenario for which a new search application is required, and then to propose the design of a search application for this situation. Your analysis and design should be based on material studied in CA4009 Search Technologies and any other relevant material which you might wish to incorporate.

In answering this question it is suggested that you address as appropriate the following elements:

- analysis of the search requirements of the end users of the system
- analysis of the domain and the search expertise of the end users
- consideration of the types of queries that might be entered by the users
- available search technologies that could be used in a new search application to address this problem
- selection of a set of required components for your new search application and how these would be combined or used within the new system
- how the new system could be evaluated, including the features of a suitable test collection and choice of evaluation metrics

These points are suggestions, you are free to include any issues, topics or materials that you wish to in your answer, but description of a **complete solution** containing all relevant elements is required to receive full credit.

Scenarios Answer this question by selecting one of the following scenarios requiring a new search application.

1. *LearnTech* are a leading supplier of online learning tools to schools, colleges and universities. They have recently launched a successful product which enables educational establishments to video record classes and lectures and archive them for later reuse. These materials are used by pupils and students for private study, to catch up missed classes or to review material which they did not understand in class, but they are also used by teachers and lecturers to review their teaching sessions, looking for things that went well or badly, so that they can work towards improving their teaching. While this application is effective for

recording teaching sessions, its playback component is very basic, only allowing playback from the beginning of the file and there is no facility for searching the archive for locations where specific material is covered. In order to build on the success of the content recording application, LearnTech has decided to develop a new search and browsing component to support better engagement with the recorded teaching archive at each teaching establishment. As a member of the software design and development team at LearnTech you are tasked with specifying the functionality of the new component and developing a plan for its development.

2. An internal review of the operation and work practices within a small enterprise employing about a dozen people has revealed that the information management practices are very inefficient and are costing the company a considerable amount of money. Employees have been found to be spending many hours each week looking for information within company archives, emails, professional reference documents, etc to enable them to carry out work on current projects. The company managers realise that all this time spent looking for information is not just wasting time and money, but also means that projects risk overrunning agreed delivery deadlines. Such overruns may cause contract penalty clauses to be triggered, costing more money, and damaging the company's reputation. The company managers propose to develop an enterprise search system to index all information within the company and make it available for online search in the expectation that this will improve the efficiency with which required information can be located. You work for a search technology firm which develops enterprise search solutions to meet the needs of individual companies. Your company is contracted by the small enterprise to provide a specification and development plan for the creation of an enterprise search solution for the small enterprise.

3. Many colleges and universities provide their students with online lecture notes for their course. These notes are often based on one or more textbooks and other resources. If a student wants to learn more about a topic covered in a lecture, they typically need to find out what these books or resources are and then to spend time studying them to try to find relevant material and to relate it to the lecture notes. The effort to do this is often off-putting to students who frequently end up relying on the lecture notes and not bothering to consult the more detailed descriptions contained the source text. In order to encourage students to read more deeply in the source material, the IT department at your local university, for which you work as a software developer, has decided to create a new application which automatically links the contents of lecture notes to related online content. The online content to which it is to be linked is either the source materials from which the lectures notes are derived, or is related to it and describes the topics covered in individual slides in more detail. You are tasked with developing a system to be made available to students which will automatically link the contents of sets of lecture notes, provided to them for the modules that they are studying, to available online digital resources, and to be able to efficiently browse retrieved content to find relevant material to help them in their studies.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

QUESTION 2 ***[TOTAL MARKS: 25]***

2(a) **[4 Marks]**

What does it mean to "jump into a hyperspace", such as the World Wide Web (WWW)?

Why can the output of a web search engine be considered to be a dynamically created landmark node?

2(b)

i. **[5 Marks]**

Retrieving relevant documents at the top of a ranked retrieval list in web search based only on query-document content matching is generally not effective. Why is this? To answer this question, recall that user queries to web search engines are typically very short and that the world wide web is very large.

How do link-based methods such as *PageRank* improve the reliability of document ranking for web search engines?

ii. **[4 Marks]**

Use a simple example to outline the principles of the operation of the PageRank algorithm.

2(c) **[7 Marks]**

i. Explain the concept of "learning-to-rank" as used in Web search.

ii. Outline **two** complex signal factors typically used in learning-to-rank for Web search. No credit will be given for sImple count based factors such as no of Inlinks, page URL length, no of matching terms between query and document.

Note: These features should be in addition to the use of standard information retrieval ranking methods and PageRank. No credit will be given for describing information retrieval ranking methods or PageRank in the answer to this part of the question.

2(d) **[5 Marks]**

What is A/B testing as applied to online web applications? Describe a simple example of the use of A/B testing to evaluate a proposed new algorithm for use in an existing web search engine.

[End Question 2]

QUESTION 3 ***[TOTAL MARKS: 25]***

3(a) **[5 Marks]**

The purpose of an information retrieval (IR) system is to satisfy a user's information need. User engagement with an IR system is essentially a cognitive process in which the user seeks to satisfy their information need by entering a query and examining documents retrieved in response to the query.

When using an IR system, users often encounter difficulties in creating effective queries due to the "non-specifiability of need"' problem.

What is the non-specifiability of need problem, and why do users encounter this problem? Use examples to illustrate your answer.

3(b) **[4 Marks]**

Give **three** examples of English stop words, and explain why they are stop words. Why are stop words often removed in IR systems?

3(c) **[5 Marks]**

What are *stemming* algorithms as used in automatic indexing for IR?

Explain using examples what is meant by *under-stemming* and *over-stemming*.

3(d) **[8 Marks]**

Term weighting is an important factor in the effective ranking of documents in best-match IR systems. One of the best known and most effective methods of term weighting is the Okapi BM25 term weighting function given by the following equation:

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where

i = the current search term
j = the current document
$cw(i,j)$ = the overall BM25 *combined weight* of search term i in document j
$cfw(i)$ = the *collection frequency weight* of search term i
$tf(i,j)$ = the within document *term frequency* of term i in document j
$ndl(j)$ = the normalised length of document j
k_1 = an experimentally determined constant
b = an experimentally determined constant

With reference to the Okapi BM25 model as described by the equation above, explain the concepts of:

- *collection frequency weighting*,
- *term frequency* weighting,
- *document length normalisation*

How do the k_1 and b factors operate in the equation for the Okapi BM25 model?

3(e) **[3 Marks]**

Recording proximity of terms within documents in an IR system enables it to take account of whether terms from a query are close together or far apart within a document.

Using an example, explain why term proximity information can be a useful factor in determining the potential relevance of a document to a search query containing multiple terms.

[End Question 3]

QUESTION 4 ***[TOTAL MARKS: 25]***

4(a) **[5 Marks]**

i. What is the purpose a recommender system? Use examples to illustrate your answer.

ii. What are the two main sources of information used by recommender systems to determine their output to the user?

4(b) **[6 Marks]**

Explain the following concepts as they apply to the goals of an operational recommender system: *relevance*, *novelty*, *serendipity*, *diversity*.

4(c) **[7 Marks]**

Describe in outline the operation of a recommender system based on *collaborative filtering*.

What are the key features of *memory-based* and *model-based* algorithms as used in collaborative filtering?

4(d) **[7 Marks]**

i. Give a concise definition of a document *summary*. In your answer contrast the possibilities for *depth* versus *coverage* in the summary generation process.

ii. Effective short snippet summaries are an important part of representing items in the output of a recommender system.

Suggest algorithmic components which could be used in the selection of material from the description of an item for use in a snippet summary of the item in the output of a recommender system.

In each case explain using an example how the component might contribute to the effective selection of suitable content to create a good snippet.

How might the quality of the snippet summaries created using these components be evaluated?

[End Question 4]

QUESTION 5 ***[TOTAL MARKS: 25]***

5(a) **[4 Marks]**

What is enterprise search? Why is enterprise search of increasing importance?

5(b) **[4 Marks]**

Outline the issue of controlling access to content in enterprise search. What are the two standard approaches to managing access control in enterprise settings?

5(c) **[6 Marks]**

i. Give the standard definitions of *precision* and *recall* as used in the evaluation of information retrieval systems.

Briefly explain what each of these metrics is designed to measure.

ii. Are precision or recall generally more important in: (a) web search, (b) enterprise search. Give reasons for your answers.

5(d)

i. **[3 Marks]**

What is meant by there being a "human-in-the-loop" in image and video search systems?

ii. **[2 Marks]**

What is the "semantic gap" in search of visual media and why is it important in image and video search?

iii. **[6 Marks]**

Automatic image analysis for multimedia information retrieval is typically classifed into three levels: *image primitives*, *iconography* and *iconology*. Explain these different levels of image processing. In your answer make clear the relative complexity of using each level, and how it relates to the semantic gap and human interpretation of images.

[End Question 5]

QUESTION 6 ***[TOTAL MARKS: 25]***

6(a) [6 Marks]

Knowledge graphs encode information extracted from source texts. A knowledge graph typically describes the relationships between entities and the attributes of the entities.

Create a simple graphical example of a knowledge graph to illustrate these features. Your example should contain at least **4** entities, each with at least **2** attributes.

6(b) [5 Marks]

How do search engine companies such as *Google* use information summary "cards" created from knowledge graphs to provide information about common entities as part of a Search Engine Results Page (SERP)?

Describe in outline how information contained in a knowledge graph might be selected for use in a summary "card".

6(c)

i. **[3 Marks]**

Sketch the standard workflow of a question answering system based on document retrieval.

ii. **[8 Marks]**

Explain how a question answering system based on document retrieval can seek to use documents to answer a question entered by a user for: (a) a small collection of documents, (b) a very large collection of documents.

iii. **[3 Marks]**

Why can a question answering system based on the use of a knowledge graph potentially answer questions that a document-based system is unable to?

[End Question 6]

[END OF EXAM]