Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2021/2022

MODULE: CA4009 – Search Technologies

PROGRAMME(S):

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

YEAR OF STUDY: 3,4,O,X

EXAMINER(S): Gareth Jones (Internal) (Ext:5559)

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. *Glasnevin Technologies* is a medium size enterprise developing bespoke interactive user interfaces for intelligent information systems. The company seeks to provide high quality solutions in an efficient and agile manner. A key component of this work process is to maximise reuse of expertise, experience and solutions gained in previous projects. Glasnevin have contacted your search technology consultancy asking you to develop a new search application for them to facilitate effective online search of their archives of documentation relating to their previous projects. You have been tasked by your manager with creating a project plan for the development of the search system for Glasnevin.

2. Music streaming has become the primary source of listening to music for many people. Many users discover new music by selecting from recommendations made to them, based on their previous listening behaviour. However, this encourages people to restrict their listening to genres and artists with which they are already familiar. *Banzai Music*, a music streaming company, believes that they can expand music consumption by encouraging their users to actively search for new music to listen to. To enable this, Banzai are planning to develop a new interactive search tool to enable their users (and other who might become their users) to explore Banzai's online music library to find interesting new items, using details of keywords, genres, artists and any other related metadata which might be available. Develop a design plan for development of Banzai's new search application.

3. Lectures are increasingly recorded for later viewing online. When watching a recorded lecture, the viewer often has questions about the material being described or would like to explore the topic in greater detail. *LearnOnline* provides an online platform to colleges using which they can upload their recorded lectures, and students can view the recordings as part of their studies. LearnOnline want to extend their application to include automatic linking of lecture content to external resources such as books, web pages, research papers, other videos, etc., on the specific topic being described at that point in the lecture, to give viewers easy access to related additional information. As a product developer at LearnOnline, you are tasked by your manager with developing a plan for the creation of a prototype application for the automatic linking functionality.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

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

2(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. 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.

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

i. Why can *relevance* of documents returned in response to a search query only be determined by a human relevance assessor, ideally the human who created the query?

ii. In what different ways can a document be *relevant* to addressing a user's information need?

iii. How can multiple successive search passes with different refinements of a search query be useful in helping a user to resolve their information need?

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

i. Give the standard definitions of the evaluation metrics *precision* and *recall* as used in IR.

What do each of these metrics seek to measure about an IR system?

ii. Based on these definitions, explain how use of the *pooling* procedure in the construction of an IR test collection can potentially make calculation of both *precision* and *recall* values unreliable.

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

What is A/B testing as applied to online web applications?

Explain in outline how A/B testing could be applied to compare the effectiveness of alternative IR systems available to be deployed in an operational setting.

What are the relative advantages and disadvantages of IR evaluation using an IR test collection and A/B testing?

[End Question 2]

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

3(a) **[6 Marks]**

What is *learning-to-rank* as used in web search?

Explain in outline how the learning-to-rank method is used in the development of a web search engine.

3(b) **[7 Marks]**

What is the *PageRank* algorithm?

Why can the *PageRank* algorithm be important in the operation of an effective Web search engine?

By means of a simple numerical example explain the principles of the calculation of the *PageRank* value of a Web page. No credit will be given for copying the example from the lecture notes.

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

i. Give a definition of *spam* content in the context of web search.

What types of content can be considered to be spam in web search?

ii. What is a *link farm*?

With reference to the PageRank algorithm, explain why web search engine companies seek to detect and exclude information from link farms from their search index files.

3(d) **[6 Marks]**

What are *click models* models as used in web search engines?

Why are click models often found to be a useful component in a web search engine?

Describe in outline any disadvantages which can arise from the use of click models in web search engines?

[End Question 3]

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

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

What is the "semantic gap" in search of visual media, and why is it important in image and video search? Use examples to illustrate your answer.

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

A summary is a condensed derivative of a source text, where the content is reduced through *selection* or *generalisation* on what is important in the source.

Explain what is meant by the concepts of *selection* and *generalisation* in this definition.

4(c)

The temporal nature of videos means that browsing complete videos can be time consuming and inefficient. Browsing video summaries consisting of important content taken from the video could make it much quicker for users to locate interesting material in the video relevant to their search query.

Video summaries can be created by combining selected video segments from within the video, similar to text summarization using sentence or phrase extraction.

i. **[5 Marks]**

A fundamental step in the preprocessing of a video to enable creation of a summary in this way is to divide the video into segments from which the summary could be constructed. One way to divide a video into segments is to identify short boundary points. Describe the process of shot boundary detection when preprocessing video data.

What problems are typically encountered in shot boundary detection for professionally edited real-world video data such as movies?

ii. **[8 Marks]**

Assuming that the spoken content of a video has been accurately transcribed and time aligned with the video content, and objects and people appearing in the video have been labelled, suggest component factors which could be combined to identify the video segments which should be included in a useful video summary created in response to a user search query.

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

How might reducing the semantic gap potentially enable construction of better video summaries?

[End Question 4]

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

5(a)

i. **[5 Marks]**

Explain why *stemming* algorithms are often used in automatic indexing for information retrieval (IR) systems?

For stemming of English language text, why do we generally want to stem suffixes, but not prefixes?

ii. **[3 Marks]**

Explain using examples why conflation by string comparison methods can be important in IR systems.

iii. **[3 Marks]**

Why can use of stemming algorithms and string comparison methods sometimes degrade search effectiveness compared to an equivalent IR system which does not use these methods?

5(b) **[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 does use of these weighting factors in combination within the Okapi term weighting function enable the computation of effective terms weights for best-match information retrieval?

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

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

Query expansion for the Okapi probabilistic IR model operates using the offer weight $ow(i)$ for a term i, which is defined as follows,

$$ow(i) = r(i) \times rw(i)$$

where

$rw(i) =$ the Robertson/Sparck Jones weight
$r(i) =$ the number of **known relevant** documents term i occurs in,

and the Robertson/Sparck Jones weight is defined as follows,

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

$n(i) =$ the number of documents term i occurs in.
$N =$ the total number of documents in the collection archive.
$r(i) =$ the number of **known relevant** documents term i occurs in.
$R =$ the total number of **known relevant** documents in the collection archive.

$ow(i)$ is calculated for all terms present in known relevant documents, which are then ranked by their $ow(i)$ values, with the highest ranking terms being added to the original query as query expansion terms.

Explain why each of the factors $r(i)$ and $rw(i)$ are useful measures of a term i's potential value as a query expansion term to enhance the user's initial query to improve the effectiveness of an IR system.

Following your arguments for the use of the factors $r(i)$ and $rw(i)$ in contributing to the selection of good expansion terms, explain why combining $r(i)$ and $rw(i)$ values to obtain $ow(i)$ is likely give a better overall estimate of the utility of a potential expansion term i than either of them in isolation.

[End Question 5]

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

6(a)

Knowledge graphs encode information extracted from source texts.

A knowledge graph typically describes the relationships between entities and the attributes of individual entities.

i. **[6 Marks]**

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. No credit will be given for copying knowledge graph examples from the lecture notes.

ii. **[6 Marks]**

Knowledge graphs contain information extracted from multiple source texts. However, knowledge graphs are often incomplete, e.g. some entity attribute values may be empty. How could a document-based question answering method be used to attempt to automatically add missing information to a knowledge graph?

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

Outline a method by which a knowledge graph could be used within a question answering system to answer user questions.

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

Semantic representation of words using word embedding can be used to overcome mismatch between words appearing in queries and documents in an information retrieval system.

Explain the principles of semantic representation of words using word embedding. Use an example to illustrate your answer.

How do embedded representations of words enable the word mismatch problem to be overcome?

Use of examples appearing in the lecture notes will receive no credit.

[End Question 6]

[END OF EXAM]