DCU

DUBLIN CITY UNIVERSITY

SEMESTER ONE EXAMINATIONS 2013

MODULE: (Title & Code)	CA437 / CA437A / CA437D / CA437F Multimedia Information Retrieval
COURSE:	B.Sc. in Computer Applications (SE Stream) B.Sc. in Computer Applications (IS Stream) B.Sc. in Computer Applications (Evening) B.Eng. in Digital Media Engineering
YEAR:	4
EXAMINERS: (Including Telephone Nos.)	Dr. Gareth Jones Ext. 5559 Dr. Micheal Manzke Dr. J. Power Prof. Finbarr O'Sullivan Prof. W. Buchanan
TIME ALLOWED:	2 hours
INSTRUCTIONS:	This paper contains 6 questions. Please answer any 4 questions. All questions carry equal marks.

Please do not turn over this page until 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

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

1(a)

The World Wide Wide (WWW) is an example of an inter-linked hypermedia or hyperspace which has no beginning or ending points.

What does it mean to "jump into a hyperspace", such as the WWW, at a selected node?

How do web search engines such as Google support users of the WWW doing this?

What is a landmark node in hypermedia?

[5 Marks]

1(b)

What is the PageRank algorithm as used in WWW search? Use a simple example to outline the principles of the PageRank algorithm.

[9 Marks]

1(c)

Why are algorithms such as PageRank important in search engines for the WWW, compared to other document collections such as a news report archive?

[3 Marks]

1(d)

Adversarial search is concerned with attempts to promote the rank position of documents in the output of a web search engines above their true rank. Adversarial methods include the use of link farms and stuffing pages with words to increase their term frequency. Explain the principle of a link farm and how it attempts to distort the behaviour of link-based methods in search engines.

[5 Marks]

Outline methods can that be used by search engines to prevent link farms and page stuffing from distorting ranks of documents in a ranked output list.

[3 Marks]

[End of Question 1]

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

2(a)

What is the purpose of an information retrieval system?

[3 Marks]

2(b)

Users can pose search requests with different types of focus. Some requests are better solved using a conventional information retrieval (IR) system, but others might be better addressed using a question answering (QA) system.

1. What is a QA system?
2. Use examples to explain why a QA system can provide a better solution for some search requests than a conventional IR system. In your answer be clear to explain why QA systems are not a general replacement for IR systems.
3. Why are some questions easy to answer using a QA system and others much harder?

[8 Marks]

2(c)

Using examples, explain how analysis of the HTML markup of a web page can be used to modify search term weights such as those in the the vector-space model or the BM25 model, in a web information retrieval engine?

[5 Marks]

2(d)

What is the purpose of the Google Panda algorithm? Outline the procedure used by Google to implement the Panda algorithm.

How can the output of the Panda algorithm be combined with other document scoring components in a web search engine based on document content and link structure for effective document ranking in the output of the search engine.

[9 Marks]

[End of Question 2]

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

3(a)

What is enterprise search?

[4 Marks]

3(b)

What is "database offloading" in enterprise search? Why can database offloading be particularly helpful in enterprise search system when users are not familiar with the structure and contents of the database which has been being offloaded?

[5 Marks]

3(c)

What is *faceted search*? Use examples from enterprise search to explain your answer.

[4 Marks]

3(d)

Give the standard definitions of the evaluation measures *precision* and *recall* as applied in information retrieval. In your answer explain why it is difficult to obtain both high precision and high recall in an information retrieval system.

[6 Marks]

3(e)

1. Why is recall generally more important than precision for an enterprise search application? Explain your answer.

   [3 Marks]

2. By contrast, why are web search engines such as Google typically designed to favour high precision rather than high recall?

   [3 Marks]

[End of Question 3]

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

4(a)

In best match ranked information retrieval,

$$cfw(i) = \log \frac{N}{n(i)}$$

is the standard equation used to calculate $cfw(i)$ where

$cfw(i)$	=	the collection frequency weight of term i in a document collection, also referred to as $idf(i)$ = inverse document frequency of term i in the same collection
i	=	the current search term
$n(i)$	=	total number of documents containing term i
N	=	total number of documents in the collection

1. Explain the underlying principle of collection frequency weighting.
2. With regard to $cfw(i)$, what does it mean to say that a search term in a best-match information retrieval system has high selectivity?
3. Would a term i have high or low selectivity if N has a similar value to $n(i)$? Why would this be the case?

[6 Marks]

4(b)

Term weighting is an important technique in best-match information retrieval. In addition to the $cfw(i)$ in part (b), the other two components generally used for effective term weighting in best-match information retrieval are term frequency $tf(i,j)$ and document length normalisation. For each of these components explain its underlying principles and why it is expected to be beneficial in improving the ranking of relevant documents in best-match information retrieval.

[6 Marks]

4(c)

Recording term proximity within documents in an information retrieval system enables it to take into account whether a pair of terms are close together or far apart within a document. Why can term proximity be a useful factor in determining the relevance of a document to a search query containing such a pair of terms? Suggest one means by which term proximity could be incorporated into the calculation of the query-document matching score in a best match information retrieval system.

[5 Marks]

4(d)

Relevance feedback methods in information retrieval are intended to improve information retrieval effectiveness.

1. What is relevance feedback?
2. Explain the fundamental principle of the Rocchio relevance feedback method as used in the vector-space approach to ranked best-match information retrieval. Your answer should make reference to the concepts of the vector-space model and how the Rocchio relevance feedback is used to modify search using the vector-space model to improve retrieval effectiveness.
3. Why can relevance feedback sometimes fail to improve the rank at which relevant documents are retrieved?

[8 Marks]

[End of Question 4]

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

5(a)

What are the differences between HTML and XML document markup. Use examples to illustrate your answer.

[5 Marks]

5(b)

XML markup can be used to define standards for textual metadata labels of multimedia content. These textual metadata labels can be used for retrieval of relevant multimedia content by using text information retrieval methods. Metadata labels can be generated either manually or automatically.

Design an XML markup standard that could be used for labeling of images or videos for use in a multimedia information retrieval system based on text information retrieval.

Your design should include examples of metadata labels to be assigned by manual examination of the content and using automatic content analysis techniques.

[8 Marks]

5(c)

How can context information such as location and time used to assist users in searching multimedia data collections?

[6 Marks]

5(d)

Explain using an example why graphical visualisation can improve the efficiency of content access in spoken document retrieval.

[6 Marks]

[End of Question 5]

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

6(a)

Give a short definition of a document summary. In your answer make clear the possibilities for coverage of the contents of the document being summarized within a summary.

[4 Marks]

6(b)

Shot boundary detection and shot keyframe identification are typically applied to videos to prepare them for interactive search of a video archive or for browsing of an individual video.

Describe a simple process of shot boundary detection. What are the problems typically encountered in shot boundary detection? Suggest one solution to these problems.

Suggest some simple methods to select a keyframe to represent a shot. What are the potential problems with these methods?

[7 Marks]

6(c)

Explain why image and video retrieval often needs to be more interactive, typically involving multiple retrieval passes with user feedback, to locate relevant items than text information retrieval.

[5 Marks]

6(d)

Describe in overview the design of a query-biased video summarization system which makes use of the output of a shot boundary detector and extracted keyframes, and also any other media stream (text, audio or speech) or metadata data sources that you wish. Your design should take into account that video is a time-based medium making it difficult to browse content rapidly.

Suggest how you would evaluate the effectiveness of your summarization.

[9 Marks]

[End of Question 6]

[END OF EXAM]