OLLSCOIL NA hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND
COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

Summer Examinations 2003
B.Sc Honours

Computer Science
CS4040: Information Retrieval and Organisation

Prof. M. Calder
Prof. C.J. Sreenan
Mr. H. Sorensen

Answer *all* questions

Paper Total: 160 Marks

All questions carry equal marks

Time: 3 Hours

1. Search Engines

You intend to develop a fully-functional (text-only) search engine for deployment on the internet. It is to support a rich set of query types, with well-defined semantics. For a given query, it will rank results based on conditions being satisfied, the number of query terms present, their specificity, where they are within the page, and on the page popularity. The robot is to be well-behaved.

(a) State what query types you intend to support and the default semantics for a multi-term query. *(4 marks)*

(b) Outline the structure of an *inverted index* suitable for support of the query types of (a). *(8 marks)*

(c) The search engine will retrieve all Web pages that exactly match a user's query words (or expression). However, it is also to reward – to a lesser extent – pages containing variants of the query terms. How exactly might this be achieved? *(5 marks)*

(d) Specify an appropriate *ranking algorithm* that might be employed by this search engine. *(8 marks)*

(e) In what way might your robot be well-behaved? *(3 marks)*

(f) Is this search engine susceptible to spoofing by Web page writers? Explain. If so, explain how such spoofing might be challenged and outline any detrimental effects your solution might have *(4 marks)*

2. IR Models

(a) In the *Boolean Model*, queries must be expressed as Boolean expressions and must be cast into disjunctive normal form (DNF) prior to evaluation.

(i) Cast the following query into DNF:

Term1 AND Term2 AND (Term3 OR Term4) *(2 marks)*

(ii) Specify how the DNF query is evaluated against the document-term representation. *(2 marks)*

(iii) Why is the Boolean Model considered unsuitable for modern-day deployment in information retrieval? *(2 marks)*

(b) The *Vector Space Model* (VSM) is a commonly used model for document and query representation. Describe this model, paying particular attention to the following features:

(i) Give both the geometric and the algebraic representation of a document vector. *(3 marks)*

(ii) Specify a *term weighting function* that takes word specificity, but not document length, into account. *(3 marks)*

(iii) State why and how the document and query vectors are generally normalised to be unit length from the origin. *(3 marks)*

(iv) Outline the basic storage structure for a collection of document representations. *(3 marks)*

(v) Specify the computational model whereby the storage structure outlined in (iv) is used to compute the ranking of stored documents relative to a query. *(3 marks)*

(vi) State what steps might be taken to increase the efficiency of the ranking algorithm. *(3 marks)*

(vii) Would you consider the VSM to be universally suited to textual information retrieval? Explain. *(3 marks)*

(c) The *Latent Semantic Indexing (LSI) Model* is a variation of the VSM that attempts to address some perceived shortcomings of the latter. State the essential difference between these models and the benefits and drawbacks of using the LSI approach. *(5 marks)*

3. Query Enhancement

(a) One approach to query enhancement involves the use of *relevance feedback*, whereby the results of one query phase can be used to adjust the input into the next phase, thereby allowing the phased (re)focusing of a query. With respect to either the *Vector Space Model* **or** the *Probabilistic Model* for information retrieval, address the following aspects of the process:

(i) Explain the basic principle involved in relevance feedback. *(4 marks)*

(ii) Derive and explain the means by which a query or similarity expression is altered from one query phase to the next. *(7 marks)*

(iii) Specify how one might objectively assess the improvement brought about by the use of relevance feedback. *(3 marks)*

(b) A second form of query enhancement involves *query expansion* – usually requiring the use of a *statistical thesaurus* so as to add synonymous terms to a query. This thesaurus itself can be constructed through either *local* or *global clustering*.

(i) State the main differences between *local* and *global* clustering. *(3 marks)*

(ii) Describe the synonym-finding model of any one local clustering method. *(7 marks)*

(iii) In global clustering using the *complete link algorithm*, the aim is to keep document clusters small and tight. Why and how? *(3 marks)*

(iv) Automatic document clustering, especially when combined with *information visualisation*, might be useful in its own right. Explain how information access might be augmented through this process. *(5 marks)*

4. **Retrieval Evaluation; Document & Query Processing**

(a) *Precision* and *recall* are the most common measures of retrieval accuracy. They are usually combined to produce a *Precision:Recall (P:R) Graph.*

(i) Define the terms precision and recall. (*2 marks*)

(ii) Specify the steps involved in carrying out experimental evaluation of P & R. *(3 marks)*

(iii) Assume that an IR system ranked documents as listed overleaf in the left hand column. Assume that the marked documents are those deemed by the user to be relevant. Construct the P:R graph for this case and comment of the apparent accuracy of the IR system. Repeat the process for the right-hand column (depicting another IR system using the same query & documents). Assume that you have prior knowledge that the collection contains 15 relevant documents. Briefly compare and contrast the IR systems involved. *(13 marks)*

(iv) IR systems can have very different types of uses – from casual to professional and from short-query to full-text, for example. For which type(s) of system is the P:R measure most appropriate? Explain. For a general-purpose search engine, what do you think would be a suitable evaluation measure? *(5 marks)*

(b) There are several reasons for – and several means by which - terms can be added to, or omitted from, document and/or query representations.

(i) State what these techniques are. *(4 marks)*

(ii) What effect would you expect them to have on precision & recall measures – and why? *(3 marks)*

(c) If you permitted regular expressions, using wildcards, in queries, what effect would you expect them to have on precision & recall measures? *(2 marks)*

Doc 9		Doc 125	*
Doc 1	*	Doc 68	*
Doc 31		Doc 90	*
Doc 125	*	Doc 11	*
Doc 12		Doc 86	*
Doc 68	*	Doc 82	
Doc 90	*	Doc 9	
Doc 88	*	Doc 1	*
Doc 77	*	Doc 10	*
Doc 56		Doc 16	
Doc 54	*	Doc 17	
Doc 33		Doc 229	
Doc 11	*	Doc 301	
Doc 8	*	Doc 2	
Doc 66	*	Doc 33	
Doc 225	*	Doc 7	
Doc 16		Doc 29	

5. Multimedia IR

(a) Outline why the GEneric Multimedia object INdexIng (GEMINI) algorithm for indexing of multimedia data is used, and how it fits in to the overall approach to Multimedia Information Retrieval. Why should the distance in feature space lower-bound the actual distance between objects? *(4 marks)*

(b) Explain how this indexing technique might be applied to *colour image* information retrieval. Pay particular attention to the following issues:

(i) Specify an appropriate distance function between colour images. *(3 marks)*

(ii) State what feature extraction might take place. *(3 marks)*

(iii) Specify how the distance in feature space might be computed. *(2 marks)*

(iv) What experiment(s) might be carried out to estimate the computational saving achieved by use of the GEMINI approach. Indicate the expected results. *(3 marks)*

(v) Explain whether your approach alleviates the *dimensionality curse* or the *cross-talk* problems inherent in multimedia retrieval. *(3 marks)*

(c) *Information visualisation* can prove useful in the case of either video or audio content analysis. With respect to any one of these media, specify how this visualisation is achieved and used:

(i) State which media feature(s) are being visualised and what display structure is employed. *(4 marks)*

(ii) Outline the computation involved in producing the data from which the display structure is derived. *(6 marks)*

(iii) State what interpretations and conclusions can be inferred from the visualisation. *(4 marks)*