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

## Autumn Examinations 2009
**B.Sc Honours**

**Computer Science**
*CS4040: Information Retrieval and Organisation*

Prof. A.F. Smeaton
Prof. J.A. Bowen
Mr. H. Sorensen

Answer *all* questions

Paper Total: 160 Marks

All questions carry equal marks

**Time: 3 Hours**

1. **IR Models**

(a) The Vector Space Model has long been a benchmark model for information retrieval. Why has it not been employed within Web search engines? Similarly, why has the Probabilistic Model not been so employed? *(5 marks)*

(b) With respect to the *Vector Space Model* (VSM):

(i) Define the *tf/idf* [term frequency / inverse document frequency] term weighting mechanism and specify why it is appropriate. *(5 marks)*

(ii) Define the *cosine similarity* measure and specify why it is more appropriate than inter-document *distance*. *(5 marks)*

(iii) Specify a data structure that might be employed for storing the document representations. Comment on any organisational strategies that might make it more efficient for the retrieval task. *(5 marks)*

(iv) Describe the effect on the stored document representations of adding new documents or changing existing documents. Which values must be recomputed? How can the strategy be slightly modified so that it is more resilient to the addition of new documents? *(5 marks)*

(v) Relevance feedback might be incorporated into the VSM in order to allow a user incrementally improve the focus of a query. Give a geometric interpretation of this process and an algebraic process by which it might be achieved. State any limitations of your technique. *(5 marks)*

(c) The *Probabilistic Model* of information retrieval presents an alternative approach to the VSM, one in which relevance feedback (implicit or explicit) is a requirement.

(i) Give the fundamental formula for the Probabilistic Model, which weights individual terms based on their probability of being relevany/non-relaevant. *(3 marks)*

(ii) Specify how vocabulary terms can be given initial weights [before a query has been evaluated] and how these weights can be subsequently altered from one iteration to the next. *(7 marks)*

## 2. Document/Query Processing Techniques

(a) *Clustering* – either at document or term level – can prove beneficial during information retrieval.

(i) Specify an algorithm by which local term *or* document clustering might be achieved. *(13 marks)*

(ii) Outline, with the aid of an example, how the retrieval process might benefit from the clustering of (i) above. *(6 marks)*

(iii) Are there any drawbacks to this particular approach? Give an alternative. *(6 marks)*

(b) Instead of using the approach of (a) above, why not simply use a *linguistic thesaurus* (synonym finder), as commonly found in word processing applications? *(3 marks)*

(c) A number of techniques exist for processing document/query content at the individual term, or word, level. With respect to the following techniques, specify (i) what technique might be applied, (ii) at what point it might be applied, and (iii) what benefits it might have.

(i) Stopword elimination *(4 marks)*

(ii) Stemming/conflation *(4 marks)*

(iii) Acronym expansion *(4 marks)*

## 3. Retrieval Evaluation; Document & Query Processing

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

(i) Why is the data usually provided in the form of a continuum – graph or table – rather than as a simple pair of values, P & R? *(5 marks)*

(ii) Assume that, for a given query, an IR system ranked documents as listed in Fig. 1 in the left hand column. Assume that you have prior knowledge that the collection contains 20 relevant documents, and that the marked documents are those retrieved that belong to this set. Construct the P:R graph for this case and comment on the apparent accuracy of the IR system. Repeat the process for the right-hand column (depicting another IR system using the same query & documents). Briefly compare and contrast the IR systems involved. *(12 marks)*

*(iii)* IR system developers sometimes use P:R measures in an attempt to tune up the ranking algorithm. If you were constructing a web search engine, how would you approach this problem? *(5 marks)*

(iv) If an IR system embodied relevance feedback, how would you measure its effectiveness? *(5 marks)*

(v) What are the characteristics of a good test collection for evaluating an IR system? *(5 marks)*

(b) When evaluating accuracy for an IR system, some content types might skew the results. Comment on the expected effect of the following content types:

(i) Numeric data *(4 marks)*

(ii) Acronyms *(4 marks)*

| *IR System 1* | | *IR System 2* | |
|---|---|---|---|
| Doc 19 | * | 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 19 | * |
| Doc 18 | | Doc 15 | * |
| Doc 77 | * | Doc 10 | * |
| Doc 56 | | Doc 33 | * |
| Doc 54 | * | Doc 17 | |
| Doc 33 | * | Doc 229 | |
| Doc 11 | * | Doc 301 | |
| Doc 8 | * | | |
| Doc 66 | * | | |
| Doc 225 | * | | |
| Doc 16 | | | |
| Doc 231 | | | |
| Doc 233 | * | | |
| Doc 15 | | | |
| Doc 37 | * | | |

**Fig. 1 – IR System Rankings**

4. **Multimedia IR**

(a) Within multimedia IR, two strategies are commonly employed: (i) *GEneric Multimedia object INdexIng (GEMINI)*; and (ii) *Spatial Access Methods*.

(i) What is implied by GEMINI? Illustrate this by means of *time series* information retrieval.

*(5 marks)*

(ii) What is implied by Spatial Access Methods? *(3 marks)*

(b) Describe a possible approach to *colour image (photographic)* information retrieval. Pay particular attention to the following issues:

(i) What type of queries do you intend to support? *(3 marks)*

(ii) Specify how the similarity of two images might be measured. *(5 marks)*

(iii) State what GEMINI feature extraction might take place and how it is achieved. *(8 marks)*

(iv) Specify how the distance in feature space might be computed. *(4 marks)*

(v) Explain whether your approach alleviates the *dimensionality curse* or the *cross-talk* problems inherent in multimedia retrieval. *(4 marks)*

(c) *Information visualisation* can prove useful in the case of *video* content analysis. 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) State what interpretations and conclusions can be inferred from the visualisation. *(4 marks)*