

AMSc IRIC exam 00/01 – questions on evaluation, and suggested answers

Question 1

a) Give definitions for (standard) **recall** and **aspectual recall**. Describe one limitation of (standard) recall and explain how aspectual recall addresses this limitation.

[4 marks]

Recall is the proportion of relevant documents retrieved out of the total number of relevant documents in the collection. Aspectual recall identifies different aspects of a query, and calculates the proportion of aspects for which at least one document has been retrieved. The main limitation of recall is the necessity to know all the relevant documents in advance. Aspectual recall addresses this limitation by only requiring that one relevant document be found for an aspect to be considered retrieved.

b) Two students, student A and student B, have a report to write on the same topic, and are searching for suitable references. They are using a small, bibliographic collection of 500 references. They submit one common query to the IR system, and both look through the resulting ranked list. However, they have different strategies:

- Student A is prepared to spend an hour judging the references
- Student B has an idea of the size of report she wishes to produce, and would like to find 15 relevant references

Assume that both students examine the documents in ranked order, and that a document takes, on average, 2 minutes to read. For the given query, 30 documents were previously found to be relevant, and the recall/precision table is as follows:

Recall	Precision
0.1	0.8
0.2	0.6
0.3	0.5
0.4	0.4
0.5	0.3
0.6	0.2
0.7	0.15
0.8	0.1
0.9	0.1
1.0	0.06

Using definitions of recall and precision, and the recall/precision table above, calculate the following (remembering to justify your answers):

- (i) How long will it take student B to find the 15 relevant references she is seeking?
- (ii) How many more relevant references will student B find, compared to student A?

[10 marks]

- (i) Student B is seeking 15 relevant documents. This corresponds to a recall level of 0.5, so the corresponding precision figure, from the given table, is 0.3. She will therefore have to examine 50 documents in order to find 15 relevant ones. Since examination of each document takes 2 minutes, the whole process will take 100 minutes, i.e. one hour and 40 minutes.
- (ii) We must first calculate how many relevant references student A will find in an hour. Since examination of each document takes 2 minutes, she will be able to examine 30 documents in this time. The quick way to the answer is to note that recall and precision therefore both have denominator 30 in this case. Since they share the same numerator (representing the number of relevant documents retrieved), they must have the same value in the table. The slow way to the answer is to work out recall and precision values at each stage of the table until the correct combination is found. Either way, the answer is that recall = 0.4 and precision = 0.4. Reminding ourselves that she will examine 30 documents, this means she will find 12 relevant documents. So student B will find 3 more relevant documents than student A.

c) Two instruments commonly employed in user-centred evaluation are **questionnaires** and **think-aloud protocols**. Explain how each of these instruments may be used and what kind of data may be gathered using them.

[6 marks]

Questionnaires come in a variety of forms (structured, semi-structured, unstructured), and may be administered before, during or after the search process. They can gather many different kinds of data, both quantitative and qualitative. They can be administered in person, or distributed and returned later. Think-aloud protocols involve the experimenter watching the searcher interact with the IR system and encouraging the searcher to voice his/her thoughts during this process. This is a particularly good method for picking up unexpected problems or errors that are likely to be forgotten quickly.

Question 2 – part ii only for evaluation!

d) SuperSearch, an internet company, has developed two interfaces to its new search engine. Both interfaces offer interactive query expansion, i.e. the ability for the user to choose which extra top-ranking terms should be added to the query. Both interfaces offer a set of 20 possible query expansion terms for the user to examine. However, the two interfaces present this set of terms differently:

- In interface A, all 20 possible terms are selected by default, so the user must deselect any terms he does not wish to be added to the query.
- In interface B, no terms are selected by default, so the user must select any terms he wishes to be added to the query.

- (i) Identify and briefly discuss two issues that you think an interface designer should take into consideration when designing an interactive query expansion interface.
- (ii) Which of the two interfaces, described above, do you think will result in greater **user satisfaction**? Define your usage of this term, and describe a small-scale experiment to test your hypothesis. Justify your answer and state clearly any assumptions you make.

[10 marks]

I'm less bothered here by what the exact answers are than by the reasoning and justification provided.

For part (ii), user satisfaction may be defined as the degree to which the user is satisfied with the overall searching experience, including both the process of searching itself and the end results obtained. For the experimental design, I would expect some basic discussion of *most* of the following issues:

- Type of experiment (laboratory, field,...)
- Type and number of users (real vs experimental subjects,...)
- Type and number of tasks/information needs/queries
- Measures to be used in the experiment (recall/precision, user satisfaction,...)
- Instruments to be used in the experiment (questionnaires, interviews, think-aloud protocols,...)
- Experimental set-up (possible pilot test, within-subject vs between-subject design,...)
- Results expected and analysis to be performed (quantitative, qualitative, statistical significance testing,...)