Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Topics in Advanced Information Retrieval

Date: Tuesday 29th May 2012
Time: 09:45 - 11:45

Please answer any THREE questions from the FIVE questions provided
For full marks your answers should be concise as well as accurate.
Marks will be awarded for reasoning and method as well as being correct.

This is a CLOSED book examination
The use of electronic calculators is NOT permitted
1.

a) A conventional information retrieval system requires a user to formulate a query to initiate a search. Cole notes that “what underlies the user’s formulation of the query is not considered” by system-oriented information retrieval system designers, who “treat the query as a given” (Cole, 2011).

i) State what notions you would identify underlying the formulation of a query that a designer should attempt to take into account, noting any problems that may arise with such notions. In your answer, also give supporting examples and justifications.

(2 marks)

ii) To what extent is it helpful to a designer to subclassify any of the notions you have identified? Give examples to support your argument.

(3 marks)

b) Consider the following document collection, where each document has a unique identifier (docn):

- doc1: US Tornadoes’ Death Toll Mounts.
- doc2: Will A-bombs generally stop tornadoes?
- doc4: “British Tornadoes bombed us on Mount Usa” – General DeAth Will.

i) Briefly explain and justify the decisions you would make regarding what you take to be an index term in these documents.

(1 mark)

ii) Based on your index term decisions, draw up the basic binary term-document incidence matrix for this collection.

(2 marks)

iii) Based on the binary term-document incidence matrix you have established, draw up the basic inverted index representation for this collection.

(2 marks)

iv) For this collection, using the inverted index you have established, what would be the result of the following query? Demonstrate how you arrive at your answer.

```
tornado AND bomb
```

(2 marks)

(If the result would be empty with respect to your inverted index, then, for the same marks, briefly explain what changes would be needed to obtain some result, assuming no change by the user to the query.)
c) Consider the two postings lists PL1 and PL2:

PL1: [3, 5, 9, 15, 24, 39, 60, 68, 75, 81, 84, 89, 92, 96, 97, 100, 115]
PL2: [3, 5, 89, 95, 97, 99, 100, 101]

PL1 has skip pointers 3→24, 24→75, 75→92 and 92→115.
PL2 has no skip pointers, as it represents the result of a previous operation.

i) List the comparisons required to merge (intersect) PL1 and PL2. For each comparison, use the form {x, y}, where x is a document ID from PL1 and y is a document ID from PL2.

(2 marks)

ii) How do we determine where to place the skip pointers? Ensure that you briefly justify your answer.

(1 mark)

d) You formulate a Boolean query for an information retrieval system based on the Boolean model, using 3 terms linked by AND. You receive some unranked results. You then formulate another query, by appending to the original query a further AND plus a fourth term. You receive no results.

i) Explain why you should not be surprised to receive no results.

(1 mark)

ii) Set out pros and cons of the Boolean model, discussing scenarios where this model may be preferred and why.

(4 marks)
2.

a)

i) Both stemming and the removal of stop words have the effect of reducing the size of a collection’s indexed vocabulary. Which, if any, of these two steps would lead to the greater reduction? Justify your answer.

(1 mark)

ii) Say if the following are true or false. Justify your decisions

a) In a Boolean retrieval system, stemming never lowers precision.

b) In a Boolean retrieval system, stemming never lowers recall.

(1 mark)

iii) In 1995, Croft expressed the view that stemming “produces reliable improvements in effectiveness” and “is well-regarded by users”.

Is this a view you would hold today? Justify your answer.

(2 marks)

b) Briefly explain how decisions on the following can affect the performance of a retrieval system, giving examples:

- Tokenisation
- Normalisation (e.g., mapping to lower case, acronym punctuation normalisation)
- Asymmetric expansion
- Transliteration
- Synonymy

(4 marks)

c)

i) Why are we interested in measuring the inverse document frequency (IDF) for a word, rather than its document frequency?

(2 marks)

ii) Account for the enduring popularity of IDF, giving appropriate justifications.

(2 marks)

d) Can the tf-idf weight of a term in a document exceed 1.00? Justify your answer.

(1 mark)
[Question 2 continues from the previous page]

e)  
   i)  Explain two ways in which cosine similarity scores are used with the Vector Space Model.

   (2 marks)

   ii)  I choose from an indexed collection, at random, 3 documents represented by vectors of weighted terms. Document 1 has 25 terms in its vector, document 2 has 304 terms and document 3 has 556 terms. A colleague suggests I should therefore apply the following algorithm to the entire collection: find the document with the shortest vector, and truncate all other vectors to be of that length. Explain why this may or may not be a good approach. If you decide it is not a good approach, explain what would be a better one.

   (1 mark)

f) “There is a continuing debate as to whether “understanding of content” is essential for effective information search or whether statistical methods will do.”

   (Wilks & den Besten, 2007)

   Set out your contribution to this debate. Justify your views and conclusions, giving appropriate examples to back up your arguments.

   (4 marks)
3.

a) Consider the following partially completed diagram:

The outer ellipse represents the document collection. Circle A represents the relevant documents for some information need. Circle B represents the result of a query attempting to answer this need.

Reproduce this diagram and clearly label appropriate parts to indicate:
• True positives
• False positives
• True negatives
• False negatives

(2 marks)

b) A lorry travels 100 miles at 25 mph, then another 100 miles at 20 mph. I calculate that the average speed of the lorry is \( \frac{25 + 20}{2} = 22.5 \) mph. My friend, however, does the following calculation: \( \frac{100}{25} = 4; \frac{100}{20} = 5; \frac{200}{(5+4)} = \frac{200}{9} = 22.22 \).

i) Explain what my friend has calculated.

(1 mark)

ii) Explain why such a calculation is of relevance in the evaluation of information retrieval systems.

(2 marks)
c) An information retrieval system under evaluation yields a list of 10 documents for a particular test query. This list of documents is ranked, with position 1 being the top-ranked item. The evaluator knows there are 5 relevant documents for the test query in the gold standard collection. He notes that the relevant documents returned in the ranked list occur at positions 2, 3, 4, and 8.

i) Give both the recall and the precision values for each of the 4 returned relevant document positions.

(2 marks)

ii) Using your results from i), calculate and show the 11-point interpolated precision for this query. This involves determining the interpolated precision value at each of the 11 standard recall levels (0.0, 0.1, . . . , 0.9, 1.0). Show your result in a table, not a graph, with the left column being a standard recall level and the right column being the interpolated precision you have calculated for that level.

(4 marks)

Hint 1: The formula for precision is:

\[
\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{number of retrieved items}}
\]

and for recall it is:

\[
\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items}}
\]

Hint 2: Interpolated precision at some recall level \(r\) is defined as the highest precision found for any recall level \(r' \geq r\).

d) I wish to evaluate my Boolean model based retrieval system in a search environment in which the collection is an aggregation of numerous repositories whose contents change on a daily basis. One colleague recommends that I use precision-at-\(k\), another that I use 11-point interpolated precision. Which of these methods should I choose to evaluate my search engine? Justify your answer, citing advantages and disadvantages where appropriate.

(2 mark)

e) Why, in \(n\)-gram based phrase indexing, is \(n=2\) the most common setting?

(2 marks)

f) Elaborate on challenges and problems that may arise in attempting, at both indexing and query stages, to apply text mining techniques of multiword term extraction and named entity recognition. In particular, in relation to query aspects, discuss what re-design implications there may be for a conventional search interface that currently involves a simple text box.

(5 marks)
4. a) Resource Description Framework (RDF) is based on triples representing binary predicates. Consider the following RDF/XML fragment:

```xml
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
 xmlns:uni="http://www.mydomain.org.uni-ns">
 <rdf:Description rdf:about="949318">
   <uni:name>David Billington</uni:name>
   <uni:title>Associate Professor</uni:title>
   <uni:age rdf:datatype="&xsd;integer">27</uni:age>
 </rdf:Description>
</rdf:RDF>
```

i) Is this a valid RDF/XML fragment? If yes, state which triples are involved. If no, state what would be required to amend it to be a valid triple-based fragment, and separately state what the triples are. In either case, when stating triples, do so using a directed graph representation, labelling the graph components appropriately to show their status with respect to a triple.

(2 marks)

ii) Explain how RDF namespaces differ from XML namespaces, stating what this difference implies for the organisation of knowledge, commenting in particular on the nature of `rdf:about`.

(2 marks)

b) “XML is to XML Schema (XMLS) as RDF is to Resource Description Framework Schema (RDFS)”. Why is this a false comparison?

(1 mark)

c) Consider the following informally expressed statements:
   “Discrete Maths is taught by Concrete Maths”
   “Room 1002 is taught by John Smith”

Explain informally how RDFS enables us to avoid nonsensical statements such as these, commenting on any limitations of RDFS in this respect that you may identify.

(2 marks)

d) RDF, RDFS, SPARQL and OWL are key enabling languages for the Semantic Web. Explain the contribution of each to rendering the web 'semantic'.

(4 marks)
c) I have developed an OWL ontology, in collaboration with a pizza store owner, that I have used to drive a structured data entry form for store personnel to record details of pizzas ordered. The owner would like his web site to offer a facility to search for types of pizza, bases and toppings. Discuss how I could re-use my ontology to satisfy this requirement and what advantages/disadvantages there are to adopting an ontology-based approach.

(4 marks)

f) Consider the following quotation:

“Because it’s a complex format and requires users to sacrifice expressivity and pay enormous costs in translation and maintenance, the Semantic Web will never achieve widespread public adoption.” (McCool, 2005)

Several years have passed since this claim was made. Discuss to what extent this view can be said to have any validity today, and give reasons for or against recommending that a company should today engage in Semantic Web-based activities. Justify your positions, giving appropriate examples to back up your arguments.

(5 marks)
5. Consider the following quotations:

“Google needs to move from words to meaning. [...] Google's long-term goal is to be able
to give you one answer, which is exactly the right answer.” (Schmidt, Google, 2009)

“Web and enterprise users will become accustomed to interacting with and exploring
information, and there will be no going back to plain-old keyword search and low-value
hit lists of search results.” (Grimes, Alta Plana)

“Despite its promise, semantic search faces numerous obstacles.” (Lawton, Computing
Now, 2010)

“Most things don’t exist in isolation. They have complex relationships to other things, and
by representing that information using verbs – for example, ‘the company that Google
acquired’ vs. ‘the company that Google competes with’ – we can represent the world more
accurately. And that means better, more meaningful responses when we search.”
(Dyson, EDventure Holdings)

“Research suggests people prefer to state their information need rather than use keywords.
[...] Information worded as questions is increasing on the Web.” (Hearst, UC Berkeley)

“Specialists and enthusiasts will define taxonomies and ontologies: data sets that describe
classes of objects and relations among them. These sets will help computers everywhere to
find, understand and present targeted information.” (Shadbolt and Berners-Lee, 2008)

Taking these as a starting point, discuss the future of search. Justify your views and
conclusions, giving appropriate examples to back up your arguments.

(20 marks)