Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Topics in Advanced Information Retrieval

Date: Tuesday 24th May 2011
Time: 09:45 - 11:45

Please answer any THREE Questions from the five questions provided
Each question is worth 20 marks
Do not answer more than the required number of questions. Additional questions will not be marked. Clearly cross out anything you do not wish to be marked.

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

a) Jansen et al. (2008) classify user needs of Information Retrieval into informational, navigational and transactional needs.

i) Briefly describe these three needs and give two examples of each kind of need.

(3 marks)

ii) In their analysis of user queries, Jansen et al. assigned each query to a single class of need. Discuss to what extent this is a reasonable methodology, bringing into your answer considerations of the notions of task, information need, intent, query and relevance. Give justifications for your position.

(2 marks)

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

- doc1: anti-government protests rock Middle East rulers
- doc2: Middle Eastern governments adopt rock-hard positions
- doc3: Governor Rock protests Middle East’s rulers’ positions
- doc4: Governor’s “50:50 East/West” R.O.C.K. adoption protest

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 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 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.

rock AND protest AND government

(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) A user gives a query consisting of two words. For one word, the postings list is:

\[4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 122, 157, 180\]

For the other word, the postings list is:

\[47\]

State how many comparisons would be required to intersect the two postings lists

i) Using the bare lists above. 

(1 mark)

ii) Using these lists augmented with skip pointers with a skip length of \(\sqrt{L}\), where \(L\) is the length of the postings list.

(2 marks)

Ensure that you briefly justify your answers.

d) A user asks a query “rockets” of an Information Retrieval system based on the basic Boolean model. The first document he gets back is the following one:

“Baking a cake is easy. It is not rocket science.”

Discuss the advantages and disadvantages of the basic Boolean model in Information Retrieval, explaining why the above result is to be expected. Exemplify your answer and justify your conclusions.

(5 marks)
2.

a) One the one hand, a stemming algorithm fails to map “adhere” and “adhesion” to a common stem. On the other hand, it unhelpfully maps both “experiment” and “experience” to the stem “experi”.

i) What are the technical terms specifically used to describe this behaviour of stemming algorithms?

(1 mark)

ii) “For IR purposes, it doesn’t usually matter whether the stems generated are genuine words or not.” (Paice)

Why is this the case?

(1 mark)

iii) “Stemming, in general, has not been an unmitigated success in improving IR.” (Kantrowitz et al., 2000)

To what extent do you agree with this view? Justify your answer.

(2 marks)

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

- Tokenisation
- Use of a stop list
- Mapping to lower case
- Normalisation of hyphenation and of acronym punctuation
- Asymmetric expansion
- Accented characters
- Transliteration
- Synonymy

(4 marks)

c) What is the inverse document frequency for a word that occurs in every document of a collection?

(1 mark)

d) Manning et al. (2008) note that “relevance does not increase proportionally with term frequency of occurrence” and that “rare terms are more informative than frequent terms”.

Explain how we can we use such notions to weight terms and to rank documents for queries.

(2 marks)

[Question 2 continues on the following page]
e) A user makes a query consisting of one term. What relevance ranking behaviour should we expect of a retrieval system using inverse document frequency? Justify your answer.

   (1 mark)

f) In relation to the Vector Space Model, three documents (D1, D2, D3) are analysed and results reported as follows:

\[
\begin{align*}
\text{cosine}(D1, D2) &= 0.94 \\
\text{cosine}(D1, D3) &= 0.79 \\
\text{cosine}(D2, D3) &= 0.69
\end{align*}
\]

i) What are these results telling us about these three documents?

   (2 marks)

ii) Why might you be interested in knowing whether or not the L2 norm had been applied during the analysis?

   (1 mark)

g) “IR research has clearly demonstrated that the user’s needs can be met by applying essentially simple but soundly based statistical techniques.” (Spärck Jones, 2003) Discuss, setting out your position, giving appropriate justification and exemplification.

   (5 marks)
3.

a) The formula for Precision is:
\[
\frac{TP}{TP+FP}
\]
and that for Recall is:
\[
\frac{TP}{TP+FN}
\]
where ‘T’ is ‘true’, ‘F’ is ‘false’, ‘P’ is ‘positives’ and ‘N’ is ‘negatives’.

Using the above, give the Precision and Recall of a system that returns 3 relevant documents and 2 irrelevant documents for a collection where there are known to be 8 relevant documents. Show your working. Express the final scores as a fraction.

(2 marks)

b) What is the advantage of using F measure, the harmonic mean of Precision and Recall, rather than the arithmetic mean?

(2 marks)

c) Consider the following chart showing 11-point interpolated precision-recall for two different retrieval systems. You have been asked to undertake a systematic review on a particular subject, thus you need to aim to retrieve all documents on that subject and be as sure as you can be that you are not missing any. Which one of the two systems would you choose to carry out this task? Justify your choice, referring to the values in the chart.

(3 marks)
d) I wish to evaluate my retrieval system in a Web search environment. One colleague recommends that I use precision-at-$k$, another recommends that I use 11-point interpolated precision-recall. Which method should I choose to evaluate my search engine? Justify your answer, citing advantages and disadvantages.

(3 marks)

e) Compare and contrast $n$-gram based phrase indexing against an indexing approach that firstly applies a linguistic filter to yield sequences of adjectives and nouns, then uses a multi-word term extraction technique to yield multi-word index entries.

(2 marks)

f) As a memory aid, a 10th century codex (manuscript book) held in the British Library (Br. Mus. MS. Or. 4445) records the counts of common words and phrases that appear at the beginning and end of verses. (It is thought that readers would train themselves to recall all occurrences of a word or phrase in this religious text, hence the need for such an aid.)

i) Identify the modern-day Information Retrieval technique that is analogous to this approach.

(1 mark)

ii) What are the advantages and disadvantages of the modern-day technique?

(2 marks)

g) “A considerable fraction of Web queries contain named entities. This […] imposes the ever-increasing need that search engines handle efficiently named entity queries.” (Stamou & Kozanidis, 2009)

Elaborate on challenges and problems that may arise in attempting to handle named entities, at both indexing and query stages, giving examples, and discuss to what extent and by which means users may be prepared to indicate named entities in their queries.

(5 marks)
4. Consider the following fragments, each taken from a different XML document instance:

<book title="RDF saved my life">
  <author>Edzell Snidgewort</author>
</book>

<writer name="Edzell Snidgewort">
  <document>RDF saved my life</document>
</writer>

<work work_author="Edzell Snidgewort" work_title="RDF saved my life"/>

i) Why would these fragments hinder interchange of information by systems? 

   (1 mark)

ii) How would you translate these fragments into Resource Description Framework (RDF) format? You may use a graphic representation, informal triple representation or RDF/XML representation.

   (3 marks)

b) RDFS is described as a “primitive ontology language” by Antoniou & van Harmelen (2008). To what extent do you agree with this view? Justify your position with appropriate argumentation and examples.

   (3 marks)

c) A multinational manufacturer with a large archive of XML instance documents of many different types, for both internal and external use, seeks advice on whether it should undertake a mass conversion of its archive to RDF format. Set out advantages and disadvantages of conversion, and give your recommendations. Justify your conclusions and recommendations.

   (4 marks)

d) The main requirements for an ontology language according to Antoniou & Van Harmelen (2008) are: a well-defined syntax; a formal semantics; convenience of expression; efficient reasoning support; and sufficient expressive power. The W3C Web Ontology Working Group has defined however three sub-languages of OWL in response to such requirements. Say what these three varieties are, why they are deemed necessary and describe the relationships among them.

   (4 marks)
[Question 4 continues from the previous page]

e) Consider the two following quotations:

“'If I can make sure that all of my trading partners and all of my systems in my divisions are speaking the same language, then that will solve 90 percent of my problems,' Ronald Schmelzer, analyst for ZapThink LLC, says of the Semantic Web. ‘People are not at the point where they need to talk to arbitrary systems that they don't know the semantics for.’” (www.techweb.com)

“Today searching Google for "Toyota used cars for sale in western Massachusetts under $8,000" returns more than 2,000 general Web pages. Once Semantic Web capabilities are added, a person will instead receive detailed information on seven or eight specific cars, including their price, color, mileage, condition and owner, and how to buy them. [...] 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)

Discuss, giving your views on the Semantic Web in terms of supporting ontology-based search as well as interoperability. Include discussions of issues of scale and feasibility, and of near-term prospects. Justify your views, giving appropriate examples to back up your arguments.

(5 marks)
5. Consider the following quotations:

“The real question is not will search be better, but rather how will it be better?”
(Mayer, Google)

“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)

“The future of search is verbs.” (Gates, Microsoft)

“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)

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)