Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Topics in Advanced Information Retrieval

Date: Tuesday 25th May 2010
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) Explain how, in Information Retrieval, \textit{query}, \textit{task}, \textit{information need} and \textit{relevance} are related, discussing pitfalls that may arise in the information seeking process up to the point that a user inputs one or more search terms to a conventional search engine.  

(4 marks)

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

- doc1: eurozone agrees bailout deal for Greece
- doc2: more cracks in the eurozone
- doc3: euro zone agrees bailout for Greece
- doc4: euro-deal boosts confidence in Greece

i) Draw the binary term-document incidence matrix for this collection.  

(2 marks)

ii) Draw the basic inverted index representation for this collection.  

(2 marks)

(In relation to i and ii, \textit{briefly} explain and justify any decisions you may have made regarding what you take to be a term.)

iii) For this collection, using the inverted index you have established, what would be the result of the following query? 

\textit{eurozone AND Greece}

Demonstrate how you arrive at your answer.  

(2 marks)

c) What is the main disadvantage of a binary term-document incidence matrix?  

(1 mark)

d) Why would I wish to keep document frequency information for a term (how many documents the term occurs in) in an inverted index?  

(2 marks)

e) “Today’s Web may be defined as the Syntactic Web.” (Breitman et al., 2007). What do you understand by this characterisation of the Web? In your answer, include:

- Explanation of how the Syntactic Web and the Semantic Web are related.
- Examples of types of task that would be hard to accomplish via the Syntactic Web.
- Discussion of how the Syntactic Web impacts upon the nature of the search experience for the user, with appropriate exemplification and justification.  

(7 marks)
2. a) Say if the following are true or false, justify your decisions.
   i) In a Boolean retrieval system, stemming never lowers precision.
   ii) In a Boolean retrieval system, stemming never lowers recall.  

(2 marks)

b) What kind of errors can occur when stemming is used?  

(2 marks)

c) In designing my indexing routines, I decide to tokenize on white space, 
not carry out case-folding, and not do stemming. Discuss the implications 
of these particular decisions, and discuss the advantages and 
disadvantages of normalisation in general.  

(4 marks)

d) Consider the following table, based on a collection where the number of 
documents N=1,000,000:

<table>
<thead>
<tr>
<th>Term</th>
<th>Document frequency of term ( (df_t) )</th>
<th>Inverse document frequency of term ( (idf_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>calpurnia</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>animal</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>sunday</td>
<td>1000</td>
<td>3</td>
</tr>
<tr>
<td>fly</td>
<td>10000</td>
<td>2</td>
</tr>
<tr>
<td>under</td>
<td>1000000</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>10000000</td>
<td>0</td>
</tr>
</tbody>
</table>

i) What is the reasoning behind the calculation of inverse document 
frequency \( (idf) \)?  

(1 mark)

ii) I now calculate how many times each term appears in each 
document, to yield \( tf_{t,d} \). I then calculate a \( tf-idf \) weight for each term 
in each document. I note that, for some term, its weight is less for 
document \( A \) than for document \( B \). What is this telling me? How 
would I use this information to rank documents for relevance in 
response to a query?  

(3 marks)

e) Explain two ways in which cosine similarity scores are used with the 
vector space model.  

(2 marks)

(Question 2 continues on the following page)
f) I form a new document by taking a copy of a document and appending it to itself. This new document evidently has the same semantic content as the original. What steps do I have to take to ensure that a cosine similarity calculation does not cause me to think that these two documents are quite different? (1 mark)

g) “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. (5 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

b) Explain why, in information retrieval, precision and recall measures are preferred to a measure of accuracy.

c) What is the advantage of using F measure, the harmonic mean of precision and recall, rather than the arithmetic mean?

d) My search engine produces ranked query results when running over a test collection with known relevance judgements for the set of queries used.

i) Explain how precision-recall curves and interpolated precision can be used to evaluate the search engine's performance.

ii) I now transfer my search engine to a web search environment. Explain the advantages and disadvantages of using precision at a fixed low level of retrieved results (precision-at-k) to evaluate the system in this environment.
(Question 3 continues from the previous page)

e) Consider the following documents, D1 and D2:

D1: A blind Venetian Doctor (Tomaso Giovanni Bertinato) has provided a new accessibility tool for Microsoft ® Office Windows ™ in under twenty four hours.

D2: Office accessibility for the on-call Doctor (Twenty Four Hours) is provided via the surgical tool store entrance. The office windows are provided with a 'Bertinato' venetian blind.

i) An information retrieval system based on a simple bag of words Boolean model with no use of weighting (e.g., via tf-idf) would return both documents for queries such as “venetian AND blind”, “blind AND accessibility”, “doctor AND bertinato AND provided AND tool” or “windows AND tool AND accessibility”. How could phrase indexing and positional indexing allow users to express their information need more precisely in queries and to achieve greater discrimination in results? (4 marks)

ii) What techniques from text mining could help users achieve more focussed search results and what problems might such techniques face when confronted with these two documents? (5 marks)
4. a) Consider the following fragments taken from two XML documents:

```xml
<course name="Discrete Maths">
    <lecturer>David Billington</lecturer>
</course>

<lecturer name="David Billington">
    <teaches>Discrete Maths</teaches>
</lecturer>
```

i) Why would these fragments represent a problem for processing by machine? (1 mark)

ii) How would you translate the fragments in 4a into Resource Description Framework (RDF) format? You may use a graphic representation, informal triple representation or RDF/XML representation. (2 marks)

b) RDF, RDFS, SPARQL and OWL are key enabling languages for the Semantic Web. Explain the contribution of each to rendering the web 'semantic'. (5 marks)

c) In a Resource Description Framework (RDF) document, when would you use “rdf:ID” and when “rdf:about”? (2 marks)

d) 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)

e) I own two companies, and have been told that named entity recognition (NER) can be used to improve the performance of their search engines. One company provides access to general news items, the other provides access to scientific journals. Write a report (adopting the viewpoint of an objective external consultant) explaining the role of NER in a search engine context, a summary of the advantages and disadvantages of NER, and recommendations on deploying NER for both search engines, for only one (state which and why) or for neither (with justification). (6 marks)
5. Consider the following quotations:

“Text mining will catch and eventually dwarf traditional information retrieval.” (Rao, Queue, 2004).

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

“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)

“Sophisticated search engines like Bing are doing a better job of understanding structure and recognizing the meaning behind words based on context and other factors.” (Prevost, Microsoft Bing, 2010)

Taking these as a starting point, discuss the current state of semantic search. You should comment, for example, on what kind of obstacles there might be, on enabling technologies that hold out promise of near-term progress in overcoming the obstacles you identify, on any areas where semantic search has shown benefits for users and areas where benefits are perhaps more elusive and why. Justify your views and conclusions, giving appropriate examples to back up your arguments. (20 marks)