Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Topics in Advanced Information Retrieval

Date: Tuesday 21st May 2013
Time: 09:45 - 11:45

Please answer any THREE Questions from the FIVE Questions provided

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.

(2 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: Near-Miss Asteroid Rushes Past Earth.
    doc2: Miss I.S. Rush "Is Meteorite Fatality".
    doc3: Rare Earths From Meteor-showers?
    doc4: "Earthstorms: Near Misses In Earth’s Fateful Past" – Meteorologist Frøm F’Atål.

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

(2 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.

(earth OR meteor) AND (fatal OR miss)

(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.)

[Question 1 continues on the following page]
c) A user gives a query: web AND arachnid. For ‘web’, the postings list is:
   \[4, 6, 10, 12, 14, 16, 18, 20, 22, 32, 47, 81, 120, 122, 157, 180\]

   For ‘arachnid’, the postings list is:
   \[47\]

   If these lists are now augmented with skip pointers with a skip length of \(\sqrt{L}\), where \(L\) is the length of the postings list, state how many comparisons would be required to intersect the two postings lists. **Ensure that you briefly justify your answer.**

   (2 marks)

d) Consider the following index terms and their postings list sizes:

   \[
   \begin{array}{|c|c|}
   \hline
   \text{Index term} & \text{Posting list size} \\
   \hline
   \text{biscuit} & 213313 \\
   \text{chocolate} & 87010 \\
   \text{lemonade} & 107914 \\
   \text{orange} & 271659 \\
   \text{sugar} & 46654 \\
   \text{toast} & 316813 \\
   \hline
   \end{array}
   \]

   A user gives the query:

   (sugar OR toast) AND (lemonade OR orange) AND (chocolate OR biscuit)

   Recommend a query processing order for this query. **Justify your recommendation.**

   (2 marks)

e) “The Boolean model is arguably the simplest model to base an information retrieval system on.” (Schütze). Discuss the advantages and disadvantages of adopting the Boolean model for information retrieval. **Give appropriate examples and justifications.**

   (4 marks)
2.

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

i) Tokenisation
ii) Normalisation (e.g., mapping to lower case, acronym punctuation normalisation)
iii) Asymmetric expansion
iv) Stemming
v) Transliteration
vi) Synonymy

(6 marks)

b) Consider the following table, based on a collection where the number of documents N=1000000:

<table>
<thead>
<tr>
<th>Term</th>
<th>Document frequency of term (df)</th>
<th>Inverse document frequency of term (idf)</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>100000</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1000000</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?

(2 marks)

How would I use this information to rank documents for relevance in response to a query?

(1 mark)

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

(1 mark)

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

(2 marks)
e) A colleague recommends that I use the *inverse of Euclidian distance* as a similarity metric for my vector-based retrieval system. Do you agree? *Explain your decision.*

   (1 mark)

f) A user gives a query where one of his query terms does not occur in the retrieval system’s indexed terms. This implies that a vector for this query would not be in the vector space of the collection. How should we adapt the representation of vector space to handle this case? *Give your reasoning.*

   (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) There are 100 relevant documents in a collection of 10000. My goal is to predict which ones are relevant. From the 10000, I select 200 documents in the hope of having predicted many relevant documents among these 200. I then note how many times I was correct or wrong in my predictions, and establish the following table:

<table>
<thead>
<tr>
<th></th>
<th>Predicted Non-relevant</th>
<th>Predicted Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-relevant documents</td>
<td>9760</td>
<td>140</td>
</tr>
<tr>
<td>Relevant documents</td>
<td>40</td>
<td>60</td>
</tr>
</tbody>
</table>

i) How many documents are false negatives?
ii) How accurate are my predictions (percentage correct)?
iii) What percentage of the relevant documents did I correctly identify?
iv) What percentage of the predictions of relevant documents was correct? (2 marks)

b) Which of the following attributes does the F-measure capture directly? Give the complete list of those you think apply. If you think none of these apply, say so.
   i) The size of the set of irrelevant documents.
   ii) The relative ranking between relevant documents.
   iii) Recall.
   iv) Precision.
   v) The relevance of the highest-ranked documents. (2 marks)

[Question 3 continues on the following page]
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 (d6):


The evaluator knows there are 6 relevant documents for this test query in the gold standard collection: d4, d13, d16, d22, d23, d31. He notes that the relevant documents returned in the ranked list occur at positions 2, 3, 5 and 10.

i) Give both the recall and the precision values for each of the returned 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. It is sufficient to be accurate to 2 decimal places, with no rounding.

(4 marks)

iii) What is the R-precision value?

(1 mark)

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) 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.

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)
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 general Web-scale access, the other provides access only to scientific journals. I retain you as a consultant and ask you to write a report 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)
4.

a) Consider the following two RDF descriptions:

```xml
<?xml version="1.0"?>
<River xmlns="http://www.geodesy.org/river#">
  <name>Yangtze</name>
  <length>6300 kilometers</length>
  <startingLocation>western China's Qinghai-Tibet Plateau</startingLocation>
  <endingLocation>East China Sea</endingLocation>
</River>

<?xml version="1.0"?>
<River xmlns="http://www.geodesy.org/river#">
  <name>Dri Chu - Female Yak River</name>
  <name>Tongtian He, Travelling-Through-the-Heavens River</name>
  <name>Jinsha Jiang, River of Golden Sand</name>
</River>
```

Explain why an attempt at aggregating from these descriptions would fail, indicating what change(s) would be needed to ensure successful aggregation.

(2 marks)

b) Consider the following (partial) RDF Schema (RDFS) document:

```xml
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xml:base="http://www.geodesy.org/water/naturally-occurring">
  <rdfs:Class rdf:ID="River">
    <rdfs:subClassOf rdf:resource="#Stream"/>
  </rdfs:Class>

  <rdfs:Class rdf:ID="Stream">
    <rdfs:subClassOf rdf:resource="#NaturallyOccurringWaterSource"/>
  </rdfs:Class>

  <rdf:Property rdf:ID="emptiesInto">
    <rdfs:domain rdf:resource="#River"/>
    <rdfs:range rdf:resource="#BodyOfWater"/>
  </rdf:Property>

  <rdf:Property rdf:ID="length">
    <rdfs:domain rdf:resource="#River"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
  </rdf:Property>
...
</rdf:RDF>
```

[Question 4 continues on the following page]
In the following separate RDF/XML document, there is no indication of the class of the resource 'Yangtze' (lines are numbered for ease of reference):

(1) <?xml version="1.0"?>
(2) <rdf:Description rdf:ID="Yangtze"
(3)   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
(4)   xmlns="http://www.geodesy.org/water/naturally-occurring#">
(5) <length>6300 kilometers</length>
(6) <emptiesInto rdf:resource="http://www.china.org/geography#EastChinaSea"/>
(7) </rdf:Description>

What inferences can we make from the RDF/XML document using the RDFS document? *Explain in detail how each inference you identify is arrived at.*

(4 marks)

c) State whether we can make the same inferences you have identified in 4b, if line 6 of the RDF/XML document is changed to:

(6) <emptiesInto>East China Sea</emptiesInto>

*Explain your answer.*

(2 marks)

d) RDFS is described as a “primitive ontology language” by Antoniou & van Harmelen (2008). Discuss, taking into account also the relationship of RDFS to other Semantic Web ontology languages you are familiar with.

(4 marks)
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. [...] special-
sists 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,
assessing to what extent you would consider the above positions of a few years ago to be
more, or less, applicable today. Justify your views and give appropriate examples to back
up your arguments.

(8 marks)
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)

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

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

“I want you to remember three words: ‘answers, not links’.” (Seth, Yahoo!)

"Asking a question isn’t the same as searching." (Ask Jeeves)

“Our goal is to build on the achievements of science and other systematizations of knowledge to provide a single source that can be relied on by everyone for definitive answers to factual queries.” (Wolfram|Alpha)

Taking these as a starting point, discuss the challenges and potential solutions for search engines that aim to handle questions, semantics and fact extraction, and to deliver appropriate answers rather than lists of documents to read. Justify your views and conclusions, giving appropriate examples to back up your arguments.

(20 marks)