Documents, Services and Data on the Web

Date: Tuesday 26th May 2015
Time: 14:00 - 16:00

Please answer any TWO Questions from the THREE Questions provided.

Use a SEPARATE answer book for each QUESTION

Each question is worth 30 marks.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]
1. **Documents on the Web**

a) Discuss strategies a search engine might use to attempt a classification of user needs as expressed via queries, commenting on advantages and disadvantages of the strategies you describe. (3 marks)

b) i) I compose the following query for a search engine that supports the Boolean model:

```
ebola AND vaccine AND africa
```

Internally, the search engine re-orders my query to be:

```
africa AND ebola AND vaccine
```

Why would it do this? (2 marks)

ii) Explain how the following query would typically be processed in terms of order:

```
spring OR surge) AND (storm OR gale)
```

(2 marks)

iii) Set out advantages and disadvantages of the Boolean model, discussing scenarios where this model may be preferred and why. (3 marks)

c) You have been asked to build a system to produce inverted indexes and are at the stage of deciding on pre-processing steps to apply to the input documents, which could come from any domain. You consider the following options:

i) Use an available off-the-shelf tokenizer. It is reported to achieve 93% F-score on business news text.

ii) Train (or write rules for) your own tokenizer.

iii) Use the Porter stemmer algorithm.

iv) Carry out case folding.

v) Carry out spell checking.

Explain what criteria you would use to help you reach a decision on which of these options to choose. Comment on advantages and disadvantages of particular choices and combinations of choices, and on the impact of potential dependencies among choices. (5 marks)
[Question 1 continues from the previous page]

d) i) 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 use such notions to weight terms and to rank documents for queries. (2 marks)

ii) 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)

e) Explain two ways in which cosine similarity scores are used with the Vector Space Model. (2 marks)

f) Consider the graph in figure 1. Probability mass is evenly distributed, in the initial state. We apply the PageRank algorithm, but do not use a damping factor and assume no random jumps. What would be the value for node B after 2 iterations? Values may be given to 3 decimal places. Show your working. (3 marks)

Figure 1: A graph with 5 nodes
[Question 1 continues from the previous page]

g) Considering the graph in figure 1, assume that we delete the following links:

- from B to C
- from D to C
- from E to D

Explain how these changes may adversely affect a subsequent PageRank calculation and describe one approach that will reduce the severity of these changes. (1 mark)

h) “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)

“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)

Discuss how techniques of semantic search, named entity extraction and fact extraction, and other techniques you may identify, can help in bringing about the evolution in search referred to in the above statements. In your answer, discuss also challenges and barriers that may arise in relation to these techniques, especially in a Web environment requiring massive indexing. Justify your views and conclusions, giving appropriate examples to back up your arguments. (6 marks)
2. **Services**

a) Enumerate and discuss three of the most important benefits of the Cloud Computing IT infrastructure paradigm. You may illustrate your discussion with examples.  

   (3 marks)

b) Suppose you are an IT professional working for a Cloud company, whose primary job is to sell to potential customers Cloud computing services. At a meeting with the Chief Executive Officer (CEO) of a mobile phone company that requires Customer Relationship Management (CRM) services, what advantages of Software as a Service (SaaS) services for CRM would you mention to persuade the CEO to buy Cloud services rather than investing in the provision of CRM services by his own company’s staff? Justify your answer by discussing each advantage.  

   (3 marks)

c) Enumerate and discuss three issues that arise from Cloud computing.  

   (3 marks)

d) Provide a definition and explain the applicability and advantages of each of following metrics in the decision as to whether or not adopt Cloud Computing to support your business: Net Present Value (NPV) and Return on Investment (ROI). Use examples to illustrate your answer.  

   (5 marks)

e) Consider the following purchase costs related to technology and personnel for a growing biotechnology company, which hosts a customer database, as well as research libraries, and sales and finance applications.

   Disk Storage: $2,000,000 (with an operational lifetime of 3 years).
   Disk maintenance: 10% of cost of Disk Storage each year.
   Firewalls: $10,000 per year.
   Network switches: $10,000 per year.
   Server Hardware: $500,000 (with an operational lifetime of 3 years).
   Server Maintenance: 10% of cost of Server Hardware each year.
   Software Licences: $200,000.
   Labour: $150,000 per year.

[Question 2 continues on the following page]
i) Calculate the Total Cost of Ownership (TCO) for the above scenario considering a period of three years.

ii) Now assume that the company is growing every year and so, its storage, software and labour needs increase 20% each year. The current costs are $500,000 for storage, $50,000 for software and $75,000 for labour. Taking this scenario as an example and considering a period of three years starting from the present time, discuss why or why not it may be a better investment to use a Cloud.

(8 marks)

f) Suppose a company that sells mobile phones is considering moving all its business processes to be performed on a Cloud environment, in view of the fact that the business is growing at a fast pace and, soon, its local data centre will probably become insufficient as well as obsolete in terms of hardware and software. Assuming that the necessary investment cost for this move to be possible is $2,000,000 and the expected income for the next few years is $700,000 per year, in how many years would the return from the investment be enough to justify it, assuming a discount rate of 10% per year? Justify your answer.

(5 marks)

g) Explain the fundamental problem of processing "big" data and how the MapReduce programming model can help minimizing this problem. (3 marks)
3. Data on the Web

a) How do Linked Data resources 'link' with each other? Explain the three types of RDF links. (4 marks)

b) Explain why re-using existing vocabularies in your linked data set would be beneficial. Provide an example. (2 marks)

c) Consider this RDF/XML document:

```xml
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:exterms="http://www.example.org/terms/">
    <rdf:Description rdf:about="http://www.example.org/index.html">
        <exterms:creation-date>August 16, 1999</exterms:creation-date>
        <dc:language>en</dc:language>
        <dc:creator rdf:resource="http://www.example.org/staffid/85740"/>
    </rdf:Description>

    <rdf:Description rdf:about="http://www.example.org/staffid/85740">
        <dc:title>James</dc:title>
    </rdf:Description>
</rdf:RDF>
```

i) Draw an equivalent RDF graph for the above RDF/XML description. (2 marks)

ii) Write a Turtle serialisation of the above RDF/XML description. (3 marks)

d) Explain the aim and motivation for the RDFa serialisation of RDF. (2 marks)

e) What is the role of RDF-Schema and OWL in the context of linked data? Why is RDF-Schema insufficient as a knowledge representation? (4 marks)

[Question 3 continues on the following page]
f) Discuss different options for storing RDF data. What are their advantages and disadvantages? (4 marks)

g) Explain what is the task that the following SPARQL query aims to address: (3 marks)

PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT ?f ?d
WHERE
{ ?f rdf:type dbpedia:Film .
  ?f dbpedia:releaseDate ?d .
  FILTER ((?d >= "2000-01-01"^^xsd:date) && (?d < "2000-02-01"^^xsd:date))
}

h) The University considers publishing its time-tableing data as open linked data. Discuss potential benefits, disadvantages and a wider impact that such a solution could have. For example, consider what one could do with such a dataset (possible applications); whether providing an Excel file with all the data would (not) be appropriate; what are the maintenance costs, etc. (6 marks)