COMP 38120

Two hours - online hybrid

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Documents, Services and Data on the Web

Date: Friday 31st May 2019

Time: 09:45 - 11:45

This is a hybrid examination with sections to be answered online and questions to be answered on paper.

Please answer ALL Questions in Section A and Section C in separate answerbooks
Please answer ALL Questions in Section B online

This is a CLOSED book examination

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

[PTO]

Section A
Answer this Section in the answerbook provided

1. Documents on the Web

a) A user has composed the following query for a search engine that supports the Boolean model:

ebola AND vaccine AND africa

However, the search engine internally re-orders the query to:

africa AND ebola AND vaccine

Why would it do this?

(2 marks)

b) Explain the tf-idf (term frequency, inverse document frequency) weight and its role in document retrieval.

(2 marks)

c)

i) Consider the graph in Figure 1. We apply the PageRank algorithm, but do not use damping factor and assume no random jumps. In the initial state, probability mass is evenly distributed. After two iterations, what would be the value for node **E**? Values may be given to 3 decimal places. Show your working and explain the process you have followed.

(4 marks)

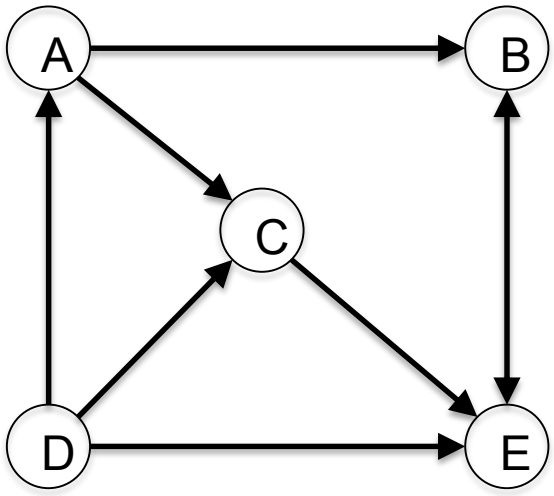

Figure 1: A graph with 5 nodes

[PTO]

ii) If a node with IN links but no OUT links is introduced in figure 1, explain what effect this could have on the PageRank calculation and how that effect could be overcome.

(2 marks)

d) There are different so-called off-line measures of quality of document retrieval. Explain *Precision at k* and discuss its advantages and disadvantages.

(2 marks)

e)

i) A user wants to query a search engine to learn about the weather in Manchester. What kind of information need does the user have? Explain your answer.

(2 marks)

ii) In order to better understand the user's information need, a search engine would like to be able to 'predict' temporality (i.e. temporal dimension) of a query. For example, they would like to predict whether a query refers to past, current or future. Discuss why would this be a relevant feature to add to the search engine and what approaches you would suggest to implement in order to support the task. Explain your answer.

(6 marks)

[PTO]

Section B
Answer this Section online

2. Services on the Web

a) Give two examples of ways in which Cloud Computing has changed how businesses and individuals interact in their collaboration efforts.

(2 marks)

b) A higher education institution decides to use services from a public cloud provider, but it is not sure about which of the three delivery models to use (SaaS, PaaS or IaaS). In your opinion, which of the three models should the institution embrace? Justify your answer, considering examples of applications that would be beneficial to the students, as well as the potential impact of distance learning.

(8 marks)

c) Consider the following purchase costs related to technology and personnel for a growing IT department, which hosts a database as well as a multitude of applications.

- Disk Storage: $1,200,000 (with an operational lifetime of 1 year).
- Disk maintenance: 15% of cost of the annual Disk Storage cost, each year.
- Firewalls: $5,000 per year.
- Network switches: $10,000 per year.
- Server Hardware: $200,000 (with an operational lifetime of 3 years).
- Server Maintenance: 10% of annual Server Hardware cost, each year.
- Software Licences: $20,000 per year.
- Labour: $900,000 per year

Calculate the Total Cost of Ownership (TCO) considering a period of three years and discuss the limitations of TCO as the sole metric for supporting the decision as to whether or not to move to a Cloud solution.

(5 marks)

d) Using the MapReduce programming model, write a Reduce function, using pseudo-code, to join two tables with one-to-one relationships, whose schemas are defined in the following. Describe the assumptions you are making for the corresponding Map function.

Schema for Table A: *(employee_id, employee_name, dept_id)*

Schema for Table B: *(departm_id, departm_name, departm_postcode)*

Join operation: *A.dept_id = B.departm_id*

(5 marks)

Section C
Answer this Section in the answerbook provided

3 Data on the Web

a) Explain how dereferencing of HTTP URIs works in terms of the steps that the 303 URI strategy involves.

(4 marks)

b) What are the different types of RDF Links, and how do they differ from each other in terms of purpose?

(3 marks)

c) Draw an RDF graph equivalent to this RDF/XML document:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:coffee="http://purl.org/ontology/coffee/"
    xmlns:milk="http://purl.org/ontology/milk/ ">

<rdf:Description rdf:about="http://purl.org/ontology/coffee/DecafAlmondCappucino">
    <rdf:type rdf:resource="http://purl.org/ontology/coffee/Coffee"/>
    <coffee:CaffeineStrength>0</coffee:CaffeineStrength>
    <coffee:MilkType rdf:resource="http://purl.org/ontology/milk/AlmondMilk"/>
</rdf:Description>

<rdf:Description rdf:about="http://purl.org/ontology/milk/AlmondMilk"
rdfs:label="almond milk">
    <rdf:type rdf:resource="http://purl.org/ontology/milk/PlantMilk"/>
    <milk:Source>Almonds</milk:Source>
</rdf:Description>
```

(6 marks)

[PTO]

d) RSS (Rich Site Summary, also known as Really Simple Syndication or RDF Site Summary) is used to deliver regularly changing web content in a standard XML format that conforms to the W3C's RDF Specification. It allows the syndication of lists of hyperlinks, along with other information that helps viewers decide whether they want to follow the link. RSS is extensible via the associated XML-namespace (http://purl.org/rss/1.0/) and has an RDF Schema (RDF(S)) that defines, among other things, classes *channel* and *item*, and properties *items* and *description.* Here are their definitions:

```
<rdfs:Class rdf:about = "http://purl.org/rss/1.0/channel"
        rdfs:label = Channel"
        rdfs:comment = "An RSS information channel.">
        <rdfs:isDefinedBy rdf:resource="http://purl.org/rss/1.0/"/>
</rdfs:Class>

<rdfs:Class  rdf:about = "http://purl.org/rss/1.0/item"
        rdfs:label="Item"
        rdfs:comment="An RSS item.">
        <rdfs:isDefinedBy rdf:resource = "http://purl.org/rss/1.0/"/>
</rdfs:Class>

<rdf:Property rdf:about = "http://purl.org/rss/1.0/items"
        rdfs:label="Items"
        rdfs:comment="Points to a list of rss:item elements that are members
                        of the subject channel.">
        <rdfs:isDefinedBy rdf:resource = "http://purl.org/rss/1.0/"/>
</rdf:Property>

<rdf:Property rdf:about="http://purl.org/rss/1.0/description"
        rdfs:label="Description"
        rdfs:comment="A short text description of the subject.">
        <rdfs:subPropertyOf rdf:resource =
"http://purl.org/dc/elements/1.1/description"/>
        <rdfs:isDefinedBy rdf:resource = "http://purl.org/rss/1.0/"/>
</rdf:Property>
```

[PTO]

and here is an example RSS feed:

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns = "http://purl.org/rss/1.0/">

<channel>
        <title>BBC Nature Video collections</title>
        <link>http://www.bbc.co.uk/nature/collections</link>
        <description>A vast array of wildlife video clips.</description>

        <items>
                <rdf:Seq>
                        <rdf:li
resource="http://www.bbc.co.uk/nature/collections/p011x1yq"/>
                        <rdf:li
resource="http://www.bbc.co.uk/nature/collections/p01btg6w"/>           </rdf:Seq>
        </items>
</channel>

<item rdf:about="http://www.bbc.co.uk/nature/collections/p011x1yq">
        <title>This is Planet Earth</title>
        <link>http://www.bbc.co.uk/nature/collections/p011x1yq</link>
        <description>Narrated by Sir David Attenborough</description>
        <pubDate>Fri, 21 Dec 2012 16:37:42 GMT</pubDate>
</item>


<item rdf:about=" http://www.bbc.co.uk/nature/collections/p01btg6w">
        <title> Sensational summer wildlife</title>
        <link>http://www.bbc.co.uk/nature/collections/p01btg6w</link>
        <description>Documents amazing opportunities to get close to the incredible wild
neighbours              on your doorstep.
        </description>
        <pubDate>Thu, 08 Aug 2013 16:56:23 GMT</pubDate>
</item>

</rdf:RDF>
```

i. Define a new subclass *vlog* of the *item* class that can be used to represent items that are video blogs.

(1 mark)

ii. Define property *duration* that can be used to represent an item's duration in terms of number of seconds. Note that this property only applies to the *vlog* class that you just created above. Make *duration* a subproperty of *http://purl.org/dc/elements/1.1/format*

(2 marks)

[PTO]

e) Refer to the following SPARQL query. Assuming that is meant to be run on a platform that contains information on people, what task is it aiming to address and what will it return?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?name ?birth ?death ?person
WHERE {
        ?person dbo:birthPlace :Manchester .
        ?person dbo:birthDate ?birth .
        ?person foaf:name ?name .
        ?person dbo:deathDate ?death .
        FILTER ((?birth >= "1915-01-01"^^xsd:date) && (?birth < "2015-01-01"^^xsd:date)) .
        FILTER (regex(?name, "son"))
}
```

Note that the <xsd:date> data type is used to represent date in the YYYY-MM-DD format. Why is it important to cast values to <xsd:date> in the above query?

(4 marks)

END OF EXAMINATION