Institutionen för datavetenskap

Linköpings universitet

* + - * 1. TENTAMEN

TDDD43 Advanced Data Models and Databases

August 29, 2015, 8-12

*Grades:* For a pass grade you need to obtain 50% of the total points.

*Instructions:* In addition to the instructions on the cover page:

* Write clearly.
* Start the answers to a question on a new page.
* If you make assumptions that are not given in a question, then clearly describe these assumptions. (Of course, these assumptions cannot change the exercise.)
* Give relevant answers to the questions. Points can be deducted for answers that are not answers to the question.
* Answer in English.
  + - * 1. LYCKA TILL!

1. **XML querying (3 + 2 = 5p)**

Study the following XML file:

1. What is the result of executing the following XPath expressions on the XML file?
2. //country/population
3. /geodata//member
4. //countries[country/@id=2]
5. Express *“Find the population of the country whose capital is Stockholm”* as an XQuery query.

<<?xml version="1.0" encoding="UTF-8"?>  
<geodata>

<countries>

<country id = “1” name=”Sweden” capital=”c1”>

<population>9,540,065</population>

</country>

<country id = “2” name=”Denmark” capital=”c3”>

<members>

<member>UN</member>

<member>EU</member>

</members>

</country>

<country id = “3” name=”Germany” capital=”c2”>

<population>81,726,000</population>

</country>

</countries>

<cities>

<city id=”c1”>

<name>Stockholm</name>

</city>

<city id=”c2”>

<name>Berlin</name>

</city>

<city id=”c3”>

<name>Copenhagen </name>

</city>

</cities>

</geodata>

1. **Data Guides (3p)**

Draw a strong data guide for the data model below.

person

friend

person

person

friend

friend

Science

Jane

Jack

123456

Jeff

interest

name

name

name

phone

1. **NoSQL databases (1 + 1 = 2p)**
2. Explain how the union operation in the relational algebra (UNION operator for relational databases) can be solved with the Map-Reduce workflow;
3. Compare vertical and horizontal scalability in the context of NoSQL databases.
4. **Information Retrieval (2,5 + 1,5 = 4p)**
5. Assume that we use the boolean model for information retrieval. Assume that we are only interested in the words ‘gene’, ‘enzyme’, ‘protein’ and ‘signal’.
   1. Give the query that represents all documents that contain 'gene', but not 'signal'.
   2. Compute the completed DNF (disjunctive normal form) of the query - make sure to show all steps in the computation.
   3. Give two examples of document representations that are answers to the query.
   4. If possible, rank the answers in question 3. If it is not possible, explain why it is not possible.
6. Assume that we use the vector model for information retrieval. Assume that we are only interested in the words ‘gene’, ‘enzyme’, ‘protein’ and ‘signal’. Assume that we have two documents in our document base. Document 1 contains 'enzyme' 5 times, 'gene' 10 times, 'protein' 0 times and 'signal' 8 times. Document 2 contains 'enzyme' 0 times, 'gene' 0 times, 'protein' 7 times and 'signal' 1 time.
   1. Explain tf and idf in the vector model.
   2. Give the document representations for Document 1 and Document 2 according to the tf-idf model.
7. **Description logics (2 + 3 = 5p)**
8. Define the semantics of:
   * *EXISTS PlaysAt.Team*
   * *FORALL PlaysAt.Team*
   * *Employee AND Team-member*
   * *ATLEAST 3 PlaysAt*
9. Does the following hold?

*Employee AND EXISTS PlaysAt.Team* is-a (is subsumed by) *Employee AND FORALL PlaysAt.Team*

Prove your answer using a tableau algorithm.

1. **Integration (3 + 1 = 4 p)**
2. Compare the global as view approach versus the local as view approach for integration of data sources regarding the following issues:

* easy/hard to add new data sources
* easy/hard query processing

Explain your answer for each issue with an example.

1. What does the query [Uniprot-description:kinase] > PDB in the SRS query language mean? Explain the constructors in the query.
2. **Ontologies (2 + 2 + 1 = 5p)**
   1. Describe from a knowledge representation point of view the different kinds of ontologies. For each kind, state which ontology components are represented.
   2. Give 4 different kinds of matchers. For each kind of matcher give an example and explain briefly what it does.
   3. Given a Tbox T, a set of atomic concepts C, an oracle Or and a set of missing is-a relations M, what is a GTAP(T,C,Or,M) in the context of debugging modeling defects?