University of Cape Town

Department of Computer Science

CSC3003s Final Exam Solution

2006

Marks **: 100**

Time **: 180 minutes**

Instructions:

- Answer all questions from Section A and 3 questions from Section B.
- Show all calculations where applicable.

SECTION A : ANSWER ALL QUESTIONS

Question 1: Databases [10]

Question 2: XML / Information Retrieval [10]

Suppose that you are a consultant designing a news website, where articles link to one another and to external sites.

a. Why would you opt for XHTML instead of HTML as a markup language? [1]

we can use XML tools to manipulate well-formed content [1]

b. You decide to build inverted files for filtering/ranking.

i. What is filtering?

ii. What is ranking?

iii. What are inverted files?

iv. How can you improve on the efficiency of storing your inverted files? [4]

filtering is the process of excluding documents that are probably not relevant [1]

ranking is the process or ordering documents according to estimated relevance [1]

inverted files are lists of the documents each term occurs in [1]

use differential encoding / compression [1]

c. After filtering, you would like to rank the documents. Name 2 possible algorithms that can be used for this purpose. Discuss one major difference between the 2 algorithms (besides execution time). [3]

Boolean/vector ranking [1/2]

PageRank [1/2]

Boolean ranking is based on term occurrences while PageRank is based on link structure. [2]

d. How would you ensure that your website is ranked highly in Google searches? [1]

make sure that external sites link to your site and your sites links elsewhere.[1]

e. You want to support non-English languages but your software only handles ASCII internally. How would you deal with this problem without violating the XML standard? [1]

use character entities for non-English characters [1]

SECTION B : ANSWER ANY 3 QUESTIONS ONLY

Question 3: Databases [10]

Question 4: Databases [10]

Question 5: XML [10]

a. Answer the following questions based on this piece of XML:

```
<exam xmlns=”http://ns1”>
   <name>CSC3003s</name>
   <venue>Jameson</name>
</exam>
```

Assume that the **name** and **venue** elements must both occur exactly once.

i. Write an XML Schema complexType type definition **examType** corresponding to the content of the **exam** element and its descendents. [4]

```
<complexType name="examType">
  <sequence>
    <element name="name" type=”string”/>
    <element name="venue" type=”string”/>
  </sequence>
</complexType>
```

[4] Minus one for each major error (incorrect attribute, incorrect structure, missing elements, etc.)

ii. Write an XSLT template to convert the **exam** node into the following structure. [4]

```
<course xmlns=”http://ns2”>
   <code>CSC3003s</code>
   <place>Jameson</place>
</course>
```

Assume your template will be placed within the following stylesheet:

```
<xsl:stylesheet version="1.0"
   xmlns:xsl=http://www.w3.org/1999/XSL/Transform
   xmlns:source=”http://ns1”
   xmlns:target=”http://ns2”>
...
</xsl:stylesheet>
```

```
<xsl:template match="source:exam">
  <target:course>
    <target:code><xsl:value-of select="source: name"/></target:code>
    <target:place><xsl:value-of select="source: venue"/></target:place>
  </target:course>
</xsl:template>
```

[4] Minus one for each major error (incorrect XPaths, incorrect structure, missing elements, etc.)

b. In future XPath versions, there is a convergence with XQuery, which supports the FLWOR construct. The letter ‘F’ represents ‘For’, which iterates over a list of nodes. Explain briefly what each of the rest of the letters in FLWOR represent. [2]

Let binds variables to values

Where specifies conditions to be satisfied

OrderBy specified an expression to be used to order the results

Return specifies the output XML fragment for each match

[2] ½ mark for each answer

Question 6: Information Retrieval [10]

a. Briefly discuss the concepts of “recall” and “precision”? Sketch a typical recall vs. precision graph. [3]

recall is the proportion of relevant documents that are retrieved [1]

precision is the proportion of retrieved documents that are relevant [1]

```
  |
  | \
  |  \
R |   \
  |    \
  +---------- [1]
       P
```

b. Briefly discuss one technique to improve on the recall of an IR system. [1]

stemming where words can converted to a canonical form, thesauri where synonyms are added as search terms, ... [1]

c. Rank the following documents using the supplied similarity measure and assuming the query Q is “apples bananas”. Show all calculation. [4]

$$Similarity(D,Q) = \frac{1}{|D||Q|}\sum_{t=1}^{n} d_t . q_t \quad \text{where } |D| = \sqrt{\sum_{i=1}^{m} d_i^2}$$

Documents:

document D1: apples bananas

document D2: apples apples apples apples pears

Assume that $|Q| = \sqrt{\sum_{i=1}^{m} q_i^2} = \sqrt{1+1} = \sqrt{2}$

$|D1| = \sqrt{\sum_{i=1}^{m} d_i^2} = \sqrt{1+1} = \sqrt{2}$ *[1/2]*

$|D2| = \sqrt{\sum_{i=1}^{m} d_i^2} = \sqrt{4^2+1} = \sqrt{17}$ *[1/2]*

$$Similarity(D1,Q)=\frac{1}{|D||Q|}\sum_{t=1}^{n}d_t.q_t=\frac{1}{\sqrt{2}\sqrt{2}}(1.1+1.1+0.0)=1$$ *[1]*

$$Similarity(D2,Q)=\frac{1}{|D||Q|}\sum_{t=1}^{n}d_t.q_t=\frac{1}{\sqrt{2}\sqrt{17}}(4.1+0.1+1.0)=\sqrt{\frac{16}{34}}$$ *[1]*

Ranking: D1, D2 [1]

d. Create inverted files for the above document collection, including per-document weights for each term (do not use differential values). [2]

apples: D1:1 D2:4 [1]

bananas: D1:1 [1/2]

pears: D2 :1 [1/2]