University of Cape Town

Department of Computer Science

CSC3003s Class Test 2

2006

Marks **: 35**

Time **: 45 minutes**

Instructions:

- Answer all questions.
- Show all calculations where applicable.

Question 1: Information Retrieval [10]

Explain what each of the following words/phrases means in the context of IR systems [4]

term, document, relevance, ranked retrieval

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

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

Documents:

doc1: apples bananas

doc2: apples apples apples apples pears

Create inverted files for the above document collection, including per-document weights for each term. [2]

UCT CSC3003s 2006 :: IR + Databases + ToA :: Supplementary Test [35 marks]

Question 1: Information Retrieval [10]

What is the difference between filtering and ranking? [2]

In the ideal case, what is the speed complexity of filtering a single word query using inverted files? [1]

tf.idf is a common part of relevance formulae – What do high tf and idf values, respectively, indicate about a term? [2]

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

Discuss 2 techniques to improve on the recall of an IR system. [2]

UCT CSC3003s 2006 :: IM :: Exam [50 marks]

Question 1: XML / Information Retrieval [10]

Suppose that you are a consultant designing a news website.

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

You decide to build inverted files for filtering. What are inverted files? How can you improve on the efficiency of storing your inverted files? [2]

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

How would you ensure that your website is ranked highly in Google searches? [2]

You want to support non-English languages but your software only handles ASCII internally. How would you deal with this problem? [2]

Question 2: XML [10]

XSLT [4]

XML Schema [4]

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 what the rest of the letters in FLWOR represent. [2]

Question 3: Information Retrieval [10]

UCT CSC3003s 2006 :: IM :: Supplementary Exam [50 marks]

Question 1: XML / Information Retrieval [10]

Suppose that you are a consultant asked to design a system for archiving student records in XML format.

If XML an efficient internal representation for data? Why or why not? [1]

What efficiency benefits are there to using a Blob representation in an XML database? [1]

Your chosen database has an IR engine built-in but the documentation says only that it uses term frequency for relevance ranking. What is relevance ranking? What is term frequency? Discuss the other factor that is typically taken into account along with term frequency. [3]

How will you ensure that your XML records that are shared with other systems do not take on ambiguous interpretations when used with XML from other sources? [2]

Noting that you need to validate data fields with special formats, such as email addresses, what validation language will you use to define your data format? Provide a hypothetical example root element showing the link between data and formal specification – exact URIs are not required. [3]

Question 1: XML [10]

XSLT [4]

XML Schema [4]

A flat tree representation in an XML database is efficient for single node queries. What operations are not particularly efficient? [2]

Question 3: Information Retrieval [10]

In the ideal case, what is the speed complexity of filtering a single word query using inverted files? [1]

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

Discuss 2 techniques to improve on the recall of an IR system. [2]