UCT CSC303 2004 :: XML/IR :: Exam [25 marks]

Answer Questions 1 AND 2.

Then answer 3 questions from among Questions 3-6.

(Questions 3 and 4 are optional – Sonia will provide 2 more questions).

Question 2 : Information Retrieval [10]

1. When building an IR system to ignore all tags and index just the textual data in a collection of XML files, which parser family is more efficient: SAX or DOM? Explain why. [3]

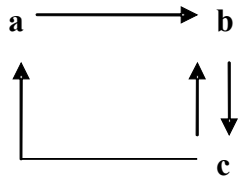
2. In preprocessing data, terms are often stemmed and stopped. What are stemming and stopping? How does stemming affect recall and precision? [3]

3. What is the difference between term frequency and document frequency? [2]

4. List 2 advantages of the Latent Semantic Analysis algorithm for IR. [2]

Question 3 : PageRank [10]

1. Apply the PageRank algorithm to the following link graph [6].



Start with equal ranks of 1/3 each. Stop after 4 iterations, or when c=5/12. Show all calculations.

(Hint: Remember that r[i](n) = Sum of r[j](n-1)/N[j] over all j in B[i])

2. Assuming that we had a collection of newspaper articles from the Cape Argus in digital form, why will the PageRank algorithm not be suitable for IR purposes? [2]

3. How does the HITS algorithm differ fundamentally from the PageRank algorithm? [2]

Question 4 : XML / XSLT [10]

1. Write an XSLT template to transform

```
<university>
<class>
<lecturer>EvilTeacher</lecturer>
<student>JoeStudent</student>
</class>
</university>
```

into:

```
<dungeon>
<torture>
<torturer>EvilTeacher</torturer>
<tortured>JoeStudent</tortured>
```

```
    </torture>
    </dungeon>
```

Assume that the values “Hussein” and “Alapan” may differ from one document to another. Assume the source namespace prefix is “source” and the destination prefix is “dest”. Assume that the number of <class> tags is unbounded in its defining XML Schema. Use the following as a starting point. [4]

```
<xslt:template match=”source:university‟>

. . .

</xslt:template>
```

2. Write an XML Schema type definition corresponding to the contents of the university node. [6]