# UCT CSC3003 2006 :: XML/IR :: Supp Solution [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 the other questions).

## Question 2 : Boolean Information Retrieval [10]

Consider the following collection of documents:

| DocumentID | Terms |
|---|---|
| Doc1 | he sells sea shells sea |
| Doc2 | he shells sea |
| Doc3 | sea shells sea |
| Doc4 | shells sells |

1. Build inverted files for this document collection. Include a term occurrence count with each DocumentID listed. [4]

| | |
|---|---|
| *he* | *Doc1:1 Doc2:1* |
| *sells* | *Doc1:1 Doc4:1* |
| *sea* | *Doc1:2 Doc2:1 Doc3:2* |
| *shells* | *Doc1:1 Doc2:1 Doc3:1 Doc4:1* |

2. If the query "sea shells" is submitted to a Boolean-AND-based search engine, which results will remain after filtering? [1]

*Doc1, Doc2, Doc3*

3. Using the following ranking formula, compute a ranking value for each result from the previous question. [3]

$$Similarity = \frac{1}{|D|} \sum_{t \in Q \cap D} (1 + \log_e f_{d,t}) . \log_e \left(1 + \frac{N}{f_t}\right)$$

Assume that:

- D is the length of the document, including all terms in the document.
- Only terms common to both query and document and considered.
- N = the total number of documents in the result set.
- $f_{d,t}$ = term frequency of term t in document d
- $f_t$ = number of documents term t appears in.

*Sim[Doc1,Q] = 1/√7 [(1+log 2)log(1+3/3) + (1+log 1)log(1+3/4)] =0.377*

*Sim[Doc2,Q] = 1/√3 [(1+log 1)log(1+3/3) + (1+log 1)log(1+3/4)] =0.577*

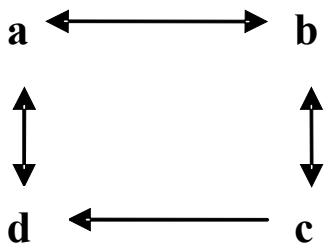
*Sim[Doc3,Q] = 1/√5 [(1+log 2)log(1+3/3) + (1+log 1)log(1+3/4)] =0.447*

4. Briefly discuss two techniques that can be used to improve on precision. [2]

*LSI [1] Stopping [1] ...*

**Question 3 : PageRank [10]**

1. Apply the PageRank algorithm to the following link graph [5].




Start with equal ranks of 1/4 each. Stop after 3 iterations, not including the initial values. Show all calculations, including N(umber of forward links) values and B(ack links) sets.

(Hint: Remember that $r[i]_n = \sum_{j \in B[i]} \frac{r[j]_{n-1}}{N[j]}$ )

| | N | B | R[0] | R[1] | R[2] | R[3] |
|---|---|---|---|---|---|---|
| A | 2 | B D | ¼ | 3/8 | 3/8 | 3/8 |
| B | 2 | A C | ¼ | ¼ | ¼ | ¼ |
| C | 2 | B | ¼ | 1/8 | 1/8 | 1/8 |
| D | 1 | A C | ¼ | ¼ | ¼ | ¼ |

*One mark for N, one mark for B, one mark for each correct iteration*

3. Why do we need to remove sinks and leaks from the link graph? [2]

*because the ranks will otherwise converge to those nodes as there are no outgoing links.*

4. Why is PageRank not applicable to a collection of letters? [1]

*there are no links*

5. Contrast the runtime performance of simple PageRank with simple HITS. [2]

*HITS calculates the focussed subgraph at runtime while PageRank calculates all ranks a priori – thus PageRank is much faster at runtime.*

**Question 4 : XML / XSLT [10]**

1. Write an XSLT template to transform

```
<basket>
   <number>123</number>
```

```
        <fruit>
           <name>apples</name>
           <type>granny smith</type>
        </fruit>
        <fruit>
           <name>grapes</name>
           <type>white </type>
        </fruit>
    </basket>
```

into:

```
    <basket>
        <number>123</number>
        <fruit>granny smith apples</fruit>
        <fruit>white grapes</fruit>
    </basket>
```

Assume that the values such as "granny smith" 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 <fruit> tags is unbounded in its defining XML Schema. Use the following as a starting point. [5]

```
<xslt:template match="source:basket">
. . .
</xslt:template>
```

```
<xslt:template match="source:basket">
  <dest:basket>
    <dest:number><xsl:value-of select="source:number"/></dest:number>
    <xslt:for-each select="source:fruit">
      <dest:fruit><xsl:value-of select="source:type"/> <xsl:value-of
select="source:name"/></dest:fruit>
    </xsl:for-each>
  </dest:basket>
</xslt:template>
```

*Minus one mark for each major error or half for minor or repeated errors.*

2. Write an XML Schema type definition corresponding to the contents of the basket node in the source document. [5]

```
<complexType>
  <sequence>
    <element name="number" type="integer"/>
    <element name="fruit" maxOccurs="unbounded">
      <complexType>
        <sequence>
          <element name="name" type="string"/>
          <element name="type" type="string"/>
        </sequence>
      </complexType>
    </element>
```

```
  </sequence>
</complexType>
```

*Minus one mark for each major error or half for minor or repeated errors.*