DIS Sample Exam (2003)

Question 1 [Overview]:

A. Explain why the search cost when using topological routing (CAN) in a d-dimensional space is $O(d\ n^{1/d})$ and not, for example, $O(n^{1/d})$.

B. What is the difference among the notions “multiple classification” and “ambigue classification” as used in the context of schemas for semi-structured databases ?

C. Is the set of predicates retained by the MinFrag algorithm for constructing the set of relevant predicates for the horizontal fragmentation of a relation monotonically increasing? Explain your answer.

Question 2 [Information Retrieval]:

A user obtained upon posing a query Q three documents D1, D2, D3 as part of the result. Assume the vocabulary consists of two terms T1 and T2 and the term frequencies of the three documents are given as

	D1	D2	D3
T1	1	x1	1
T2	3	x2	1

where $0 < x1, x2 \leq 3$.

The user now identifies to the query system D1 as being a relevant document. The query system generates a modified query Q’ by applying Roccio’s formula for query modification based on user relevance feedback, using parameters $\alpha=0$, $\beta=1$ and $\gamma=1$.

A.Compute the modified query Q’.

B. Can you find a pair of values for x1 and x2 such that the three documents will be ranked by the modified query Q’ as D1 > D3 > D2 ?

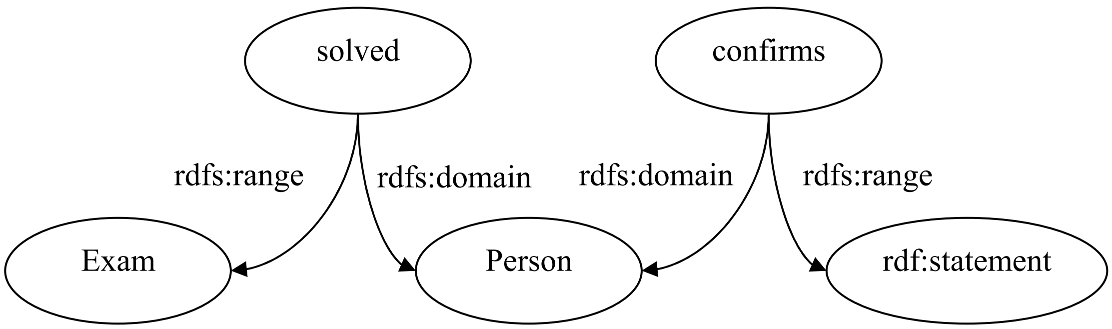
Hints:
The modified query vector could contain negative values and thus produce also negative similarity values.
To simplify computation you should assume that idf(T1)=idf(T2)=1 and need not to normalize similarity values, i.e. sim(Q, D)=QD.

Question 3 [XML and RDF]:

A. Express the following statement as RDF graph:

“X of type Person confirms the statement «Y of type Person solved each problem of Z, where Z is consisting of a sequence of four problems P1, P2, P3, and P4»”

You should use the RDF predicates and types given in the following RDF schema fragment:



B. Provide and XML encoding of the RDF graph.

Question 4 [Mobile Data Management]:

Given a broadcast schedule

A B C A D E A B F A D G

A. Of how many broadcast disks does the broadcast schedule consist and what are the frequencies of the broadcast disks?

B. Assume that data items are requested in the following order

D A G A G A

Given a cache that can store at most 2 data items, describe the cache state for the LIX strategy at each step when reading the data broadcast, till all requested data items have been obtained. In the running average computation

$$p_i := \frac{c}{t - t_i} + (1-c)\,p_i$$

of the access probability estimate use c=½ as weighting constant.

C. Does the PIX strategy need fewer steps?

Hints:
Assume that in case the PIX/LIX values of two data items are equal, the data item that is already in the cache remains in the cache.
Multiple requested data items can be accessed at the same time instant sequentially, if they are already present in the cache.