EXAMINATION FOR
“MULTIMEDIA INFORMATION EXTRACTION AND RETRIEVAL”
WINTER TERM 2010/2011,
MARCH 31, 2011
PROF. DR. RALF MÖLLER

You are not allowed to write down solutions before the examination is started. Once the examination is officially ended, you are not allowed to write down solutions either. Violations of these rules count as attempts to cheat and will lead you to fail the exam.

Name: ____

Student id: ____

Course: ____

Signature: ____

a) **Please put your student identification as well as a passport/official id card on the table. We need to check these.**

b) The exam will take **90 minutes**.

c) The exam is **closed book.**

d) The symbol “◷” will give you hints on the **recommended time for solving a task.**

e) We have more paper, should you need some, ask. Once you received additional sheets of paper, write down your name and student id.

1 General Multimedia Systems

13 🕒
10 P

a) Explain briefly the notions of *dictionary files* and *posting files*.

b) Describe briefly what is *metadata* and give two different motivations for attaching metadata to content.

c) What is the idea of *tokenization*? What are the major issues/problems?

2 Indices

8 🕒

8 P

a) Describe the idea of biword-indices. What are they used for? Can biword-indices be used to answer n-word phrase queries, e.g. „Hamburg University of Technology“? What is the notion of *false positives* in this case?

b) Given the following posting file:

- fools: doc2: (1,17,74,222); doc4: (8,78,108,458); doc7: (3,13,23,193);
- fear: doc2: (87,704,722,901); doc4: (13,43,113,433); doc7: (18,328,528);
- in: doc2: (3,37,76,444,851); doc4: (10,20,110,470,500); doc7: (5,15,25,195);
- rush: doc2: (2,66,194,321,702); doc4: (9,69,149,429,569); doc7: (4,14,404);

Which document(s), if any, match the following queries, where each expression within quotes is a phrase query?

- „fools rush in“
- „fools in fear“

c) Compute the inverse document frequency of the term „good“ with respect to the following three documents:

- doc1 = „Today is a good day.“
- doc2 = „Hello and good morning“
- doc3 = „Is this car any good?“

3 Similarity

15 🕒

15 P

a) How can one represent documents in a vector space? How big is that vector space potentially? What problems occur?

b) Name one method to reduce the number of axes in the document vector space.

c) Determine the similarity of each the following texts(=documents) with respect to the cosine similarity measure. It is sufficient if you write down the formula for each pair of documents. You don’t need to compute the actual value.

- „Hey you.“
- „Who are you?“
- „How are you doing?“

d) Why is the representation of documents as vectors particularly interesting? Think of the original intention to answer queries!

4 Media Analysis

10 🕒

14 P

Consider the following universe of documents: $D1, D2, ..., D10$. For a certain query, documents $D1, D2, D3, D4$ are relevant. However one information retrieval system returns $D3$, $D4$ and $D10$.

a) Calculate precision and recall for this example.

b) Compute the F_4-measure for that example. What does F_4 mean here?

c) Why is an information retrieval system offering 100% recall not useful without further information about the system? How about 100% precision?

17 🕒

15 P

5 Probabilistic Information Retrieval

a) What is the role of Bayesian Networks in Information Retrieval and how can they be used?

b) Model the following scenario as a Bayesian Network: Suppose that there are two events which could cause grass to be wet: either the sprinkler is on or it's raining. Also, suppose that the rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on).

c) Write down the probabilities (formulas are sufficient) for the following conditions:

- What is the probability that it is raining, given the grass is wet?

- What is the probability that it is grass is wet, given that it is raining?

d) What are the main problems with the probabilistic extension of Datalog?

6 Multidimensional Data Structures

10 ⏲ / 10 P

a) In the following, some information about the location of german cities is given. Draw a *point-quad-tree* with respect to the following city list (the tree should be dreated stepwise in the given order)

- Berlin (20,25)
- Hamburg (5,28)
- Munich (18,3)
- Stuttgart (4,4)
- Frankfurt (5,10)

7 Rules and abduction

17 🕒

17 P

a) Suppose you are given a family knowledge base in Datalog as follows:

$$
\begin{array}{l}
Person(X) : -Male(X). \\
Person(X) : -Female(X). \\
Animal(X) : -Dog(X). \\
Animal(X) : -Cat(X). \\
Male(homer). \\
Male(marge). \\
Male(bart). \\
Male(abe). \\
Dog(slh). \\
Cat(snowball). \\
has(lisa, snowball).
\end{array}
$$

Answer the following questions with respect to the family knowledge base.

- Write down a Datalog rule, which defines a *FemaleCatOwner* as “a female person, who has a Cat”

- Write down additional Datalog facts to express that
  - $mr.burns$ is a Person
  - $lisa$ is Smart
  - $homer$ works for $mr.burns$.
  - abe is the father of $homer$
  - $homer$ is the father of $bart$

- Create a rule to obtain all pairs “X is a grandfather of Y”

- Create a rule to obtain all *happy grandfathers*, that is, all grandfathers who have a smart grand child.

- Abduce the fact that “abe is a happy grandfather”. Describe in detail every step of the abduction process. Give at least two possible explanations for the desired fact, and outline which explanation you would prefer.