Multimedia Information Extraction and Retrieval
Summer Term 2012
Exercise Sheet 6

Ralf Möller, Karsten Martiny

Exercise Session:
Thursday, June 07, 2012, 8.00-8.45, SBS95-D1025

1. General: What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance is available)?

2. Bayesian Networks: Model the following scenario as a simple bayesian network:

   „I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?"

   Draw the network. Which events/variables are independent from each other?

3. For the previous alarm-exercise calculate the probability (the formula is enough, since we have no values vor the variables) for $P(J, M, A, \neg B, \neg E)$.

4. Datalog: We are given two directed graphs G_{black} and G_{white} over the same set V of vertices, represented as binary relations. Write a datalog program P that computes the set of pairs $\langle a, b\rangle$ of vertices such that there exists a path from a to b where black and white edges alternate, starting with a white edge.

5. Probabilistic Datalog: give ideas on how to model the following extensions to the probabilistic datalog example from the lecture (EDB: $term, link$, IDB: $about$):

   - Author information for documents.
   - Different types of documents based on class hierarchies (journal article, articles, poster, conference article, long journal articles).
   - Thesaurus of related words.

6. What are the main problems with probabilistic Datalog?

7. Binary Independence Retrieval: For a query q, the BIR model results in the following list of documents after the initialization step:

d_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
x_2	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0
relevance	R	R	R	R	N	R	R	R	R	N	N	R	R	R	N	N	N	R	N	N

The table further contains the binary vectors of the documents (only 2- dimensional: x_1 and x_2 for each of the 20 documents) and the relevance with respect to the query (R denotes relevant, N denotes not relevant). Given the relevance assessments, compute the new c_i-values as described in the script. Finally, sort the documents based on the new relevance ordering.