Multimedia Information Extraction and Retrieval SoSe 2010
Exercise Sheet 5

Prof. Dr. Ralf Möller, Sebastian Wandelt

02.06.2010

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)?

   Solution:

   A few differences between standard vector space tf-idf weighting model and the BIM probabilistic retrieval model on the first iteration are:

   - *tf-idf weighting is directly proportional to term frequency of the query term in the document whereas the BIM just takes into account the absence or presence of term in the document. Consider the query 'India Information Technology' on the document set:*
     Document1: India's Information technology sector is booming very fast relative to technology sectors of India.
     Document 2:The Information technology sector of India has grown drastically over the last few years.
     Now the tf-idf weighting will give more relevance to Document 1 whereas the BIM model puts them on the same relevance level.
   - *The idf part of the tf-idf weighting keeps the frequently occurring words (the stop words which cannot distinguish between documents) out of the relevance decision making part as for them idf is approximately is 0. On the other hand, the BIM treats all terms alike and every word has an equal say in deciding the relevance.*

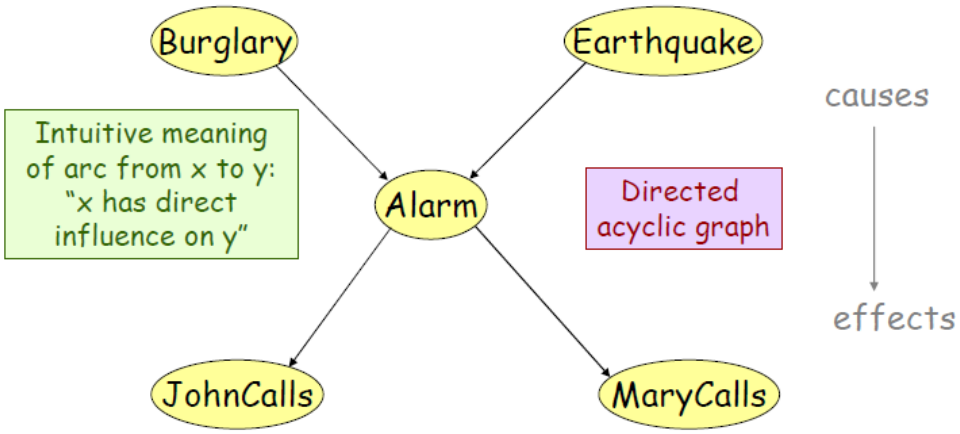
- *While calculating the relevance of a document, the tf-idf says that the words occurring in the query but not present in the document have a zero contribution in the relevance value of the document whereas the BIM counts their contribution by the fraction of such other documents in the collection (which are relevant but do not contain this term).*

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?

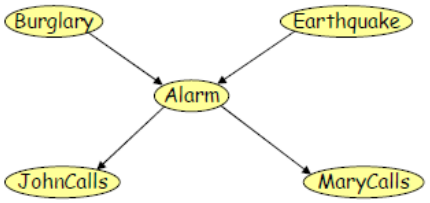
   Solution:



   The beliefs JohnCalls and MaryCalls are independent given Alarm or $\neg Alarm$

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

Solution:



- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
  $= P(J \wedge M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  $= P(J|A, \neg B, \neg E) \times P(M|A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  (J and M are independent given A)
- $P(J|A, \neg B, \neg E) = P(J|A)$
  (J and $\neg B \wedge \neg E$ are independent given A)
- $P(M|A, \neg B, \neg E) = P(M|A)$
- $P(A \wedge \neg B \wedge \neg E) = P(A|\neg B, \neg E) \times P(\neg B|\neg E) \times P(\neg E)$
  $= P(A|\neg B, \neg E) \times P(\neg B) \times P(\neg E)$
  ($\neg B$ and $\neg E$ are independent)
- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J|A)P(M|A)P(A|\neg B, \neg E)P(\neg B)P(\neg E)$

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.

   Solution:

   - $p(X, Y) :- white(X, Y)$
   - $p(X, Z) :- white(X, Y), p2(Y, Z)$
   - $p2(X, Y) :- black(X, Y)$
   - $p2(X, Z) :- black(X, Y), p(Y, Z)$

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. **Solution:**

  Just add a author-predicate for each document.

- Different types of documents based on class hierarchies (journal article, articles, poster, conference article, long journal articles). **Solution:**

  $PX\ isA(journalarticle, article).$
  $PX\ instanceOf(D, Class) : -instanceOf(D, SubClass), isA(SubClass, Class).$

- Thesaurus of related words. **Solution:**

  Hard to model. In a naive way, all words will be related in the end by transitive closure.

6. What are the main problems with probabilistic Datalog?

   Solution:

   All statements are either independent or disjoint. No intervals. Reasoning is hard.

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.

   Solution:

The evaluation of the c_i-values is depicted in the following table. It is: $l=12$ (number of relevant documents) and $k=20$ (number of presented documents). To obtain the c_i-values, the following equation was used:

$$c_i = ln(r_i(1-n_i))-\backslash ln(n_i(1-r_i))$$

i	l_i	k_i	r_i	n_i	c_i
1	8	11	2/3	3/8	1.20
2	7	11	7/12	1/2	0.34

This leads to an ordering of the documents based on their binary vector representation: (1,1) > (1,0) > (0,1) > (0,0). As one solution, the BIR model ranks documents just in order of their document ID, i.e., $d_1 > d_2 > \ldots > d_{20}$.