Lab session introduction

Uncertain data management

Antoine Amarilli and Silviu Maniu

December 19th, 2017

This document describes the beginning of the lab session, with exercises that must be performed by hand. We will implement them on the computer later.

We consider a *truth finding* application, where we have extracted facts from several sources (i.e., websites). A first table, Trust, indicates which sources are trustworthy. Another table, Claims, indicates which sources support which fact (each fact is given as a numeric identifier). Both tables are uncertain: we do not know which sources are correct, and we are not sure of whether a source supports a statement (i.e., errors may have been made when extracting facts from sources).

We represent the uncertainty on these two relations using the TID model. We consider the following instance:

Trust

source	proba
Legifrance	0.95
Wikipedia	0.8
Doctissimo	0.4
Gorafi	0.2
Onion	0.1

Claims

source	claim	proba
Legifrance	42	0.7
Wikipedia	42	0.9
Gorafi	42	0.6
Wikipedia	51	0.7
Doctissimo	51	0.4
Wikipedia	66	0.3
Gorafi	66	0.9

1 Query evaluation by hand

Question 1. We wish to determine the answer to the following query: "Which sources in our dataset are trustworthy and make at least one claim?"

Write the query in the relational algebra using the select, rename, product and project operators.

Write the query again in the relational algebra using the project and join operators, in two equivalent ways (each way should use exactly one project and one join).

Question 2. Consider the deterministic instance obtained from the given TID instance by removing the proba column. What is the result of evaluating the previous query on this deterministic instance?

Question 3. Compute the probability according to the given TID instance that the source Gorafi makes at least one claim. Deduce the probability that Gorafi is a trustworthy source that makes at least one claim.

Hint: Use a calculator for numerical computations.

Question 4. Generalizing this process, compute the answer to the query of Question 1 with the correct probabilities according to the TID model. The result should be a probabilistic annotation of the result of Question 2. Do this in two steps: first, compute for each source the probability that it makes a claim, and then combine this with the Trust table to obtain the final answer.

Question 5. Consider the two relational algebra queries of Question 1 written with the join and project operators. Extending the relational algebra operators to manipulate probabilities as seen in class, which query would yield the output table of Question 4 with the correct probabilities? Explain why.

Would the other query give the correct output tuples (up to the probabilities)? How would the probabilities compare to the correct output? Explain why.