Synthetic Dataset Generator for Truth Discovery Scenarios

Documentation

Mouhamadou Lamine Ba and Laure Berti-Equille

Hamad Bin Khalifa University

Creation date : November 5, 2015
Revision date : December 7, 2015

Contents

1 Introduction

In order to provide to users the ability to test and compare multiple truth finding algorithms on various dataset settings, some of them being not easy to obtain in practice, we provide the DAFNA API together with a synthetic dataset generator. The synthetic dataset generator allows to simulate a large variety of scenarios where sources present different behaviors in terms of coverage, error rate, reliability level, conflicting information, ect. More importantly, the generator guarantees a ground truth for the evaluation of the accuracy of compared truth finding algorithms.

2 JAVA Set Up

The synthetic dataset generator has been implemented under JAVA version 7. The corresponding jar file is made available for free download on `http://daqcri.github.io/dafna/#/dafna/exp_sections/realworldDS/synthetic/syntheticDs.html` with the name **DAFNA-DataSetGenerator.jar**. Kindly, verify you have installed JAVA JRE 1.7 or a higher version in your computer before using the jar file.

3 Types of Datasets

The synthetic dataset generator can produce various types of datasets in function of the source coverage, ground truth distribution, distinct values distribution, and level of similarity between distinct values of the same data item. It also offers the ability to capture both extreme optimistic and pessimistic scenarios by playing on ground truth distribution controller.

4 Generator Usage

The synthetic dataset generator is used with the following command line, e.g. on a UNIX terminal.

```
# java -jar DAFNA-DataSetGenerator.jar param_1 param_2 ...param_n
```

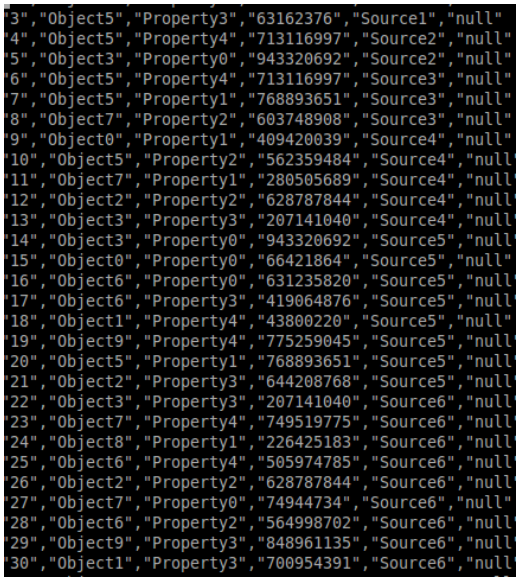
param_1, param_2, . . . , param_1 correspond to the different input parameters of the generator.

4.1 Generation Example

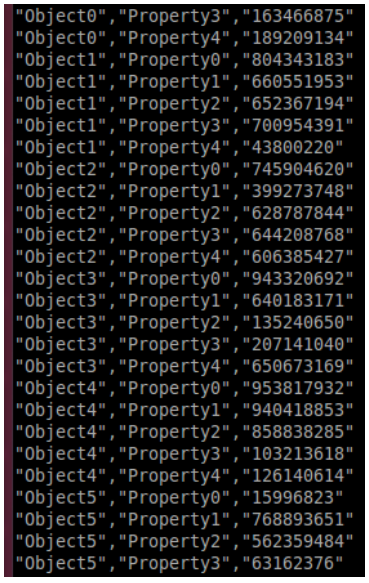
The use of the synthetic dataset generator requires to instanciate all the parameters in a certain order, as we will see in the next. As an example of a generation, suppose we would like to obtain a dataset consisting of 10 sources which fully cover 10 objects having 5 properties. We also want to constrain the distribution of the distinct values per data item and the ground truth distribution to be both exponentials. We do not care about similar distinct values. Finally, we would like to set the default directory for the output as “./Test". We obtain the needed configuration by running the synthetic dataset generator as follows.

```
   # java -jar DAFNA-DataSetGenerator.jar -src 10 -obj 10 -prop 5 -cov
1.00 -ctrlC Exp -ctrlT Exp -v 3 -ctrlV Exp -s dissSim -f "./Test"
```

Figures (a) and (b) respectively show an excerpt of the dataset file and the corresponding ground truth (GT) dataset produced by the generator after completion.



(a) Excerpt of the dataset file



(b) Excerpt of the GT dataset

4.2 Input Parameters

The synthetic dataset generator considers as input 10 parameters. All parameters are mandaratory for a successful generation. Each parameter controls a given

characteristic of the dataset to be generated. Below the details of each parameter.

4.2.1 Number of sources

-src specifies the number of sources providing claims. Its value must be a non null integer.

4.2.2 Number of objects

-obj specifies the number of objects covered by the sources. Its value must be a non null integer. Note that it should correspond to the higher bound of the amount of objects covered by the sources.

4.2.3 Number of properties

-prop specifies the number of properties describing each object. Its value must a be a non null integer.

4.2.4 Data item source coverage

-cov corresponds to a value from 0 to 1 representing the uniform percentage of data items covered by sources in the Uniform coverage Distribution.

4.2.5 Source coverage distribution

-ctrlC sets the source coverage distribution which is either **uniform** or **Exp** (i.e. Random). An uniform distribution means that the number of values provided by the sources is uniformly distributed given the source coverage value. In the same spirit, the number of values provided by the sources is exponentially distributed when its coverage distribution is exponential.

In the case of an exponential distribution, the coverage of a given source i is defined as follows.

$$\mathrm{Cov}(i) = 1 + (|D| - 1) \times \frac{\mathrm{e}^{\frac{4 \times i}{(|S|-1)}} - 1}{\mathrm{e}^4 - 1} \; \forall i = 0, \ldots, (|S| - 1)$$

where S and D are respectively the set of source and the set of data items.

4.2.6 Ground truth distribution per source

-ctrlT specifies the Ground Truth Distribution per Source. It can be set to one of the following distribution:

R: Random Distribution in which the number of true positive claims per source is randomly chosen.

Uniform: Uniform Distribution where each source provides the same number of true positive claims.

FP: Fully Pessimistic Distribution which considers that 80% of the sources provide always false claims while 20% of the sources provide always true positive claims.

FO: Fully Optimistic Distribution which considers that 80% of the sources provide always true claims while 20% of the sources provide always false positive claims.

80P: 80% Pessimistic Distribution in which 80% of the sources provide 20% true positive claims whereas 20% of the sources provide 80% false positive claims.

80O: 80% Optimistic Distribution which considers that 80% of the sources provide 80% true positive claims and only 20% of the sources provide 20% true positive claims.

Exp Exponential Distribution in which the number of true positive values provided by the sources is exponentially distributed. Such an exponential distribution ground truth is defined and estimated for a given source i as follows.

$$\mathrm{GT}(i) = |D_{s_i}| \times \frac{\mathrm{e}^{\frac{i}{|S|}} - \mathrm{e}^{\frac{1}{|S|}}}{\mathrm{e} - \mathrm{e}^{\frac{i}{|S|}}} \; \forall i = 1, \ldots, |S|$$

in where S and D_{s_i} are respectively the set of sources and the set of data items covered by the source i.

4.2.7 Number of distinct values per data item

-v sets the number of distinct values per data item. The specified value will be used as a constant for the uniform model, and as the maximum number of distinct values in the exponential model.

4.2.8 Distinct value distribution per data item

-ctrlV specifies the distinct values distribution (or conflict distribution) per data item. It can be set either uniform or exponential.

Uniform: All data items have the same number of distinct values claimed by the set of sources.

Exp: Each data item has a number of distinct values that is exponentially distributed. In such a case, many data items have very few conflicts, while few data items have lots of conflits. An exponential distinct value

distribution for data item j is defined as follows.

$$\mathrm{NbDistinctV}(j) = (\mathsf{maxNbDistinctV} - 1) \times \frac{e^{\frac{2 \times j}{|D|} - 1}}{e^{\frac{2 \times (|D|-1)}{|D|} - 1}} + 1$$

where D is the set of data items.

4.2.9 Value similarity

-s: Similarity level between the different values of the same data item. It must be set either to **sim** or to **dissSim**.

Sim: Distinct values for data items will be highly similar.

dissSim: Distinct values for data items will be highly diss-similar.

4.2.10 Output folder

-f specifies the output folder where the dataset will be created and saved.

Important: When running the dataset generator, **all the parameters above need to be valued and listed in the required ordering**. Otherwise, the generator will not work properly.

4.3 Output Datasets

Given the required input parameters, the synthetic dataset generator proceeds to the generation of the dataset with the desired characteristics. After completion, it creates the two following folders into the main directotry specified by the user with the parameter **-f**.

- **claim/** Folder which contains the dataset file
- **truth/** Folder which contains the ground truth file

The dataset file produced by the generator contains the sources together with the set of claims. It corresponds to a csv file in which each line contains the following fields:

```
ClaimId, Object, Property, Claim, Source, TimeStamp
```

where claimId, Object, Property, Claim, Source, and TimeStamp are respectively the identifer of the claim, the real world object, one of its properties, a claim about this property, the source who made the claim, and the time when the claim has been made (Cf. Figure a) above).

The ground truth dataset contains the ground truth associated to the generated dataset. It is also a csv file in which each line contains the following fields:

`Object, property, Trueclaim` where Object, property, and TrueClaim represent respectively an object, one among its set of properties, and the true claim about this given property (Cf. Figure b) above).

5 Questions or Feebacks

For any questions or feebacks about the generator, please feel free to contact us or visit the website of the DAFNA project for more details by following the link: http://dafna.qcri.org.