User's Manual

Pre-requisites

A Java Runtime Environment must be installed and properly set up to use our program.

Usage

Our benchmark generator is a command-line utility with the usage depicted here:

```
java -jar generator.jar -load-pre-defined-settings QueriesDir=String
ChanceOfSortingQueriesInFiles=double SizeGen=RandomVariateGenInt
ProbabilityOfPureRecursion=double PercentageOfFilesWithDTD=double
PercentageOfExactQueriedElements=double
PercentageOfTextGen=RandomVariateGenInt TextFile=String
AverageMixedContentLevelGen=RandomVariateGenInt
NumOfAttributesGen=RandomVariateGenInt
ChanceOfAggregateQueryInFiles=double
DesiredExactMatchLevels=RandomVariateGenInt
NumberOfElementsGen=RandomVariateGenInt
ChanceOfRelativeOrderedQueryInFiles=double
ProbOfRelationalPattern=double RecursiveQueryPerFile=double
InterestingElement=String QuantifyQueryPerFile=double OutputDir=String
IntermediateResultQueryPerFile=double ChanceOfTextQueriesInFiles=double
NaturalTextFile=String ProbabilityOfPurelyRecursiveElement=double
NumberOfGeneratedFiles=int ProbabilityOfLinearRecursion=double
FanoutGen=RandomVariateGenInt ChanceOfAbsoluteOrderedQueryInFiles=double
AbsoluteOrderPercentage=double
SimpleMixedContentElementsPercentage=double
PercentageOfMixedElementsGen=RandomVariateGenInt
```

All parameters have a form of `ParameterName=type`. The user is responsible for filling the `type` with the value of respective parameter. Different types of possible values are explained in the next section. Every parameter is described in the section Parameters.

At the place of `-load-pre-defined-settings` can be pre-defined settings string. See Pre-defined Settings for details.

Types of Parameters

There are four types of parameters. Following catalogue lists them all:

- `String`: A string value is expected. It is usually a path to a file or a directory.
- `Int`: Integer value is expected.
- `Double`: A floating point number is expected. Actual type is Java double. A typical example would be 3.14.

- `RandomVariateGenInt`: A discrete probability distribution is expected. This is a list of possible distributions:
    - `uniform(int, int)`: Represents a uniform distribution. The first parameter is `a`, the second parameter is the `b`. For more details, see the thesis.
    - `binomial(double, int)`: Represents a binomial distribution. The first parameter is the `p`, the second one is the `n`. For more details, see the thesis.

Parameters

List of possible parameter of the benchmark generator:

- `QueriesDir`: Directory where generated queries will emerge.
- `ChanceOfSortingQueriesInFiles`: Probability for each file, that it will be queried by a sorting query.
- `SizeGen`: Distribution of size in bytes of XML documents.
- `ProbabilityOfPureRecursion`: Probability that an element will be a part of a pure recursion.
- `PercentageOfFilesWithDTD`: Percentage of XML documents that will have a generated DTD schema.
- `PercentageOfExactQueriedElements`: Percentage of all generated elements in all files that will be queried by an exact match query.
- `PercentageOfTextGen`: Distribution of percentage of text in the
- `TextFile`: File containing words for markup of XML document.
- `AverageMixedContentLevelGen`: Distribution of an average level where the mixed content should appear.
- `NumOfAttributesGen`: Distribution of the number of the attributes per element.
- `ChanceOfAggregateQueryInFiles`: Probability of aggregate query per XML document.
- `DesiredExactMatchLevels`: Distribution of levels where exact match queries will apply.
- `NumberOfElementsGen`: Distribution of the number of elements in XML files.
- `ChanceOfRelativeOrderedQueryInFiles`: Probability of relative order query per XML document.
- `ProbOfRelationalPattern`: Probability of occurrence of a relational pattern.
- `RecursiveQueryPerFile`: Probability of a recursive query per XML document.
- `InterestingElement`: Type of element of your interest. Can be one of the:
    - `common`: most common element will be queried for each queried XML document
    - `mixed-content`: most mixed-content element will be queried for each queried XML document
    - `recursive`: element that occurred most times in recursion will be queried for each queried XML document
    - `leaf`: element that occurred most times as a leaf in XML tree will be queried for each queried XML document
- `QuantifyQueryPerFile`: Percentage of files that will be queried by quantification query.
- `OutputDir`: Outpu directory for generated XML documents and their schemes.
- `IntermediateResultQueryPerFile`: Percentage of files queried by intermediate result queries.

- `ChanceOfTextQueriesInFiles`: Percentage of files queried by text search queries.
- `NaturalTextFile`: A file containing natural text. This text will be used when creating text content.
- `ProbabilityOfPurelyRecursiveElement`: Probability that an element in a pure recursion will be the recursive one.
- `NumberOfGeneratedFiles`: Number of generated XML documents.
- `ProbabilityOfLinearRecursion`: Probability of a linearly recursive element.
- `FanoutGen`: Distribution of fan-out per XML document.
- `ChanceOfAbsoluteOrderedQueryInFiles`: Probability of an absolute order query per XML document.
- `AbsoluteOrderPercentage`: Percentage of absolute ordered queries' index. For instance a 100 would mean to ask for the last element in every XML document.
- `SimpleMixedContentElementsPercentage`: Percentage of mixed-content in simple elements.
- `PercentageOfMixedElementsGen`: Distribution of mixed-content elements through files.

Pre-defined Settings

There are six possible pre-defined sets of parameters that can be loaded:

- `-data-centric` is equivalent to parameters:

```
NumberOfGeneratedFiles=89 PercentageOfTextGen=binomial(438,0.1)
NumberOfElementsGen=binomial(604,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(1,3) TextFile=text.txt
PercentageOfMixedElementsGen=binomial(2,0.1)
PercentageOfFilesWithDTD=0 OutputDir=./generated/dat
ProbOfRelationalPattern=0.15
SimpleMixedContentElementsPercentage=0.559
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0.0003
ProbabilityOfPureRecursion=0.0003
QueriesDir=./generated/dat/queries
PercentageOfExactQueriedElements=0.01
ChanceOfAggregateQueryInFiles=0.4
ChanceOfSortingQueriesInFiles=0.03
ChanceOfTextQueriesInFiles=0.5
ChanceOfRelativeOrderedQueryInFiles=0.2
ChanceOfAbsoluteOrderedQueryInFiles=0.3
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.08 RecursiveQueryPerFile=0.075
IntermediateResultQueryPerFile=0.09
ProbabilityOfLinearRecursion=0.0006 NaturalTextFile=text.txt
```

- `-document-centric` is equivalent to parameters:

```
NumberOfGeneratedFiles=305
PercentageOfTextGen=binomial(816,0.1)
NumberOfElementsGen=binomial(6162,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(3,5) TextFile=text.txt
PercentageOfMixedElementsGen=binomial(765,0.1)
PercentageOfFilesWithDTD=0 OutputDir=./generated/doc
ProbOfRelationalPattern=0.03
SimpleMixedContentElementsPercentage=0.794
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0.1876
ProbabilityOfPureRecursion=0.1876
QueriesDir=./generated/doc/queries
PercentageOfExactQueriedElements=0.01
ChanceOfAggregateQueryInFiles=0.4
ChanceOfSortingQueriesInFiles=0.03
ChanceOfTextQueriesInFiles=0.5
ChanceOfRelativeOrderedQueryInFiles=0.2
ChanceOfAbsoluteOrderedQueryInFiles=0.3
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.08 RecursiveQueryPerFile=0.075
IntermediateResultQueryPerFile=0.09
ProbabilityOfLinearRecursion=0.1992 NaturalTextFile=text.txt
```

- `-exchange` is equivalent to parameters:

```
NumberOfGeneratedFiles=9 PercentageOfTextGen=binomial(363,0.1)
NumberOfElementsGen=binomial(26772,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(3,5) TextFile=text.txt
PercentageOfMixedElementsGen=binomial(87,0.1)
PercentageOfFilesWithDTD=0 OutputDir=./generated/ex
ProbOfRelationalPattern=0.15
SimpleMixedContentElementsPercentage=0.996
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0.2248
ProbabilityOfPureRecursion=0.2248
QueriesDir=./generated/ex/queries
PercentageOfExactQueriedElements=0.01
ChanceOfAggregateQueryInFiles=0.4
ChanceOfSortingQueriesInFiles=0.03
ChanceOfTextQueriesInFiles=0.5
ChanceOfRelativeOrderedQueryInFiles=0.2
ChanceOfAbsoluteOrderedQueryInFiles=0.3
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.08 RecursiveQueryPerFile=0.075
IntermediateResultQueryPerFile=0.09
ProbabilityOfLinearRecursion=0.3257 NaturalTextFile=text.txt
```

- `-reports` is equivalent to parameters:

```
NumberOfGeneratedFiles=1491
PercentageOfTextGen=binomial(62,0.1)
NumberOfElementsGen=binomial(464314,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(3,5) TextFile=text.txt
PercentageOfMixedElementsGen=uniform(0,0)
PercentageOfFilesWithDTD=0 OutputDir=./generated/rep
ProbOfRelationalPattern=0.48
SimpleMixedContentElementsPercentage=0
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0
ProbabilityOfPureRecursion=0 QueriesDir=./generated/rep/queries
PercentageOfExactQueriedElements=0.0001
ChanceOfAggregateQueryInFiles=0.004
ChanceOfSortingQueriesInFiles=0.003
ChanceOfTextQueriesInFiles=0.005
ChanceOfRelativeOrderedQueryInFiles=0.002
ChanceOfAbsoluteOrderedQueryInFiles=0.003
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.0008 RecursiveQueryPerFile=0.00075
IntermediateResultQueryPerFile=0.0009
ProbabilityOfLinearRecursion=0 NaturalTextFile=text.txt
```

- `-research` is equivalent to parameters:

```
NumberOfGeneratedFiles=153
PercentageOfTextGen=binomial(331,0.1)
NumberOfElementsGen=binomial(170,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(3,5) TextFile=text.txt
PercentageOfMixedElementsGen=binomial(101,0.1)
PercentageOfFilesWithDTD=0 OutputDir=./generated/res
ProbOfRelationalPattern=0.11
SimpleMixedContentElementsPercentage=0.02
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0
ProbabilityOfPureRecursion=0 QueriesDir=./generated/res/queries
PercentageOfExactQueriedElements=0.01
ChanceOfAggregateQueryInFiles=0.4
ChanceOfSortingQueriesInFiles=0.03
ChanceOfTextQueriesInFiles=0.5
ChanceOfRelativeOrderedQueryInFiles=0.2
ChanceOfAbsoluteOrderedQueryInFiles=0.3
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.08 RecursiveQueryPerFile=0.075
IntermediateResultQueryPerFile=0.09
ProbabilityOfLinearRecursion=0.0065 NaturalTextFile=text.txt
```

- `-semantic-web` is equivalent to parameters:

```
NumberOfGeneratedFiles=26 PercentageOfTextGen=binomial(549,0.1)
NumberOfElementsGen=binomial(30308,0.5) FanoutGen=uniform(3,5)
AverageMixedContentLevelGen=uniform(3,5) TextFile=text.txt
PercentageOfMixedElementsGen=binomial(24,0.1)
PercentageOfFilesWithDTD=0 OutputDir=./generated/sem
ProbOfRelationalPattern=0.20
SimpleMixedContentElementsPercentage=0.03
NumOfAttributesGen=uniform(0,4)
ProbabilityOfPurelyRecursiveElement=0.0146
ProbabilityOfPureRecursion=0.0146
QueriesDir=./generated/sem/queries
PercentageOfExactQueriedElements=0.01
ChanceOfAggregateQueryInFiles=0.4
ChanceOfSortingQueriesInFiles=0.03
ChanceOfTextQueriesInFiles=0.5
ChanceOfRelativeOrderedQueryInFiles=0.2
ChanceOfAbsoluteOrderedQueryInFiles=0.3
AbsoluteOrderPercentage=0.5
DesiredExactMatchLevels=uniform(4,5) InterestingElement=common
QuantifyQueryPerFile=0.08 RecursiveQueryPerFile=0.075
IntermediateResultQueryPerFile=0.09
ProbabilityOfLinearRecursion=0.0252 NaturalTextFile=text.txt
```

Example

The example usage would be:

```
java -jar generator.jar -data-centric
```

This will generate data-centric XML documents in the `./generated/dat` folder and queries in the `./generated/dat/queries` folder.