Gestão e Tratamento de Informação

3rd Project

Deadline: 21 Dec. 2014 Online submission at IST/Fénix

Exercise 1

Consider an XML representation for a list of politicians, as shown in the next figure.

```
<politicians>
 <politician name="José Sócrates de Sousa" party="PS" />
 <politician name="Pedro Passos Coelho" party="PSD" />
 <politician name="Paulo Portas" party="CDS" />
 <politician name="António Costa" party="PS" />
 <politician name="António José Seguro" party="PS" />
 <politician name="Passos Coelho" party="PSD" />
 <politician name="Aníbal Cavaco Silva" party="PPD/PSD" />
 <politician name="Cavaco Silva" party="PSD" />
 <politician name="José Seguro" party="PS" />
 <politician name="Paulo Portas" party="CDS/PP" />
 <politician name="José Socrates de Sousa" party="PS" />
 <politician name="Pedro P. Coelho" party="PSD" />
 <politician name="Mário Soares" party="PS" />
 <politician name="Aníbal António Cavaco Silva" party="PPD/PSD" />
 <politician name="Catarina Portas" party="BE" />
 <!-- list of remaining politicians -->
</politicians>
```

1.1. Write an XQuery function for detecting pairs of politicians whose names are highly similar, and are thus likely to be duplicates.

In your function, to estimate the similarity between pairs of politicians, you should use the Jaccard similarity coefficient between character bi-grams extracted from the names of the politicians. Pairs of names with a similarity score above 0.5 should be considered as duplicates.

1.2. Write an XQuery function that, using as input the results of the function developed in 1.1, computes the clusters of politician records that are likely to be duplicates between themselves. For computing the clusters, you should implement the transitive-closure approach.

1.3. Write an XQuery function that, using as input the original list of politicians and the results from the function developed in 1.2, returns a "cleaned" dataset where clusters of duplicate politicians are consolidated into a single record. When consolidating multiple duplicate records, you should use the largest politician name, and the shortest party abbreviation.

Exercise 2

Consider the following three tables:

```
speech (name, date, topic)
politician (name, party)
review (name, date, topic, score)
```

The relation `speech` stores data about speeches of politicians, the relation `politician` associates the politician name to a party, and the relation `review` stores the score given to each speech (that ranges from 1 to 5).

2.1. Express in Datalog the following queries using only the three base relations:

`GreatPoliticians(p)` returns the names of politicians who made speeches that receive a score greater than 4

`GreatParties(p,pt)` returns the names of the politicians, and corresponding parties, who made a speech with a score greater than 4

2.2. Consider the following queries:

```
Q1(n,t) :- GreatPoliticians(n), review(n,d,t,s), t = 'Education'
Q2(p) :- GreatPoliticians(p), review(p,d,t,s), t = 'Education'
```

Indicate which containment and equivalence relationships occur. Justify.

2.3. Consider the schema mapping language *Global-As-View* (GAV) and the source relations `speech`, `politician` and `review`. Consider also the two relations of the mediated schema `GreatPoliticians` and `GreatParties` as defined in 2.1.

Consider the query:

```
Q3('Joaquim Silva,pt) :- GreatPoliticians(p),
                          GreatParties(p,pt), p='Joaquim Silva'
```

Show how can this query be answered using the source relations.

Exercise 3

Consider the following two strings:

HEPOCARTES
EPOCRATES

3.1. Compute the similarity between the two strings using the Jaro measure.

3.2. Compute the similarity between the same two strings using the Jaccard measure with basis on 2-grams.

3.3. Compare the two similarity values obtained and comment on the similarity between the two strings and the characteristics of the two similarity measures.

Exercise 4

Consider the query discovery algorithm used by the CLIO system to compute the schema mappings (data transformations to be applied to source data in order to produce target data) from a given set of correspondences or schema matches.

Suppose the following two schemas:

- **Source schema:**

```
Politician (nameP, party)

Journalist (nameJ, newspaper)

Article (nameJ, date, title, opinion)
     nameJ: FK(Journalist)

Speech (nameP, date, topic, opinion)
     nameP: FK(Politician)
```

- **Target schema:**

```
Speaker(name, party, opinion)
```

Consider also the following correspondences between them:

```
C1: Politician.nameP ≈ Speaker.name
C2: Journalist.nameJ ≈ Speaker.name
C3: Article.opinion ≈ Speaker.opinion
C4: Speech.opinion ≈ Speaker.opinion
```

Apply the algorithm to find the mappings that will be shown to the user. Show the intermediate and final results.

Exercise 5

Explain briefly, using **your own words** (you can use an example if it helps):

5.1. Why is it challenging to find the transformations (schema mappings) that must be applied to a given data source in order to produce data to be inserted in a target schema, once the correspondences between the two schemata (schema matching) are given?

5.2. What is the purpose of data quality dimensions?

(Note that we do not want a transcription of what is written in the slides or book)

Submitting the project

The solutions to the project should be submitted through the *Fénix* system, in the form of a .zip file containing a PDF report with the solutions for the exercises, as well as individual text files with the solutions for each of the exercises (i.e., documents with the XML, XSD, XSLT or XPath/XQuery code).

In the course Webpage, you can find a Microsoft word template for the project report.

In the theoretical class following the electronic submission, a printed copy of the report, with the solutions for each exercise, should also be delivered.

We will not accept deliveries through e-mail, with reports not conforming to the supplied template, or without the text files with the solutions for the exercises.

Good luck!