wuh95_liy23_jiah71_Assignment10

Task 1. Develop a project idea

1 . Have a look at the given data set *Literature-data_TU-Darmstadt.txt*

The data set “Literature-data_TU-Darmstadt.txt” is a database of 1071 rows where an article can be preliminarily viewed as the entity because all other information such as DOI, Quelle, Autor, Titel, Jahr, Journal, Type, etc, can be viewed as organized around it, namely the attributes of the articles.

Some cells are empty, even DOI, which was soon to be observed as clearly the primary key for the entity “article”.

Quelle, or source, is where an article can be sought. Most of the sources are WOS, web of science.

Author, title are clearly directed towards articles. Where column “author” can and usually do contain many items.

Year and journal are when and where the article is published.

Type indicates what type the article belongs to, book chapter, interview, or something else.

DOI is for uniquely identifying an article, thus the best choice for the primary key as we proceed to transform this dataset into an ER gram.

“Gelesen” has the value of either Görge or NULL. Which seems a not generally valuable attribute.

Empirisch contains (quasi) Boolean values, likely to categorize if an empirical method of analysis had been used in the writing of the article.

Ausschlusspunkt marks a certain part of the article, such as abstract, volltext, etc.

Ausschlussgrund points out some key traits of the article.

Kommentare, or comments, are some short comments about the article.

Feld, or field, indicates which field the article is mainly categorized into.

Thema, or theme, is more specific about the discipline of the branch or applicative prospects.

HMD, abbr for head-mounted device, contains the model of such devices as values. Together with “interaction” and “participants” they seem to describe the experiments, if any, that were led in the writing of the article.

Then the columns behind are all of the values either NULL or cross, forming a certain rating system for articles or some aspects of the articles.

2. Develop a project idea by combining another data set with the given one and visualize the data appropriately.

We think the description, categorization, and rating of the articles are considerably complete, and even if not, we are not likely to be able to dig deeply into this very specialized field of VR and provide further professional perspectives. So we decided to merge this dataset with some information about the background of the researchers or their affiliations, in order to look into the social-environmental factors that influenced the research productions in this field.

One of the main dimensions of the social environment is the country. Which country has produced the most research papers among the set of articles we have? Does it have anything to do with its investments in the R&D (research and development) yearly? How much is the R&D investment and what percentage has it accounted for in the GDP of these countries? That's what our main project idea is based on.

To find the affiliation information we looked into a more comprehensive database of scientific papers through an API (**Crossref REST API**) we found on Github. And we wrote a python program to extract the affiliation names and write them into a file together with DOI, our primary key. And the visualization is about the statistical result of the countries where these affiliations are located. We manually transcribed the affiliation name into countries since the number of them are not very large and luckily many affiliations claim their bases in their names.

The challenge was, that each article have normally more than one author, and different authors can have different affiliations. And to decide which affiliation to go with was beyond what our python program can do. So we decided to make extinct the status of the first author (or lead author) and use only their affiliation as the samples in our report and presentation.

Task 2: Data schema and database set up

1 Create a meaningful data schema for your data set(s) and draw an according Entity-Relationship model (ERM).

The ERM was a relatively simple part. Naturally "article" is an entity that has many attributes. Here we neglected some attributes that were listed in the original data file to achieve a more concise view for analysis and understanding.

At first, we didn't tell apart the author and lead author. But we then realized, after working on the relational model, that only keeping lead authors' information is not true to the data we received. So we separate the lead author into a new entity and completed the table of all authors and modified the ERM.

2 Transform your ERM into a relational model.

Transforming an ERM into a relational model can be a process of recognizing the inappropriate design in the ERM, just as in the story mentioned above. But also, ERM can be very enlightening in writing the tables of the relational model. An entity, together with its attribute(s), if any, forms a table, and these tables can be put into and processed in postgresql.

3 Set up a PostgreSQL database with your relational model.

To create tables in postgresql we needed a tutorial for it.

CREATE TABLE Public.AUTHOR_WRITE_ARTICLE(DOI VARCHAR(200), LEAD_AUTHOR VARCHAR(500)) is an example.

Task 3: Pre-processing and import data .

1 Pre-process (data cleansing) the data set(s).

In pre-processing we reduced the duplicate tuples (we rule duplicate through DOI) to one and removed all DOI with NULL or unrecognizable (for example, ‘?’) values because they cannot be processed and counted as a legal entry.

2 Import the data set(s) into your database.

```
COPY Public.AUTHOR_WRITE_ARTICLE FROM
'C:\Users\**\Desktop\database_system\database_project\AUTHOR_WRITES_ARTICLE.csv'
DELIMITER ',' CSV HEADER;
```

Task 4: Develop a Web application (Details in ZIP File)

1 Develop an interactive Web application to access your database.

Here we need to learn R shiny to achieve the interactive functions of a web application.

2 Visualize the data in a problem-oriented way.

We made a line graph with year as a dimension and investment as another. Users can pick which country or countries’ data they want to see.

And histogram via as an example:

```
ARTICLE<-
read.csv("C:/Users/Yumeng/Desktop/database_system/database_project/New
folder/ARTICLE_C.csv")
x<-ARTICLE$Feld
require(ggplot2)
q<-qplot(x)
q+theme(axis.text.x=element_text(angle=90,vjust=0.5,hjust=1))
```

3 Develop sorting and filter functions for your visualized data.

We developped the filtering functions in the ZIP file. Customers can choose which part of information they would like to show, and check the corresponding boxes. For more details, see in the ZIP file.