Lab course tasks

Notes

Group work and project planning

Each group plans their work and divides tasks evenly between themselves. This plan should be demonstrated at the 3rd week of lab course. During the course each group presents the results of its work after each task. The dates of the presentations will be set. At the end of lab course each group provides the overall presentation, the source codes and all the working datasets.

The schedule

Task number	Task name	Submission deadline	Presentation deadline
0	First start	22.04.2016	25.04.2016
1	Literature review	6.05.2016	9.05.2016
2	Preprocessing	20.05.2016	23.05.2016
3	Rewriting UDFs	3.06.2016	6.06.2016
4	Getting domains	10.06.2016	13.06.2016
5	Clustering and distance function	8.07.2016	11.07.2016
6	Final presentations	15.07.2016	18.07.2016

Task 0. First start

Submission deadline: 22.04.2016, 8:00 a.m.

Need to be done by 25.04.2016

1. Install Eclipse, get the source codes, build them, and make sure you do not have compilation errors.
2. Become familiar with SkyServer database, we expect you to have some domain knowledge
3. Install Oracle SQL developer: http://www.oracle.com/technetwork/developer-tools/sql-developer/downloads/index.html
4. Have a look at the query log in Oracle database.
5. Run test samples (project)

The full description here: Tasks\0_First_Start.docx

We expect as a result:

1) Output file with clusters
2) Cluster interpretation. What do these clusters mean?
3) Change threshold parameters. Evaluate the influence of thresholds to the final result.
4) Short presentation regarding the output results

Note: we really expect you to ensure that you have all the data you need for the future work.

Task 1. Literature review

Submission deadline: 6.05.2016, 8:00 a.m.

Need to be done by 9.05.2016

> Read the article H. V. Nguyen et al. “Identifying User Interests within the Data Space – a Case Study with SkyServer”. In: *EDBT*. OpenProceedings.org, 2015, pp. 641–652. http://dbis.ipd.kit.edu/download/IdentifyingUserInterests.pdf

We expect as a result:

The review which should consist of:

- The short description of the content (≈3 slides)
- Limitations of the approach (so many slides as you need). Please, provide argumentation with examples why do you think it is a problem
- Suggest solution for limitations you found

Task 2. Preprocessing

Submission deadline: 20.05.2016, 8:00 a.m.

Need to be done by 23.05.2016

1. Extract data (SQL log) from Oracle database. For our experiments we need 4 sets of data:
    1) The log with UDFs, copy-patterns and antipatterns
    2) The log without UDFs, but with copy-patterns and antipatterns
    3) The log with UDF and copy-patterns, but without antipatterns
    4) The log with UDF but without copy-patterns and antipatterns

Note: the way you collect data is not important but I suggest to write simple script to do so. Otherwise you will spend **too much time for this –** and this is not the thing you supposed to do for a long time.

Log type	How to collect
The log with UDFs, copy-patterns and antipatterns	Query table in Oracle database. Important: we need not less than 2 million queries. Three groups should collect **DIFFERENT** data. Suchwise, they should **reach an agreement** of what part of data they intend to collect. For example, the first group get data from 1.01.2003 to 1.07.2004, the second group – from 1.07.2004 to 1.04.2005 etc.
The log without UDFs, but with copy-patterns and antipatterns	The full list of SkyServer user define functions you can find here: http://skyserver.sdss.org/dr1/en/help/browser/browser.asp We need to collect the data for THE SAME period of time, but without any UDFs.
The log with UDF and copy-patterns, but without antipattern	The table “PARSED_STATEMENTS” from which you are supposed to collect queries consists of the column “can_be_stifle” and “stat_id”. The other table “FROM_WHERE_STATEMENTS” has columns “ID”, “count” and “distinct_ips_count”. “PARSED_STATEMENTS”.”stat_id” = “FROM_WHERE_STATEMENTS”.“ID”. You need to collect the data for THE SAME period of time where can_be_stifle = 0
The log with UDF but without copy-patterns and antipatterns	You need to collect the data for THE SAME period of time where can_be_stifle = 0 and STAT_ID which has a little amount of distinct IPs (less than 20) and large amount of count (more than 1000)

The full description here: Tasks\1_Preprocessing.docx

We expect as a result:

- 4 datasets with query log described above
- The evaluation of the datasets (short presentation, some slides): how do they differ (the size) one from the other? What will it change in terms of runtime? (Let’s assume that for one comparison we spend the time t) Some graphics would be nice.

The results of the preprocessing step should be included to the final presentation as well

Note: if you have any questions regarding this task please ask them at 16.05.2016

Task 3. Rewriting UDFs

Submission deadline: 3.06.2016, 8:00 a.m.

Need to be done by 6.06.2016

The other step of our processing is converting data to intermediate format. This includes rewriting queries with UDFs to ordinary queries.

1. From the data YOU use investigate the most popular UDFs. Provide statistics regarding UDFs usage. Implement the first 10 UDFs which mean rewriting them to non-UDFs format. If there is no accurate implementation do approximate one
2. Test your transformation and present the result
    - Provide the transformation rule for each UDF
    - Show at the real data from Sky Server that we get the same (or approximately the same) result with UDFs and with rewritten queries. The results will be checked by me. Please make sure the rewritten queries do not occur syntax errors.
3. Analyze the improvement in the input data when we include UDFs in it.

The full description here: Tasks\2_UDFs.docx

We expect as a result:

- 4 datasets in intermediate format
- The evaluation of datasets (short presentation, some slides)

Task 4. Getting domains

Submission deadline: 10.06.2016, 8:00 a.m.

Need to be done by 13.06.2016

At this step we collect columns distributions for the future analysis. So far we have 4 “types” of columns (please, note this division is not related to the actual type of the column):

Name	Description
Identificator	$count(columnName) = count(*)$
DictionaryField	$count(columnName) <= MxCtbD\ (now\ it\ is\ = 2000)$
DistributedField	$count(columnName) < count(*)$, evenly distributed
DistributedFieldWithEmissions	$count(columnName) < count(*)$, BUT not evenly distributed

1. Make suggestions how else could we divide columns
2. Run the code which get distributions for the columns
3. Find columns where we did not get the distributions. Investigate why was it happened? How to fix this problem? Provide your solution and implement it
4. Get familiar with the current implementation
5. Run clustering process for all 4 datasets

The full description here: Tasks\3_Domains.docx

We expect as a result:

- The file with columns and their distributions
- The short presentation regarding what difficulties you faced with, how you fixed them.

Task 5. Clustering and distance function

Submission deadline: 8.07.2016, 8:00 a.m.

Need to be done by 11.07.2016

1. Get familiar with the current implementation
2. Run clustering process for all 4 datasets
3. Evaluate the runtime. Make suggestions how could it be reduced.
4. Evaluate the quality of clusters. Change the parameters of the thresholds, evaluate the difference. Find the optimal thresholds. Substantiate your choice
5. Make suggestions how to improve the distance function; implement it; test it; prove it

The full description here: Tasks\3_Clustering.docx

We expect as a result:

- The files with clusters for each dataset, for each implementation of the distance function
- Present comparison of all the datasets and the results
- Present YOUR SOLUTION for the distance function. Please, prove your point of view, give examples; we expect also the presentation of results of your improvement

Note: This is the most important part of your study. Think carefully of what you intend to do; you need to provide at least one suggestion how to improve the distance function. The meetings 21.06.2016, 27.06.2016 and 4.07.2016 will be devoted to discussions of your ideas.

Task 6. Final presentations

Submission deadline: 15.07.2016, 8:00 a.m.

Need to be done by 18.07.2016

By this time we expect to have the overall presentation, the source codes and all the working datasets from you. The presentations should last for about 20 minutes. The other teams could ask questions after the presentation.