5. Web Analytics/Web Usage Mining

Task

- Use an existing dataset from a travel agency wum_dataset_hw.zip
  - clicks.csv
    - LocalID - local identifier of an event
    - PageID - identifier of a visited page
    - VisitID - session identifier
    - PageName - page label
    - CatName, CatID - page type (Navigation), general information
    - ExtCat,ExtCatID - page type (Content), more specific information
    - TopicName, TopicID - topic
    - TimeOnPage - time spent on the page. Last page of the session = 30s.
    - PageScore - weight of the page using the following heuristic: (ln(o)+1)*t
    - SequenceNumber - page order within a session
  - visitors.csv
    - VisitID - session identifier
    - Referrer - anonymized referrer
    - Day - day of the visit
    - Hour - hour of the visit
    - Length_seconds - visit length in seconds
    - Length_pagecount - visit length as number of visited pages
  - search_engine_map.csv
    - Referrer - anonymized referrer
    - Type - type of the referrer domain
- Tasks
  - Execute data preprocessing
    - Design a suitable data representation for the analysis - association rule mining + any other analysis of your choice
      - e.g. file where each row represents one user visit/session and columns including all interesting descriptions or summaries of the visit
        - user-transactions matrix, pageview-feature matrix, transaction-feature matrix
    - Remove too short visits
    - Implement any other data cleaning mechanism
    - Identify conversions in data
      - main conversions
        - APPLICATION (reservation of the trip) or CATALOG (request the printed catalog) in the PageName (or category) attribute of clicks
      - micro conversions
        - DISCOUNT, HOWTOJOIN, INSURANCE, or WHOWEARE in the PageName attribute of clicks

- Implement pattern extraction
  - Identify interesting association rules in the data (e.g. conversions in the consequent)
  - Realize any other analysis of the data of your choice (e.g. users, visits clustering etc.)

Instructions for submitting

In your private namespace on EDUX provide the following information:

- Provide general statistics about the dataset
  - e.g. visits, users, conversions, …
- Describe your preprocessing/cleaning operations and demonstrate all steps using examples
  - Describe the final dataset suitable for the data analysis
- Perform the association rule task
  - Describe the input and task settings
  - Present the outputs and try to interpret/explain
- Perform any other data analysis task
  - Describe the input and task settings
  - Present the outputs and try to interpret/explain
- Provide the link to your implementation
  - You can use: https://gitlab.fit.cvut.cz [https://gitlab.fit.cvut.cz] or https://github.com/ [https://github.com/] or https://bitbucket.org [https://bitbucket.org]
- Comment on
  - issues during the design/implementation
  - ideas for extensions/improvements/future work

/mnt/www/courses/MI-DDW.16/data/pages/hw/05/start.txt · Poslední úprava: 2017/04/25 18:45 autor: kuchajar