ISEL - ADEETC

FINAL PROJECT – B

Subject: AMD – Aprendizagem e Mineração de Dados

Course: Mestrado em Eng. de Redes de Comunicação e Multimédia Year: 2012 / 2013

Goal: Explore the relation among the notions of data, information and knowledge. Apply those concepts in specific scenarios and experiment (and develop) the techniques that allow to "move" from data into knowledge. Explore the classification, association rules and clustering that are nuclear to data-mining, knowledge discovery, machine learning and information retrieval. Techniques resort to statistics (e.g., Bayes rule), induction of association rules (e.g., Apriori), decision trees (e.g., J4.8/C4.5, ID3) and instance analysis (e.g., KNN with KDTree support). Develop the competence of analyzing, modeling and validating a data-mining project. Use tools to manage data (e.g., PostgreSQL), to discover knowledge (e.g., Orange data mining) and to implement specific algorithms (e.g., via Python).

ScenarioB. The "SoftKnow" company had a meeting with the "We-Commerce" company; the meeting was focused in the e-commerce business of "We-Commerce". Their innovation group asked "SoftKnow" for technical advises on how to generate knowledge (i.e., business value) from the large volume of data that "We-Commerce" has been storing throughout time. The "We-Commerce" captures and stores each event generated by visitors that navigates through their e-commerce Web sites. After the meeting the "We-Commerce" wrote a report about the process they use to collect data and sent it to "SoftKnow"; they also sent the dataset with some of the collected data.

The "We-Commerce" report. The first time a visitor arrives at one of our Web pages we store a cookie in its browser. The visitor's cookie value is assigned a "globally unique identifier – GUID" (implemented by "We-Commerce" and named GUI). The next time the visitor arrives at one of our pages we retrieve (from that visitor's browser) the value of the stored cookie and we insert, the value of that cookie, in our database (in the "cookie_id" attribute). It is important to remark that a new session identifier is created whenever a visitor arrives at one of our pages. The session expires when the browser closes or the visitor has been inactive for some period of time. The session identifier ("session_id") is inserted in our database (along with the "cookie_id" value). Therefore, each "cookie_id" may get associated with a list of different "session_id" values. Additionally, for those visitors that are already subscribed users we also insert (in the database) that user's unique identifier ("user_gui"). Another important information that we keep record is the product unique identifier ("product_gui") that gets visited (by a visitor within a session). We remark that occasionally we get useless information inserted in the "product_id" attribute; e.g., some of the useless values are: "open", "home", "/customer/account/login/", "/customer/account/forgotpassword/" (we are improving but at this time the dataset still contains such useless values). We also register the Internet Protocol (IP) address of each visitor but as the same visitor may get the IP dynamically this information is less useful than the one stored in the "cookie_id" attribute. Each record of data is assigned a unique identifier ("tracking_record_id" attribute) and we say that each record represents an "event" (or a visit). Each event is described by additional attributes such as: a timestamp ("date_time" attribute), the name of the company ("company") that provides the product ("product_gui") associated with that same event, the URL of the visited Web page ("link" and "refer" attributes), the identification of the visitor's browser ("browser" attribute), the identification of a promotional campaign ("campaign_id") whenever a product belongs to a campaign.

The "We-Commerce" dataset. Attach to our report (cf., above) we also send a dataset with the events that were collected (during JAN.2012); the file is named "`z_dataset_201201.csv`" with around 420.000 registered events. We also send a smaller sample ("`z_datasetSample_201201.csv`") with the first 30 events taken from that dataset.

Project Items:

Recall that "We-Commerce" asked "SoftKnow" for technical advises on how to generate knowledge (i.e., business value) from the large volume of data that "We-Commerce" has been storing throughout time.

Assume that your working team represents "SoftKnow" company and that you teacher represents "We-Commerce". In order to reply to the "We-Commerce" challenge your company ("SoftKnow") decides to develop the following items:

ISEL - ADEETC

FINAL PROJECT – B

Subject: AMD – Aprendizagem e Mineração de Dados

Course: Mestrado em Eng. de Redes de Comunicação e Multimédia Year: 2012 / 2013

1. Analyze the dataset file (or the smaller sample) and write down the meaning of each attribute; consider the report received from "We-Commerce" and also your own (à-priori) assumptions about their meaning.

2. The first goal is to search for association rules among products ("cf., "product_gui" attribute). Formulate this goal as a "market-basket analysis" problem.

3. Given the dimension of the dataset start by generating an aggregated view of the data distribution. Get the number of total events, the total number of visitors (i.e., the cardinality of "cookie_id" domain), the distribution of visitors and sessions (i.e., number-of-visitors versus number-of-sessions), and the distribution of sessions and visitors (i.e., number-of-events-per-session versus number-of-visitors). In order to validate your results you can use the information provided in the file "`z01_aggregatedView.txt`" that was generated from the "`z_dataset_201201.csv`" file after being imported into PostgreSQL and manipulated via SQL statements; for testing purposes, the file also contains the "cookie_id" (or "visitor_id") value of the visitors with 18 sessions.

4. In order to reduce the complexity (computation time) of the "market-basket analysis" generate a subset of the most relevant data. In this case we may focus our analysis in the subset that contains all the events generated from visitors with a number of sessions within a certain range; e.g., with $5 \leq$ number-of-sessions ≤ 30.

5. Generate a dataset file with the recorded transactions (i.e., `TRANSACTION_ID;PRODUCT_ID`) that were filtered according to the criteria defined in the previous item. *Hint*: use the "`psql`" "`\COPY FROM`" command.

6. Process the file with the recorded transactions with the following two goals: a) "normalize" the strings that describe each product, e.g., eliminate white spaces, eliminate accents, set to lower case, and b) generate a "`.basket`" Orange file. *Hint*: use the provided "`_goPy_transform.py`" code and implement the remaining.

7. Use Orange to graphically generate a "market-basket analysis" report (to send to "We-Commerce").

8. Implement the code to automatically execute the previous items 4, 5, 6 and to programmatically (i.e., using Orange API) make the "market-basket analysis".

Important dates (deadlines) and deliverables:

- ***until the day of first exam of this subject*** – deliver all elements described below (cf., "Rules" paragraph).

Rules:

- Deliver a project report, in printed form, with at most 10 pages; the first page must include the discipline and course names, a title, the working group number and each student's name and number.
- Deliver, in electronic format (file named AMD_XX.zip, where XX is the working group number), the project report and all the information regarding the implemented system; the ".ppt" (powerpoint), ".bat", ".exe", ".py", ".tab", ".basket" (data) and any other file needed to properly execute your solution.