Intelligent systems - Data Exploration

BEAUMARD Colleen, COSTA Claésia, HENG Soklay,
PIBRIL Peter, PRIYA Shalini

January 2022

Introduction

As a machine learning students, we are contacted by a retailer company "Diginetica". The company wants us to build a recommendation system to recommend the best product ranking to the customers. The data is provided by the company, which will be used for our model training dataset.

Diginetica Train Dataset

Before going into the trainings, we had to analyse precisely the dataset. It contains several files, each one was focusing about a precise question. We checked all of them:

- *ItemID* is unique to each item *
- *UserID* is unique to each customer *
- *SessionID* is unique to each session started by a customer *
- *Duration* is the time spent on the results page of a query
- *Time Frame* is the time between the start of the session and the first query
- *Event Date* is the date when the session was opened
- *CategoryID* is the category of an item (not unique) *
- *QueryID* is unique to each customer (queries for an object are differents) *
- *Searchstring.tokens* tokens for the query (comma are used as separators for tokens)
- *is.test* if the query was a test (TRUE) or not (FALSE)
- *pricelog2* is the log transformed item price

- *product.name.tokens* is the name of the item in token (comma are used as separators for tokens)
- *ordernumber* is the ID of the purchase for each customer (if the customer bought several products, each product will have the same *ordernumber* ID)
- *Items* is the itemIDs returned by the ranking algorithm (must be re-ranked)

The columns followed by a * are serial.

Column	Number of unique items
queryId	**923127**
sessionId	**573957**
userId	**232931**
timeframe	**1251323**
height*duration*	**7392**
eventdate	**169**
searchstring.tokens	**26137**
categoryId	**1218**
items	**97108**
is.test	**2**
itemId	**184049**
ordernumber	**13506**
pricelog2	**13**
product.name.tokens	**182392**

Table 1: Number of unique items per column

Additional information:

- In majority, customers who know what they want can spend about the same amount of time on a page (**1626 ms**) than customers who do not know what they want Knowing what they want (**1680 ms**)
- The category 0 is more viewed than anyone else (**35195 views**)

Future work

Our first idea is to use content-based filtering approach. We want to build a user profile, and then the items will be recommended to the users based on their behaviors (what they click on and what they purchase). We are also interested in item-to-item filtering by matching each of the user's purchase with a list of similar items to what they purchased. From the table above, we have found most useful columns for our recommendation system building, such as: ItemID, CategoryID, UserID, and Time Frame in this meantime.